# The Basics of Statistical Design and Analysis of Experiments

Rickie J. Domangue
Department of Mathematics and Statistics
James Madison University

June 23, 2015

# Contents

# Preface

The purpose of writing this text is to provide a presentation of statistical methods and concepts associated with the design and analysis of experiments geared toward undergraduate students with only an introductory background in probability and statistics and with an interest in an applied, projects-orientation to the subject.

The text arose from lecture notes used in multiple classes of an undergraduate level course in experimental design and analysis of variance (ANOVA), Math 321, taught at James Madison University over a number of years. The class meets three times per week for approximately 15 weeks in a semester. The students major in a variety of disciplines, including mathematics, statistics, psychology, biology, health science, and environmental science. The prerequisite for the course is an introductory course in statistics, calculus or non-calculus based. Students usually take the class in either their junior or senior level, but there are occasions when students are in their sophomore year.

In the first few years of teaching the course I used several undergraduate level texts. While these texts have their strengths, I did not find them totally suitable for the class given the background and goals of the class. Some books were geared more toward specific disciplines, such as engineering and agriculture. One book was compatible in many ways but I wanted a little more mathematical rigor given that some students in the class were mathematics and statistics majors.

The specific goals of the text are as follows:

- The first chapter should be relatively math-free introducing students to the ideas of good experimental design, such as randomization, blocking, and replication within the context of a variety of applications, some of which are in the areas of interest of the students.

- I felt like the text had to have a significant formal review of basic concepts in statistics (Chapters 2 and 3) since some students were entering the class with only one introductory course with maybe having taken that one course two or three years earlier. These two chapters should provide a smooth transition to the rest of the text on the more advanced ANOVA concepts and analyses. I typically spend about three weeks on these two review chapters.

- The text also needed to have examples and exercises from a variety of disciplines because of the varied background of the students and the need to keep the course relevant for them.

- The mathematical level had to be accessible to students with only a first course in statistics with no calculus background. The deas of analysis of variance should be presented at a basic level, although design issues could and should be challenging and thought-provoking.

- The text should provide examples of experiments conducted by students in previous semesters and report the results of those experiments. Seeing results of experiments done by students in previous classes should make students more confident in their abilities to perform their own experiments. I regularly require students to do their own experiments in my classes. As Box and Liu noted, " The art of investigation cannot be found just by playing with someone else's data." ("Statistics as a Catalyst to Learning by Scientific Method Part I - An Example," Journal of Quality Technology, Vol. 31, No 1, January, 1999) In this regard the text is similar to the test of Dean and Voss ([6]) but on an undergraduate rather than a graduate level.

- The number of chapters should be relatively small (total of 10 chapters) and cover the basic concepts. Other textbooks have larger numbers of chapters with topics that cannot possibly be covered in a one-semester course for students with only a minimal background in statistics.

- The text should not be tied to a particular computing package. I used SAS in class and have SAS code at the ends of the chapters. However results of analyses are not given as output from particular packages. This allows other instructors to use the text with their favorite package.

- Checking of assumptions of analysis of variance methods is undertaken in a single later chapter in the text in order to first concentrate on the details of the analyses.

This work was supported by the James Madison University Program of Grants for Faculty Assistance. I am extremely grateful to JMU for the opportunity to complete a career goal.

I would like to thank the many students who took my Math 321 classes over the years for the interesting conversations about statistics and for providing me with many real data sets from their experiments, some of which are used in the text. Students whose experiments and data are used in the text are referenced by name followed by the semester and year that the students took my class.

I would like to thank Steve Patch at UNC-Asheville and Dinesh Sharma, Lihua Chen, and Hasan Hamdan of JMU for giving the text trial runs in their classes and providing feedback. I would like to thank all of the other statistics faculty in the Department of Mathematics and Statistics for giving me the chance to teach my favorite course so many times.

I would like to thank my wife, Reinhild, for her love, support, and patience through so many working weekends, summers, and holidays, allowing me to complete this long endeavor.

Rickie J. Domangue

# Chapter 1

# The Nature of Experimentation and Analysis of Variance

## 1.1 Types of Statistical Studies

In this book we are going to be concerned with statistical studies in which groups of individuals or objects associated with different conditions are compared in terms of some characteristic. We will be mainly interested in experiments but will occasionally investigate observational studies as well. The two kinds of studies, experiments and observational, are described below.

- **Experiment:** A study in which the conditions are deliberately (and usually randomly ) assigned by a researcher to individuals/objects/time slots for the purpose of seeing the effect that these assigned conditions have on some characteristic. The assigned conditions are called **treatments** . The characteristic is called the **response** variable . The individuals or objects are called the **experimental units** .

- **Observational Study:** A study in which the conditions are not assigned/controlled by the researcher but simply observed. The conditions are inherent characteristics of the subjects/objects/time slots. Interest still lies in comparing the groups defined by the conditions in terms of the response variable.

  In this text it is assumed that the response variable is quantitative. The conditions may be categorical or quantitative.

## 1.2 Examples of Experiments

In this section examples of experiments are given and some basic terminology is introduced.

a. In a study to determine if number of calories consumed affects longevity, 60 mice were given diets differing by number of calories. Twenty mice were randomly assigned to a low calorie diet, twenty to a medium calorie diet, and twenty to a high calorie diet. The number of months that the mice lived was recorded (mice have an average lifespan of about 2 years).

   This study is an experiment. The diets that the mice get are controlled or assigned by the researcher. The different diets-low, medium, and high calorie are the treatments. The response variable is the lifespan of a mouse measured in months. The experimental units are the 60 mice.

b. Pop-up ads are advertisements that pop-up on your computer when you are visiting or leaving a website. An internet service provider conducted a study to see if reducing the number of pop-up ads would improve satisfaction with their service. A group of 1000 subscribers were randomly selected. Half of them, randomly selected, saw roughly half the usual number of pop-up ads when visiting the website. The other half saw the usual number of pop-up ads. After two weeks the 1000 subscribers were asked to fill out a satisfaction survey regarding how they feel about the provider.

   This study is an experiment. The number of ads, “usual” or “half” are conditions or treatments assigned to the subscribers, the experimental units. The response variable is the satisfaction survey score.

c. Medical research has explored the medicinal uses of garlic. In one study 60 mice were fed high-cholesterol diets. Thirty of the mice, randomly selected, were given allicin, one of garlic’s active ingredients. These 30 mice developed fewer fatty deposits in their arteries than the 30 mice not receiving allicin. The experimental units are the 60 mice. The researchers determined which mice received allicin and which did not. The treatments are “received allicin” and “did not receive allicin.” The response variable is the number of fatty deposits in the arteries.

   The 30 mice making up the group not receiving allicin is called a control group. A **control group** is a group that gets a standard treatment, no treatment at all, or a sham treatment. The control group serves as a basis of comparison.

d. In a study of a new headache relief medicine 100 headache sufferers were divided at random into two groups, with one group getting the new headache relief medicine and the other group a **placebo**, an inactive substance designed to look like the new headache medicine.

The placebo group is a type of control group, a comparison group, used to control for the **placebo effect**. In medical studies with human subjects, often patients respond positively to any treatment, even dummy treatments, presumably due to attention being paid to them. This response is called the placebo effect.

To determine if a new treatment is truly beneficial or just the placebo effect at work, another group is given a placebo, rather than nothing at all. If the new group receiving the new medicine is really beneficial, then it should do better than the group getting the placebo.

e. In some animal health studies treatments are medicines provided in feed or water for the control of certain infections. Animals may be housed in pens. The medicine is then in the feed or water of the penned animals and animals eat or drink from the same source. In such studies the experimental units are pens of animals rather individual animals since the medicines are assigned at random to pens. The response variable may be a summary of the pen of animals such as average daily weight gain (lbs/pig/day) for all animals in a pen or average feed consumption (lbs/pig/day) over some period of time for all animals in a pen.

f. An experiment was conducted to determine the differential effects of three different fat/carbohydrate diets A, B, and C on how much time (seconds) subjects could stay on a treadmill until exhaustion. Each of six subjects received all three diets at different time periods, the order being randomly determined. The conditions in this experiment to be compared are the three diets. However the three diets are not assigned to different groups of subjects. Each subject received all three diets. So the (random) assignment in this experiment is the assignment, for each subject, of the three different diets to be used at the three different time periods. So **time periods** are the experimental units with a total of eighteen, 3 for each of the 6 subjects. The response variable is the amount of time until exhaustion for a particular subject on a particular diet at a particular time period.

g. An experiment was conducted to examine the effects of external distractions (none, constant, changing) and type of words (fruit, mixed, nouns) on the ability to memorize words. There are nine (3 x 3) combinations of external distraction and type of word. Thirty-six subjects were assigned at random to the 9 combinations, with four subjects per combination. A list of 30 words was prepared for each of the three word types. Each subject studied his/her randomly assigned word list under his/her randomly assigned distraction type for a fixed amount of time. The response variable was the number of words correctly remembered out of 30. The experimental units are the 36 subjects. The treatments are the 9 combinations of external distraction and type of word.

h. An experiment was conducted to examine the effects of depth perception (both eyes open, left eye covered, right eye covered) and distance from a basketball goal (free throw distance and 1/2 free throw distance) on the number of successful throws at the basket. There are six (3 x 2) combinations of depth perception and distance from the goal. Twelve students with some experience in playing basketball were recruited. The twelve students were randomly assigned to the six combinations, with 2 students per combination. Each student took 30 shots at the basket using one of the randomly selected depth perceptions and distances. The response variable was the number of successful attempts among the 30 shots conducted by a student. The experimental units are the twelve students. The treatments are the six combinations of depth perception and distance from the goal.

i. Mark Bergenholtz, Michelle Clower, and James Skiba in 2014 investigated the effects of stove temperature (levels of 4, 6, 10) and type of meat (chicken, beef, and pork) on the amount of time for a piece of meat heated in a skillet to reach an internal temperature of 165°F. The experiment was conducted over time using one skillet as follows. At a particular heating time period a randomly selected stove setting and type of meat was determined. A piece of meat for the randomly selected type of meat was randomly selected. The piece of meat was heated in a skillet at the selected setting and the amount of time (minutes) for the piece of meat to reach the desired internal temperature was determined. Pieces of meat within type of meat and across types were similar in size. This activity was repeated over 45 different time slots/heatings in a completely random order, five time slots/heatings per combination of heat setting and type of meat. Experimental units are the 45 time slots/heatings. There are nine treatments corresponding to the nine combinations of heat setting and type of meat. Each treatment is being applied 5 times.

## 1.3 Examples of Observational Studies

In this section examples of observational studies are given, emphasizing the difference between these studies and experiments.

a. When backing out of a parking space in a lot, do people take longer when someone is waiting for them as compared with no one waiting? There have been studies that looked at this question. Suppose a researcher does a study by observing people getting into their cars in a university parking lot. The researcher records whether or not someone is waiting to obtain the parking spot and also how long it took the driver leaving to depart. The response variable is the amount of time to depart from the time that the person stepped into his/her car until he/she moved forward. Also observed was whether or not there was someone waiting to take the person's spot.

   This is not an experiment. It's an observational study because the conditions "someone waiting" and "someone not waiting" are not assigned by the researcher but observed.

b. Researchers wanted to know if IQs of children related to whether or not they were breast-fed. Researchers measured the IQs of a large number of first graders in a large city. The researchers also asked the mothers of these first graders whether or not they had breast fed their children. The researchers found that IQs of children who had been breast-fed were greater on average than those children who had not been breast-fed.

   This is an observational study. The conditions that are being compared, "breast-fed", "not breast-fed" were not assigned by the researchers to the children. These conditions were presumably selected by the mothers of the children. The response variable is IQ of a child. The "experimental", or more accurately, **observational units** are the children.

c. Suppose you want to compare reading level by way of sentence length for two magazines, **People** and **Teen People**. You randomly select 100 sentences from an issue of People and 100 sentences from an issue of Teen People. For each sentence you determine the number of letters and punctuation signs and then compare the average sentence length for the two magazines.

   This is an observational study. The conditions associated with each sentence, **People** and **Teen People** are not assigned, but are inherent characteristics of the sentences. The observational units are the sentences and the response variable is sentence length.

## 1.4 Variables in an experiment

A **variable** is a characteristic of a person or object that varies from person to person or object to object. So examples of variables are height, eye color, population of a city and color of a car. The possible **values** of a variable can be **quantitative**, such as for height, or **categorical**, such as for eye color.

The response variable in an experiment has been previously defined as the characteristic associated with the experimental units which is compared for different groups that have been assigned treatments. In this text we will be concerned with studies where the response variable is quantitative, such as longevity of a mouse or number of fatty deposits in the arteries.

In some experiments the treatments are one-dimensional and are the values of a single variable called the **factor** in the study. In the longevity study the single factor of interest is diet. There are three values or levels of the factor diet: low, medium, and high calorie, which are the treatments. The purpose of an experiment is to determine if the factor affects the response variable. Many of the studies in this text have categorical factors. However factors can be quantitative, such as dose level of a drug.

In some experiments there are **two (or more) factors** under study. For example in the short term memory example in the last section there were two factors of interest, distraction type and word type. The levels of distraction type were: none, constant, and changing. The levels of word type were: fruit, mixed, and nouns. **Treatments in a two-factor study refer to the combinations of levels of the two factors.** In the short term memory example there were two factors with 9 treatments. In the basketball study from the last section the two factors were depth perception with 3 levels and distance from the goal with 2 levels. The treatments were the six combinations of the levels of the two factors depth perception and distance from the goal.

There are typically other variables in an experiment that researchers need to take into consideration when designing an experiment. An **extraneous variable** is a variable not of main interest in the study but believed to be associated with the response variable.

A student performed a class experiment to determine whether microwaving oranges results in more juice being squeezed from the oranges. The factor of interest is categorical with two levels: microwaving and not microwaving. The response variable is amount of juice squeezed from an orange. An extraneous variable in this study would be the size of the orange since larger sizes would presumably result in more juice than smaller sizes. Another extraneous variable would be the amount of pulp in the orange. The color of the orange or how many dimples on the peel, while variables, are not extraneous variables.

In the study of different fertilizers on the effect of amount of tomatoes (in pounds) grown on a plant, extraneous variables include variety of tomato, amount of water or sunlight the plant receives, and soil fertility. In the short term memory study of the last section an extraneous variable would be the natural ability or inability of subjects to memorize words. In the basketball shooting example of the last section an extraneous variable would be the basketball shooting experience of an individual.

## 1.5 What's affecting the response variable?

Extraneous variables in an experiment are important to recognize and control since differences on the response variable across the treatment groups may be the result of extraneous variables, not the factor of interest.

Consider an experiment designed to compare two fertilizers (A,B) on the amount of tomatoes grown. Suppose that ten plants of about the same size and variety are used. The ten plants are assigned at random to ten plots in a garden. Five are randomly assigned fertilizer A and the other five receive fertilizer B. The resulting yields in pounds are given below:

Fertilizer A: 45, 50, 47, 57, 52
Fertilizer B: 48, 52, 53, 48, 56

Note that fertilizer B yields appear to be slightly larger than those for fertilizer A. The mean yields for fertilizer A and B, respectively, are 50.2 and 51.4 pounds. Can we say conclusively that fertilizer B is better? Note that WITHIN the same treatment or fertilizer group (A or B) the yields of the tomato plants vary because of presumably extraneous variables. The different plots will have slightly different fertilities. The plants, while all of the same variety, will have slightly different genetic makeups. This variation in values of the response variable for identically treated plants is referred to as **experimental error**. So maybe the slight differences seen in yields BETWEEN the two groups are not due to fertilizer, but are actually due to variation resulting from extraneous variables or due to experimental error. Perhaps just by chance (random assignment) the plants receiving fertilizer B were placed in plots that were a bit more fertile or these plants were genetically predisposed to greater production. In order to judge whether treatments really differ it is necessary to have some idea of expected differences in groups simply due to experimental error. Methods of Analysis of Variance or ANOVA are concerned with the measurement of differences **between** treatment groups, measurement of differences **within** treatment groups, and the relative comparison of the two types measurements.

There are various sources of experimental error , such as natural variation in experimental units, inability to identically treat the units in the same group, inability to measure precisely. In general all extraneous variables contribute to experimental error.

The key to designing a good experiment is to “control” the variation resulting from extraneous variables. Controlling doesn’t mean getting rid of the effects of the extraneous variables altogether, although sometimes that can be done. For example, variety of tomato is an extraneous variable that we can control by using the same variety. Controlling means NOT letting the effects of the extraneous variables enter in a “systematic” way but only in a “random” way.

A **systematic effect** of an extraneous variable would be an effect which generally goes one way: for or against a particular treatment. For example, if we only watered the fertilizer A plants, that would be a systematic effect of watering. This activity would **bias** the comparison in favor of fertilizer A.

A **random effect** of an extraneous variable would be an effect which sometimes favors fertilizer A and sometimes favors fertilizer B. For example, if we randomly assigned the plants to fertilizer A and fertilizer B, then the genetic predisposition for larger tomato production of a particular plant might sometimes favor A and sometimes favor B. Overall the effects of this extraneous variable would be mostly canceled out and thus we would have a fair comparison in terms of genetic predisposition.

Extraneous variables whose effects enter an experiment in a systematic way result in an association between the extraneous variable and the factor, in addition to the (possible) association between the extraneous variable and the response variable. Then it’s impossible to tell whether any differences in the response variable between the groups is because of differences in the treatments or differences in the extraneous variable. The extraneous variable then becomes a **confounding variable** and we say that the effects of the treatments are confounded with the effects of the extraneous variable. So in the example above water would be a confounding variable, whose effects are confounded with the effects of fertilizer.

## 1.6 Confounding and Observational Studies

In designing experiments it is often possible to ensure that the effects of extraneous variables are controlled and enter only in a random way. Section 1.8 presents some principles for doing this.

Researchers doing comparative observational studies often hope to show that the factor of interest causes changes in the response variable. However in an observational study groups determined by the factor levels may also differ in other ways not controllable by the researcher. That is there may be confounding variables which are influencing the response variable and resulting in differences in the "treatment" groups.

Based on observational studies, it has been found that suicide rates in the military are higher than in the general population. Is there something about being in the military and its strict discipline that drives people to commit suicide? Maybe not. A potential confounder here is socioeconomic status. Many people who enlist in the military come from poor, unstable families and maybe this is why the rate is higher. The point is that in observational studies group membership is not under the control of an experimenter and treatment groups may differ in other ways besides the factor of interest.

If in a medical study involving human subjects, one group gets the new treatment and the other group no treatment at all, then the placebo effect is a potential confounding variable. That is, whether subjects got something or not could be related to the response and also whether they got something or not is certainly related to group membership. In fact it defines group membership. Thus the placebo effect is a potential confounder. The way to eliminate the placebo effect is for the group not getting the treatment to get a dummy treatment, or a placebo. Then both groups are receiving a "treatment."

## 1.7 Blinding

In medical studies knowledge of what treatment a subject is getting is a potential confounder. If I know I'm getting the real treatment as a compared to the dummy treatment, then I may act in ways that affect the response. Thus subjects in a medical study should not only be assigned at random to treatment groups, but should not have knowledge as to what treatment they are receiving. If this is true it is said that subjects are **blinded**.

A physician or evaluator's knowledge of who received what treatment may also be a confounder. The evaluator may subconsciously give better scores to those subjects in a group whose treatment he/she believes to be better. Thus often the physician or evaluator is blinded as well as the subject. This situation is then called **double blinding**.

## 1.8 Principles of Experiment Design

In this section we layout principles that researchers consider when designing experiments to eliminate/reduce the potential biasing effects of variables and/or increase the precision of the comparison of the treatments. The difference between **experimental** and **measurement** units is also discussed.

### 1.8.1 Four Principles of Experimental Design

a. **Randomization**

Randomization should be used to assign treatments to experimental units. As an example of randomization consider the microwaving oranges example from Section 1.4. Suppose that there are 40 oranges available for the study. Suppose these are labelled 1, 2, 3, ..., 40. The randomization could be conducted as follows. Write down the labels on 40 slips of paper and mix thoroughly. Pull out 20 slips of paper. These twenty slips of paper identify 20 oranges to be microwaved and the remaining slips identify the remaining twenty oranges that do not get microwaved. This approach is called a **completely randomized design**.

Randomly assigning experimental units to the treatment groups ensures that the effects of extraneous variables enter the experiment in a random fashion. Random assignment should, at least for larger group sizes, create groups that are balanced with regard to extraneous variables associated with the units. Thus the average size of the oranges in the microwave group should be about the same as the average size of the oranges in the non-microwaved group, thus preventing size from becoming a confounding variable. For experiments that use a small number of experimental units, complete randomization may not produce balanced groups and "blocking" (described below) should be used to achieve better balance.

Sometimes an experiment has to be performed over time and **experimental units are time slots**. In this case randomization should be used to balance out potential time effects. For example suppose that I wanted to know which of two types of softballs, A or B, I could hit further with my favorite softball bat. I buy 4 type A softballs, which I label A1, A2, A3, A4 and 4 type B softballs, which I label B1, B2, B3, and B4. Since I can only hit one ball at a time, this experiment will have to be done sequentially. To control for time effects, such as fatigue, I will randomize the order that the balls are pitched to me. In this way the effects of fatigue will sometimes disfavor type A and sometimes type B softballs. Note that the experimental units in this experiment are **time slots** corresponding to the times when balls are hit. The types of softball, A and B, are not assigned to the balls but are inherent characteristics of the balls. The types of softballs are however assigned to time slots.

The heating experiment (Section 1.2, part [i.]) involved the process of heating different pieces of meat of similar sizes at different heat settings over time. The experimental units are time slots for heating sessions. Randomization was used to select the particular treatments (type of meat and heat setting) over the 45 heating sessions. The purpose of the randomization was to balance out across the treatment groups any time effects such as changing environmental conditions associated with the sessions. An extraneous variable in this study was also the inability to obtain pieces of meat that were exactly the same size. Potential bias, that is some treatment groups having larger pieces or smaller pieces of meat than other groups, either intentional or unintentional, was prevented by randomly selecting a piece of meat of the randomly chosen type at each heating. In this way comparisons of heat settings for each type of meat should be roughly balanced regarding the size of the pieces. Additionally comparisons of types of meats at each heat setting should be roughly balanced regarding the size of the pieces.

b. **Blocking**

i. **Basics with Orange Juice Microwave Example**

Blocking refers to a statistical technique that attempts to eliminate potential confounding by grouping experimental units into **blocks** or groups with similar values on an extraneous variable and then randomly assigning treatments within each of the blocks, independently from block to block. The procedure may also result in a more precise comparison between the treatments on the response variable.

Reconsider the orange juice example. Size is an extraneous variable. One way of randomly assigning oranges to treatments is completely at random to the two treatments without regard to size as noted at the beginning of this section. In theory, randomization will, on average over many replications of the experiment, balance out the effect of extraneous variables. However for a particular experiment, and for small group sizes, the average size of the oranges for the microwave group may not be about the same as the average size of the oranges in the non-microwaved group. Thus any differences that are seen in the average amount of juice between the two groups could be due to random differences in size or other extraneous variables not fully balanced by the randomization, and not necessarily due to microwaving. Note that the randomization with blocking is totally different than the randomization conducted previously under the completely randomized design. There was no grouping of oranges first with the completely randomized design. In the block design there is grouping and then within each group randomization to treatments is carried out.

An alternative way of designing the experiment is as follows. Before any kind of randomization, sort the twenty oranges by size, such as weight, from largest to smallest and then “block” or group them into pairs. The first pair would be the two largest and would be about the same size, the 2nd pair would be the next two largest and would be about the same size, ..., until the last two which are the two smallest oranges, about the same size. Within each pair of oranges flip a coin(use randomization) to decide which of the two oranges gets microwaved and which does not. The result is two groups of oranges with a greater likelihood of balance on the size variable since the two groups were created by selecting from blocks where the size variable was about the same. This is called a **block design**. The analysis for such designs consists of comparing the amounts of juice within each block (pair) and then pooling these comparisons. Since in theory the comparison of the two treatments within each block is not affected by size (size of the two oranges is about the same within each block) the pooled comparison between the two comparisons may be more precise than in the completely randomized design. Greater precision is achieved by eliminating one of the extraneous variables.

Note that the randomization with blocking is totally different than the randomization conducted under the completely randomized design. There was no grouping of oranges before the random assignment of the twenty oranges to the two treatments under the completely randomized design. In the block design there is grouping of the oranges by pairs based on size and then within each group or pair randomization to treatments is carried out. This is called **restricted randomization**. In both designs two groups of oranges, one for microwaving and another for not micro-waving, will be identified.

ii. **Agricultural Origin of Blocking**

The idea of blocking first arose in agricultural settings. An agricultural researcher wants to compare the response variable yield for three different varieties of wheat. The varieties of wheat are denoted by A, B, and C. The experimental units are 12 plots of land arranged in four rows and three columns. The researcher could perform the experiment by assigning the three varieties of wheat to the 12 plots in a **completely randomized design**, with 4 plots per wheat variety. The result might be as in the following table:

| | | |
|---|---|---|
| A | B | C |
| C | B | C |
| B | C | A |
| A | A | B |

Extraneous variables include soil fertility, amount of light, and pH of the soil. Suppose that the layout of the plots is such that the plots in the various rows have similar soil. Thus it makes sense to have all three varieties used in each row and compare the varieties within each row. Thus each row of plots would be regarded as a block of plots and the three varieties would be randomly assigned within each block/row. The experimental arrangement with blocking might then look as follows:

| | | |
|---|---|---|
| B | A | C |
| B | C | A |
| A | C | B |
| C | A | B |

Notice here that all three treatments appear in each row or block. Data analysis would take this structure into account. The yields for the three varieties of wheat would be compared within each block where soil is the same and then the results pooled to draw an overall conclusion.

Note that blocking is a grouping of the experimental units **BEFORE** randomization is performed. There is also a grouping of the plots by treatment, here wheat variety, but this grouping occurs **AFTER** randomization.

iii. **Other Examples of Blocking**

Sometimes blocks are natural groupings of the experimental units. For example, several pairs of twins may be used in a study to compare the effects of two drugs. A block is one pair of twins. The individuals of the twin pair are randomly assigned to the two drugs with one twin getting drug A and one twin getting drug B. This is then repeated for several pair of twins.

**Meat Heating Example Revisited**. In experiments involving processes conducted over time blocking may be a grouping of time slots. In the original meat heating experiment described earlier (Section 1.2, part [i.]), it was assumed that the treatments (heat setting and type of meat) were randomly assigned completely at random to 45 time slots/heating sessions and that the experiment was conducted over one long time period. Because of time or other constraints suppose it was decided to conduct the experiment over 5 days with one full set of the 9 treatment combinations per day. Thus on each day there would be 9 heating sessions, one for each of the 9 treatments. On each day randomization would be used to decide which particular treatment of the 9 is applied first, which second, and so forth. This is an example of blocking, here a grouping of the experimental units, heating session, by Day, and independent randomisation done on different days. There are 5 blocks or days in the study.

**Baking Cookies**. A study was conducted to compare the diameters after baking of two types of Toll House pre-sliced cookies: chocolate chip and sugar. There were a total of 10 oven runs with the same oven temperature used for each run. At each oven run two cookies, one of each type was placed at random locations on a baking sheet on the middle shelf. The experimental units are the 20 cookies. The factor of interest is the type of cookie, chocolate chip and sugar. The response variable is diameter of a cookie (mm). The experimental units are grouped in the process of baking by oven run. So oven runs are the blocks with 10 blocks. The design is a block design. In a completely randomized design the 20 cookies would be baked one at a time over 20 oven runs, a rather inefficient design. Blocking allows comparison of the diameters of the two cookies at each oven run, where two cookies are baked under similar conditions.

**Chocolate and Endurance: Blocking by Reusing Subjects**. In some experiments subjects are given not one treatment, but all treatments over time, the order of which is random. Thus subjects are **reused** over time. Experimental units are the time slots at which the subjects are reused. This is another example of blocking where the grouping is of time slots by subjects. In the article "Chocolate Milk as a Post-Exercise Recovery Aid" (International Journal of Sports Nutrition, [2006], 16, 78-91) researchers compared three treatments for post-exercise recovery: chocolate milk, a fluid replacement drink (Gatorade), and a carbohydrate replacement drink. Subjects were 9 male, endurance-trained cyclists. Each subject performed an interval workout followed by 4 hours of recovery, and then an endurance trial to exhaustion on three separate days. On each day following the first exercise bout and 2 hours of recovery, subjects drank one of the randomly assigned drinks. One of the response variables was time to exhaustion (minutes) in the endurance trial. For each subject the three drinks are assigned to time slots corresponding to the exercise trials. This is a type of blocking in that the total of 27 time slots or experimental units are grouped by subject, three per subject. Each subject is **reused** on 3 different days. In a completely randomized version of this experiment the nine subjects would be assigned completely at random to the 3 drinks, 3 subjects per drink, with each subject consuming only one of the drinks.

### c. Direct Control

**Direct control** refers to control of an extraneous variable by using experimental units that all have the **same value** on the extraneous variable.

For example, in the tomato production study cited earlier, variety of tomato is an extraneous variable, but was directly controlled by using only one variety. Assuming that all tomato plants were planted in an open field, then amount of sunlight, another extraneous variable, is also directly controlled. In the orange juice example, in theory, size could be controlled by using oranges all of the same size.

Note that direct control can limit the scope of the conclusions. If only one tomato variety is used, then the conclusions pertain only to that variety. Blocking could be used in the tomato production study to extend the scope of the study by including more than one variety.

In the meat heating example an attempt was made to directly control the sizes of the pieces of meat across the meat types. The same burner was used to heat all 45 pieces of meat. The pieces of meat were placed at the same location in the pan, that is direct control of the effects of location on the response.

In the basketball free throw example the study was conducted in an indoor facility for all trials to control the effects of environmental conditions on the throwing accuracy. The same person did all of the shooting.

In the endurance trial for cyclists described in the last section all subjects were male, healthy, and highly-trained cyclists.

d. **Replication**

**Replication** of a treatment refers to a series (2 or more) of repetitions of the treatment to different independent experiment units. The experimental units could be individuals, groups of individuals, time slots, or runs of some process, such as a baking or cooking process. The multiple repetitions are referred to as **replicates**.

There would not be replication in the orange juice microwaving example if only one orange is used for each of the microwaving and no microwaving treatments. Because of extraneous variables and their effects on the response variable, replication is obviously important. A difference in orange juice using only one orange for each treatment could be the result of a difference in size of the two oranges or some other extraneous variable. Only with sufficient replication is it possible to conclude that there are "true" differences in amount of juice between microwaving and not microwaving.

In the longevity study involving mice the three different diets were randomly and independently assigned to the 60 mice, 20 per diet. There are 20 replicates or replicate mice for each diet.

To appreciate the benefits of replication consider an observational study to compare the heights of adult males and females. If we only sampled one male and one female at random (no replication) we might just by chance obtain a taller female and then make the wrong conclusion that females are taller than males. Obviously this would be wrong. If we replicate the "treatments" (male and female), that is sampled independently many males and many females, the "true" pattern would emerge.

Recall from an earlier discussion that differences in the values of a response variable between treatments must be judged in terms of the amount of difference that can be expected from the effects of extraneous variables alone, that is from experimental error. Replication, that is multiple observations on independent experimental units **within** each group allows us to measure variation in the response from the effects of extraneous variables alone. (Recall from your first course in statistics that a standard deviation measures variation of quantitative values in a group). This information in turn allows us to determine the expected difference **between** treatment groups that could be due to the effects of extraneous variables alone. The identification of the experimental units and subsequent replication is extremely important in the design of experiments since replication allows the experimenter to measure the extent of differences between treatments that could conceivably be due to error alone.

### 1.8.2 Experimental Units versus Measurement Units

Suppose that an experiment is to be conducted to compare two different recipes (A,B) on the response variable moisture of a cake. The baking process is as follows. Cake batter is prepared with one of the recipes chosen at random and then the batter is baked in an oven for a certain amount of time, with the resulting being a baked cake using that recipe. An experimental unit would refer to a cake or cake-making run, that is the process of making the cake batter and then baking the cake. Numerous extraneous variables associated with the cake-making run might influence the moisture of the resulting cake. Replication in this setting means multiple independent preparations of cake batter after random assignments of recipe with subsequent baking of cake batter in the oven. Suppose that the experimenter wants 5 observations on moisture for each recipe. Baking one cake with recipe A and one cake with B and then taking 5 observations on moisture from 5 locations on each cake does not constitute replication of a recipe. In fact in this case there would only be 1 experimental unit/cake for each recipe and statistical analysis could not be conducted to compare recipes. The 5 observations for each cake from the 5 locations constitute a **subsample** of the experimental unit, cake, and are called **pseudo-replicates**. The 5 locations on each cake are called **measurement units** as compared to the experimental unit, cake. Statistical analysis, such as the independent samples t test could be used to compare the two sets of 5 moisture levels made on each cake but the conclusion would pertain only to moisture levels for those two cakes, not the two recipes. To compare the two recipes the process of using a recipe, making the cake batter, and then baking the cake has to be repeated or replicated. That is there needs to be 5 cakes per recipe. Five moisture levels might still be obtained from each of the five cakes per recipe. These five moisture levels per cake could be averaged and then the 5 cake averages per recipe compared for the two recipes. Or the individual location moisture levels could be used in the analysis, with the analysis taking into account that these are sub-samples of the experimental units, not true replicates.

In many studies it it relatively easy to determine the number of replicates of a treatment because this corresponds to the number of observations on the response variable. However this is not always the case. In many animal health studies treatments are provided in the feed or water source for the animals in a pen and all animals in the pen eat or drink from that same source. Independent random assignments of treatments are done at the pen level. The experimental units are pens of animals. Individual animals within a pen are called measurement units. Replication of a treatment refers to different pens (at least 2) of animals being independently assigned to that treatment.

Suppose that two antibiotics (A,B) are being compared for control of an intestinal bacterium on average daily gain of pigs. The antibiotics are provided in the feed of the pen of pigs. Suppose that 8 pens, each with 3 pigs, are available for study. The 8 pens are randomly assigned to the two treatments, with 4 pens per treatment. There would be 12 pigs per treatment group but only 4 replicates (pens) per treatment group. The animals in the pen are called **measurement units**, not experimental units or treatment replicates. Some responses are measured at the pen or experimental unit level. For example, feed consumption, would be measured at the pen level, since individual feed consumption would normally not be measured. A comparison of feed consumption between the two anti-biotics would be compared using the four pen feed consumptions for the two groups. Other characteristics, such as weight gain, might be measured at the pig or measurement unit level. The three measurements per replicate/experimental unit represent a **subsample** of the treatment replicate/experimental unit. The analysis may proceed in different ways. If appropriate an average weight gain for a pen might be calculated and pen average weight then compared for the 4 values/replicates for each group. Alternatively weight gain might be analyzed individually or at the measurement unit level, 12 values per group, with pen as the experimental unit or replicate properly accommodated for in the statistical analysis. This type of analysis with **subsampling** will be explored in Chapter 10.

If in the pig study there was only one pen per anti-biotic group, 3 pigs per pen, then the study would lack replication. The 3 pigs in the pen are called pseudo-replicates. A statistical test, such as a t test could still be conducted to compare the 3 weight gains for the two pens. However any differences found would pertain to only those two pens and not in general to the two anti-biotics. Treatments have to be replicated (2 or more pens for each group) in order to draw conclusions about the two antibiotics.

In some studies **pseudo-replication manifests itself as repeated measures on an experimental unit**. Suppose that an experiment is designed to compare three different paper airplane designs on flight distance of planes made with the design. The experimental procedure is as follows. A paper airplane is constructed after identifying a randomly selected design. The plane is then thrown and the distance travelled (inches) is measured. This process is repeated for a total of 30 different paper airplanes, 10 replicates per treatment. The experimental units are the 30 different time slots corresponding to the construction of a paper airplane and subsequent flight. Suppose that the experimenter, wanting to increase precision of the comparisons of the designs, flew each plane three times rather than once, for a total of 90 flight distances, 30 distances per design. These second and third flights do not constitute additional replicates for each plane design but are pseudo-replicates, here repeated observations on the experimental unit. The additional flights would need to be taken into account in the analysis as pseudo-replicates rather than true additional replicates.

## 1.9 Examples of Some Standard Experimental Designs

The **design of an experiment** has two major components:

a. **Treatment Structure**

   This refers to the treatments and how they are formed. In some studies there is only one factor - the treatments are just the levels of the one factor. In a two factor study treatments usually refer to the the different combinations of the levels of the two factors. If in a two factor study treatments are formed by all combinations of the levels of the two treatments then the treatment structure is referred to as a **factorial treatment structure.**

b. **Design Structure**

   This refers to how the treatments are assigned to the experimental units, whether it is in a completely randomized fashion or whether there is some restriction on the randomization, such as in blocking. Design structure also refers to whether repeated measures are taken on units and whether there is sub-sampling.

Below are examples of some standard experimental designs that are studied in more detail in the text.

- **One Factor Completely Randomized Design: Freeze-Dried Strawberries and Serum Cholesterol Level**

  In the article “Freeze-Dried Strawberries Lower Serum Cholesterol and Lipid Peroxidation in Adults with Abdominal Adiposity and Elevated Serum Lipids” (Journal of Nutrition, [2014]) researchers compared four beverages consumed daily for change in serum cholesterol level over a 12 week period. The four beverages were 1) low dose (LD) beverage with powdered strawberries, 2) control beverage for LD beverage, 3) high dose (HD) beverage with powdered strawberries, and 4) control beverage for HD. The control beverages were matched for calories and total fiber. Sixty subjects with abdominal adiposity and elevated serum lipids were assigned completely at random to the four beverage treatments with 15 per group. Serum cholesterol level was measured before treatment and at the end of 12 weeks. The 15 values of the response, change in serum cholesterol level over the 12 week period,were compared for the four beverage groups. The study has one factor, that being beverage type. The experimental units are the 60 subjects assigned completely at random to the four beverage groups. There was no grouping or blocking of subjects before being assigned to the beverages.

- **Two Factor Completely Randomized Design: Rubber Band Strength Experiment**

Rebecca Aaron and Rebecca Redman in 2014 investigated the breaking strength of two brands of rubber bands (Staples, CLI) under three different temperature conditions (freezer, room, heated) for the rubber bands. They measured how far rubber bands stretched (cm) until breaking. The experiment has two factors, brand of rubber band and temperature exposure. The experiment was conducted as follows over time, one rubber band at a time. At a particular rubber band stretching session a brand and temperature condition was selected at random. A randomly selected rubber band of the particular brand was selected at random and then subjected to the selected temperature condition. The experimental units were time slots/stretching sessions. This process was repeated for a total of 30 rubberbands, 5 replicates rubberbands per combination of brand and temperature. The design is a completely randomized design since combinations of brand and heating sessions were assigned completely at random to the 30 sessions. There was no grouping or blocking of the stretching sessions before randomization. The randomization in this example could be implemented as follows. On 30 slips of paper write down each of the 6 treatment combinatons, 5 per treatment combination. At each of the 30 time slots/stretching sessions, draw a slip from the lot of slips - this will then determine the particular combination to use. This is an example of a two-factor completely randomized design, to be studied in more detail in Chapter 6.

- **One Factor Randomized Complete Block Design: Thawing Meat**

  In the article "Effect of Rapid Thawing on the Meat Quality Attributes of USDA Select Beef Strip Loin Steaks" (Journal of Food Science, [2011]: Vol. 76, Issue 2, pages S156-S162) researchers compared three methods for thawing frozen beef strip loin steaks on various quality characteristics of the meat. Each of 24 beef strip loins was cut into 3 steaks for a total of 72 steaks. The three steaks from each strip loin were randomly assigned to the three thawing methods. There was one conventional method (18 to 20 hrs, 4°C), and two rapid thawing methods (20 min, 20°C) or very fast (11 min, 39°C). The rapid thawing methods were conducted in a circulating water bath. The design has one factor of interest, thawing method. The 24 strip loins serve as blocks or groupings of the 72 steaks. The randomization of the thawing methods to steaks was done independently from block (loin) to block. The 72 steaks were not assigned completely at random to the three thawing methods. The response variables are the quality characteristics.

- **Two Factor Randomized Complete Block Design: Liquid Evaporation Experiment**

  Jill Yocum, Kristin Marucci, and Colleen Brookfield in 2010 investigated the amount of evaporation of three types of liquid(orange juice, rubbing alcohol, cola) in Tupperware containers with three different base areas (15, 25, 45 square inches). The experiment had two factors of interest: type of liquid and container base area. The response variable was the amount of liquid that evaporated (mL) after two days. There were 5 replications of each treatment combination of liquid and base area with one complete replication (all 9 treatments) being tested every 2 days. The experiment used the same 9 containers, 3 for each of the three sizes, every 2 days. At the beginning of each testing cycles of 2 days the 9 treatments (combinations of liquid type and container type) were randomly assigned to 9 locations on the floor in a 3 x 3 grid. For the selected container type a particular container was randomly selected. The experimental units are the nine floor locations with containers. There are 5 blocks in the experiment corresponding to the 5 2-day testing cycles or 5 sets of 9 experimental units per testing cycle.

- **Two Factor Split Plot Design: Baking Experiment**

  The following experiment is an example of a split plot design. The split plot design has two factors but there are two types of experimental units, one for each factor.

  John Szarka and Zamda Lumbi in 2004 investigated the effects of type of flour (white, wheat, bread) and length of time in the oven (5, 10, 15 minutes) on the change in height of dough after baking. Three rolls of dough were made with each type of flour for a total of 9 rolls. Each roll was made using the same ingredients except for the type of flour. Each roll was divided/split into 3 equal parts and the 3 parts put into the oven. One part was baked for 5 minutes, another part at 10 minutes, and another for 15 minutes. One run of the oven involved one roll with its 3 parts. There were 9 oven runs altogether with the particular flour for the roll used at an oven run being selected at random. Random assignment determined the amount of time each of the three parts stayed in the oven. The two factors in the study were type of flour used for the roll/oven run and amount of time in the oven for parts.

  The randomization was conducted in two stages. At one stage randomization was used to assign type of flour to roll/oven run so roll at an oven run is the experimental unit for type of flour. At a second stage randomization was used to determine the number of minutes each of the parts for a given roll stayed in the oven. So the experimental units for the number of minutes factor are the parts of the roll or 1/3 splits of the roll.

  The rolls are also called **whole units** while the parts are also called **split units**, since the rolls result from a splitting of the whole units rolls. Hence the name **split plot design**. Type of flour is called the **whole unit factor** and amount of time in the oven is called the **split unit factor**.

  The rolls or whole units are also blocks since each roll is a grouping of the parts of 27 rolls.

  The data from this example is analyzed in Chapter 9.

- **One Factor Completely Randomized Design with Sub-Sampling: Feeding Fish**

  Suppose that an experiment is conducted to compare four different diets on growth for a certain species of fish. It is difficult to feed fish individually so the feeding is done by depositing the food in tanks and then the fish within a tank feed on the food. There are a total of 120 fish available altogether for the study. The fish are weighed individually before the start of the study and then randomly assigned to 12 tanks, with 10 fish per tank. The four diets are assigned at random to the 12 tanks, with 3 tanks per diet. After several weeks the fish from all tanks are removed again and weighed. Weight gain is calculated for each fish. There is one factor, diet. The experimental units for diet are tanks of fish, not individual fish, since the diets are assigned to tanks. Thus there are 3 replications per diet, not 30. The 10 fish in each tank are sub-samples of the experimental units/tanks and would have to be taken into account in the analysis. The experimental units are tanks which were assigned completely at random to the diets. Measurement units are the individual fish. Thus the design is a one-factor completely randomized design with sub-sampling.

- **Two Factor Split Plot/Repeated Measures Design: Obsessive Compulsive Disorder**

  Two different medications and a placebo were compared for efficacy in the treatment of Obsessive Compulsive Disorder (OCD). The level of the patients' illness was measured using the Yale Brown Obsessive Compulsive Scale (YBOCS) at baseline before treatment and at weeks 1, 2, 3, 4, 6, 8, 10, and 12 weeks. The YBOCS score ranges from 0 to 40 with higher values indicating more extreme cases. Thirty OCD patients were assigned completely at random to the 3 treatments, 10 to each of the two medications and the placebo. Researchers were interested in the time profile of each treatment and how the three treatments compared over time.

  The design is similar to a split plot design. There are two factors with two types of experimental units. One factor, the whole unit factor, is treatment for OCD. The experimental units or whole units for the OCD treatments are the 30 OCD patients. The whole units are assigned at random to the levels of the whole unit factor. The other split unit factor is number of weeks after start of medication, whose levels are baseline, 1, 2, 3, 4, 6, 8, 10, and 12 weeks. The split units are the time points or occasions in time that correspond to the levels of the split unit factor when the YBOCS scores are measured. One can think of the split units resulting from a splitting of the large time frame when the experiment was conducted. However unlike the usual split plot design the split units are not assigned at random to the levels of the split unit factor, but are inherent characteristics of the time points. Thus with this design there is randomization of whole units to levels of the whole unit factor but no randomization of the split units to the levels of the split unit factor.

The term “repeated measures” refers to the fact that the whole units, here subjects, are repeatedly measured on the response YBOCS on the different time points or occasions.

## 1.10 Scope of the Conclusions of an Experiment

Experimental units in an experiment should ideally be selected at random from some relevant population and then assigned at random to treatments in order to be able to draw valid conclusions about the population. However random selection from some population will usually mean a fair amount of variation in the subjects and perhaps a large number of extraneous variables. This in turn, can mean imprecise comparisons and thus not being able to say much at all with regard to the comparison of the treatments. Blocking can be used to minimize the problem, that is group the subjects that are similar and then make comparisons within each block. Thus we can have our cake and eat it too!

The subjects in experiments involving humans are usually not selected at random from some population but are volunteers who have consented to being part of the study. Volunteers are necessary because of the nature of experimentation in which people are treated in some kind of way. While volunteers can be assigned at random to treatment groups, generalizing to some population may require judgements from people who are familiar with the subject area. For example, in a study which uses college students can we generalize to the general population?

# Problems for Chapter 1

1.1* Osteoarthritis of the joints affects a large number of senior citizens. One study looked at the perceived benefits of arthroscopic surgery for osteoarthritis by giving some patients a real knee operation, while others underwent a sham surgery. Patients were assigned at random to either receive arthroscopic surgery or a sham surgery. One response variable was the speed of walking after surgery.

a. Is this study an experiment or an observational study? Explain.

b. What is the factor? What is the response variable?

c. What is the purpose of one group receiving a sham surgery?

1.2* Health experts suspect that re-circulated air in aircraft carries more germs and causes more colds than on aircraft that pumps in fresh air. An article in the New England Journal of Medicine reported the results of questionnaires given to 1100 passengers leaving the San Francisco area and traveling to Denver between January and April 1999. Some of the passengers had been aboard aircraft which used re-circulated air and others aboard aircraft which circulated fresh air. A week after their flights, 21% of the fresh-air passengers and 19% of the re-circulated air passengers reporting having a cold.

a. What are the conditions in this study? Are they controlled or simply observed? Explain.

b. What is the response variable?

c. The researchers noted that the incidence of colds in both groups was higher than that of non-travelers which is about 3%. Give some possible reasons for this difference besides cabin air.

1.3* Proponents of massage therapy believe that massaging some or all parts of the body affect psychological and physical health. In designing an experiment involving children with cancer, one group received massages from their parents at bedtime, while another group received no such massage. One critic claimed that any benefits might be due to the "attention" being given to the kids in the massage group and not the massage itself. How should the experiment be conducted to control for the "attention" effect?

1.4* In a study of 4600 young people aged 12-19 females with body piercings (other than the ears) were 2 1/2 times more likely to have sex and 2 1/2 more likely to have smoked than those who did not have body piercings. Boys had similarly high risks. [3].

a. What is the factor of interest in this study? What is (are) the response variable(s)?

b. Is this study an experiment or an observational study? Explain.

c. Can we conclude that body piercing leads to more sexual activity and more smoking? Explain.

1.5* A recent article in the Lancet medical journal reported the results of a study to determine if the implantation of a patient's own bone marrow stem cells into their leg muscles could create new vessels. If successful this could eliminate pain from bad circulation due to clogged arteries and help prevent gangrene or amputations. Twenty subjects, in whom both legs were starved of blood flow, participated in the study. They had their bone marrow stem cells injected into one leg, randomly chosen, and regular blood injected into the other leg. The legs that got the stem cells had more improvement than the others on a test comparing blood pressure in the ankle with that in the arm before and after the treatment. Similar results were seen in a second circulation test that measured differences in oxygen inside and outside tissues.

a. What are the treatments in this experiment?

b. What are the response variables?

c. What was the purpose of randomization?

d. Was there any blocking in the experiment? Explain.

1.6* A survey of 232 elderly patients who had recently undergone heart surgery was undertaken. The patients were asked, among other items, whether or not they derived strength or comfort from religion. Patients were followed for a number of years. Those patients who said they derived strength or comfort from religion lived longer than those who said they did not. ([3])

a. What is the factor of interest in this study? What is the response variable?

b. Is this study an experiment or an observational study? Explain.

c. Name some potential confounding variables.

1.7* An educational researcher is interested in comparing two different methods of memorizing material to see if they differ with regard to retention. Thirty subjects are available for the study. Explain how blocking might be used in this study.

1.8* An experiment in Dean and Voss ([6], page 62) compares balloons of different colors in terms of amount of time needed to blow them up. One individual blew up all 20 balloons of 4 different colors, 5 balloons of each color.

a. What are the treatments in this experiment?

b. What are the experimental units?

c. Discuss how randomization would be used in this study and the purpose of the randomization.

d. Is there direct control of any extraneous variables in the study? Explain.

1.9* In the article "Bioevaluation of garlic on growth, haemotological and serum characteristics of growing pigs"(*African Journal of Biotechnology [2013]: Vol. 12(25), pp. 4039-4043*) researchers compared average daily weight gains(kg/day) under three dietary treatments for pigs: 0, 100, and 200 g of sun dried garlic powder per 100 kg of feed. Eighteen grower pigs aged 70 days were randomly assigned to 6 pens with 3 pigs per pen. The three pigs in a pen consumed feed from the same feeding trough. The three treatments were randomly assigned to the 6 pens with 2 pens per treatment.

What are the experimental units in this study? Explain.

1.10* In the article "Randomized Trial of Exercise Therapy in Women Treated for Breast Cancer"(*Journal of Clinical Oncology [2007]: Vol. 25(13), pp. 1713-1721*) researchers studied the effects of aerobic exercise therapy on quality of life (QoL) as measured by a score on the Functional Assessment of Cancer Therapy-General (FACT-G). Scores on the FACT-G range from 0 to 100 with larger values indicating a higher quality of life. The abstract of the article reported that "a total of 108 women who had been treated for breast cancer 12 to 36 months previously were randomly assigned to supervised aerobic exercise therapy ($n = 34$), exercise-placebo (body conditioning, $n = 36$), or usual care ($n = 38$). Women in the aerobic exercise therapy group met on a one-to-one basis with an exercise specialist, 3 times per week, eight weeks total, for moderate-intensity aerobic exercise. Women in the exercise-placebo group also met three times per week, each session 50 minutes long, for eight weeks, for light-intensity body conditioning/stretching exercises. The usual-care group continued with their lives as usual.

a. This study is an experiment. Explain.

b. What are the experimental units?

c. Is blinding of subjects possible in this study? Explain.

1.11* In the article "Quality of Life and Functional Health Status of Long-Term Meditators"(*Evidence-Based Complementary and Alternative Medicine vol. 2012, Article ID 350674, 9 pages 2012. doi:10.1155/2012/350674*) researchers compared the quality of life and functional health of a sample of 334 long-term meditators to that of population norms for Australia. Participants completed the Medical Outcomes Study Short Form 36 (MOS SF-36).

Is this study an experiment or an observational study? Explain.

1.12* The American Statistical Association holds an annual poster and project competition for students from grades K-12. Winners receive a monetary award and a plaque. One of the winners in the 2013 competition conducted an experiment to answer the question: Do dryer balls reduce drying time? The student conducted the experiment in response to an ad that claimed that dryer balls "reduce drying time by up to 25%." The student randomly assigned the next 40 of his family's wash loads to either be dried with dryer balls added to the dryer or not. Only one washer and dryer was used. The student weighed each load prior to drying and recorded how long it took the load to dry (in minutes) using a stop watch. The dryer has a sensor that detects when the clothes are dry.

a. What are the treatments in this experiment?

b. What are the experimental units?

c. Give some extraneous variables and explain how they are controlled.

1.13 John Ellis and John Tran in 2013 studied the effect of composition of pasta (white, whole grain) and length of pasta (1.25, 6, 12 inches) on the change in mass (grams) of the pasta after boiling the pasta for 10 minutes in 1 liter of water. There were 5 replications of each treatment combination with 1 replication being conducted on each of 5 days. The 6 boilings of pasta on each day were conducted one a time using the same boiling pot and stove burner.

a. What are the two factors in the study?

b. What are the treatments in the study?

c. What are the experimental units?

d. Is there blocking in this study? Explain.

e. Explain how the necessary randomization could be physically implemented

1.14 Nick Granered, Ryan Saba, and Charlie Watt investigated the effect of type of cookie (Chips Ahoy Chocolate Chip, Oreo's, Nutter Butters) and type of liquid (milk, orange juice, water) in which the cookie was dipped on the percentage increase in the weight (g) of the cookie. Forty-five cookies were tested at the same time using 45 cups, 5 for each combination of type of cookie and type of liquid. Type of cookie and liquid type were assigned completely at random to the 45 cups.

a. What are the two factors in the study?

b. What are the treatments in the study?

c. What are the experimental units?

d. Is there blocking in this study? Explain.

e. Explain how the necessary randomization could be physically implemented.

1.15 Emily Shrader, Ashley Sawyer, and Lisa Kleinschmidt in 2009 investigated the effects of type of cup (styrofoam, paper) and type of liquid (water, water with lemon, water with salt) on the temperature of the liquid 10 minutes after it had been heated to 160 degrees Fahrenheit. The experiment was conducted as follows. A type of liquid was randomly selected. Twenty ounces of the liquid was heated to 160 degrees in a pot. The twenty ounces of liquid was then poured out, half into a styrofoam cup and the other half into a paper cup. After 10 minutes the temperature of the liquid in the two cups was measured. The procedure was repeated eleven other times, resulting in 4 replications per combination of type of liquid and type of cup. This is an example of a split plot design.

a. What are the two factors in the study?

b. What are the whole units? split units?

c. Is there blocking in this study? Explain.

d. Explain how the necessary randomization could be physically implemented.

1.16 This example is based on an exercise in McClave and Sincich ([19], page 480) from the article "Vulnerability of Canada Geese to Taxidermy-Mounted Decoys" (Journal of Wildlife Management, 59(3):474-477) A study compared the effectiveness of three decoy types - taxidermy-mounted decoys, plastic shell decoys, and full-bodied plastic decoys on the attraction of Canadian geese to sunken pit blinds. Three pit blinds in three different locations were used. Each pit blind was used on several days with all three of the decoy types being used. The response variable was the mean daily percentage of goose flocks attracted to a blind over the days when the decoy type was used.

a. The experiment uses blocks. What are the blocks?

b. What is the factor of interest?

c. Explain what kind of randomization would be used in the experiment.

1.17 A student project examined the effects that different amounts of salt had on the boiling temperature of 3 quarts of water in a pot. The student found in his literature search an example where adding salt to water increased the boiling temperature and they wanted to see if they would obtain similar results. Three levels of salt were used (0, 1, 2 tablespoons) in the experiment. Thirty different trials were conducted with 10 trials per amount of salt. The student randomly selected a salt level and then added that amount to three quarts of water. The pot was then placed on the stove and heated. The amount of time to reach boiling was based on visual inspection by the student when there was a "consistent boil." Discuss any possible bias in the measurement process.

1.18 Reconsider the meat heating experiment from Section 1.2. Suppose the original proposed design was to heat 5 pieces in the pan at one time, all pieces being of the same type and using the same heat setting. This was to be done for the 9 combinations of type of meat and heat setting. Thus the 45 values on the response, amount of time, would be obtained with only 9 heating runs.

a. What are the experimental units for this design? How many experimental units are there?

b. What are the measurement units? How many measurement units are there?

c. How many replicates are there of each treatment combination of type of meat and heat setting?

d. Is this a valid design? Explain.

1.19 An experiment was conducted to compare three different watering regimens on the growth of Marigold plants. Forty-five plants were purchased at a local nursery and randomly assigned to 15 pots, with 3 plants per pot. All 15 pots used Miracle Gro Soil. The 15 pots were randomly assigned to the three watering regimens, with 5 pots per regimen. The three regimens were: 1) pot receives 1/2 cup of water twice per cup, 2) pot receives 1 cup of water twice per week, 3) pot receives 1 1/2 cups of water twice per week. The experiment was carried out over a period of 4 weeks. The response variable was growth of a plant during the period (height at end of 4 weeks minus height at beginning of study).

a. What are the experimental units in this study? How many are there?

b. Give two extraneous variables associated with the experimental units?

c. How many measurement units are there? How many are there?

d. Is there pseudo-replication in this study? Explain.

1.20 This example is based on an experiment described in the article “Music and its effect on sports” (www.all-science-fair-projects.com) A group 10 students, 5 males and 5 females, aged 16, physically fit, with no health problems, were used to compare distance run (no unit of measurement given) on a treadmill for a 10 minute period. Each student ran on the treadmill on two different days, on the 1st day without listing to music and then on the 2nd day listening to fast paced songs on the MP3. On each day each student had the opportunity to warm up for 5 minutes and an additional 5 minutes to familiarize himself/herself with the treadmill. The results showed that students ran longer distances on average when the music was played.

a. What are the treatments in this study?

b. What is the response variable?

c. Name two extraneous variables that are directly controlled in this study.

d. This is a block design. Explain.

e. Note that there was no randomization of the treatments to the two days when the students were tested. Because of this, is there potential for bias in the comparison of the amounts of time for the two treatments? Explain.

f. How would this experiment have been conduced using a completely randomized design instead of a block design?

# Chapter 2

# Basic Concepts and the One Sample Problem

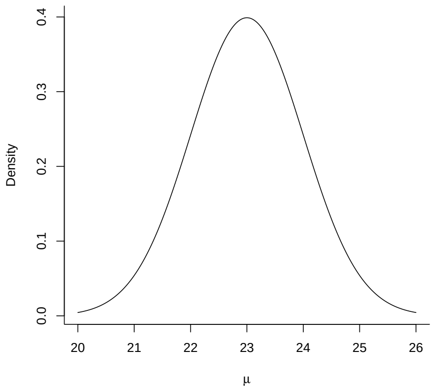
This chapter reviews the basic statistical concepts associated with inferences about a single population based on a random sample selected from that population. The notions of estimation of population parameters, standard errors of estimators, and hypothesis testing are discussed.

## 2.1 Population versus Sample

Most statistical studies are concerned with the drawing of conclusions about **populations** based on **samples** selected from those populations. In this chapter we will concentrate on the one sample/one population case. Inferences assume that the samples are randomly selected from the population of interest. Often in practice samples are not randomly selected and thus judgement must be exercised to determine if conclusions reached can be validly applied to some population. Suppose $y$ represents some quantitative variable in the population with mean $\mu$ and variance $\sigma^2$. The mean is also referred to as the expected value of $y$, denoted by $\mu = E[y]$. The variance is defined as $\sigma^2 = E[(y-\mu)^2]$, the expected value of the square of the difference between $y$ and $\mu$. The standard deviation of $y$ is defined to be $\sigma = \sqrt{\sigma^2}$. The mean $\mu$, variance $\sigma^2$, and standard deviation $\sigma$ of $y$ in the population are examples of population **parameters**.

The normal population is a type of population that should be familiar to the reader. Much of the theory of the analysis of variance is based on the assumption of normal populations. The histogram of a variable $y$ in a normal population is symmetric, bell-shaped with center at $\mu$. Figure 2.1 provides a histogram of a basic normal population with mean of $y$ equal to $\mu = E[y] = 23$ and standard deviation equal to $\sigma = 1$. The vertical axis (density) has been scaled so that total area under the curve is equal to 1 and areas of regions under the curve represent proportions in the population. The normal curve is an example of a **density curve**. Recall that the area under a normal curve above the interval

Figure 2.1: Normal Curve: $\mu = 23$, $\sigma = 1$



$(\mu - \sigma, \mu + \sigma)$ is about 0.68, above $(\mu - 2\sigma, \mu + 2\sigma)$ is about 0.95, and above $(\mu - 3\sigma, \mu + 3\sigma)$ about 1. Thus in this example about 68 percent of the values of $y$ in this normal population are between 22 and 24. The areas under the curve can also be regarded as probabilities. Suppose one value is selected at random from this population. The variable $y$ is then called a **random variable** and the curve is then referred to as the probability distribution for $y$. Areas under the curve are then regarded as probabilities about $y$. Thus we can say before sampling that there is about a 95% probability that the selection will result in a value of $y$ between 21 and 25.

The *standard normal* population is that normal population with mean $\mu = 0$ and $\sigma = 1$. Fact 2.1 shows how the standard normal variable is related to an arbitrary normal variable.

**Fact 2.1** *If $y$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$, then $z = (y - \mu)/\sigma$ has a standard normal distribution.*

Fact 2.1 generalizes. Subtracting from a normal random variable its mean and dividing by its standard deviation results in a variable that has a standard normal distribution. This general result will be applied in the next section. Table $A$.1 in the Appendix provides right tail areas or probabilities for the standard normal distribution.

## 2.2 Sample Mean and Standard Deviation

In practice the population mean $\mu$ and standard deviation $\sigma$ are unknown and interest is usually in estimating these parameters.

Let $y_1, y_2, ..., y_n$ represent a random sample of size $n$ from a population $y$ with mean $\mu$ and variance $\sigma^2$ (standard deviation $\sigma$). Statistically $y_1, y_2, ..., y_n$

Table 2.1: Sample Mean and Standard Deviation Calculation

| Student $i$ | Weight $y_i$ | $(y_i - \overline{y})$ | $(y_i - \overline{y})^2$ |
|---|---|---|---|
| 1 | 180.1 | 17.9 | 321.8 |
| 2 | 157.7 | -4.5 | 20.1 |
| 3 | 142.4 | -19.8 | 393.1 |
| 4 | 155.5 | -6.6 | 44.2 |
| 5 | 153.8 | -8.4 | 70.2 |
| 6 | 131.1 | -31.1 | 969.1 |
| 7 | 194.4 | 32.2 | 1034.5 |
| 8 | 157.3 | -4.9 | 24.1 |
| 9 | 181.3 | 19.1 | 363.3 |
| 10 | 168.4 | 6.2 | 38.7 |
| Sum | 1621.9 | 0 | 3279.1 |

represent random values which are independent, identically distributed as the population, each with mean $\mu$ and variance $\sigma^2$.

The sample mean, denoted by $\overline{y}$, is an estimate of the population mean $\mu$ and the sample variance, denoted by $s^2$, is an estimate of the population variance $\sigma^2$ (sample standard deviation $s = \sqrt{s^2}$ is an estimate of the population standard deviation $\sigma$). The sample mean is defined as $\overline{y} = (\sum_{i=1}^{n} y_i)/n$. The sample variance, $s^2$, is defined as $s^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)$ with the sample standard deviation $s = \sqrt{s^2}$.

**Example 2.1** *Suppose that the weight of male students at a university in a given semester is a normal random variable with population mean* $\mu = 170$ *pounds and population variance* $\sigma^2 = 225$ *(population standard deviation* $\sigma = \sqrt{225} = 15$ *pounds). In practice neither the population mean nor the population variance (standard deviation) would typically be known. Suppose that a random sample of* $n = 10$ *students is selected. Table 2.1 gives the* 10 *weights, the deviations of the weights from the mean* $(y_i - \overline{y})$ *and the squares of the deviations.*

The sample mean weights of the 10 students is $\overline{y} = 1621.9/10 = 162.2$ and the sample variance is $s^2 = (3279.1/(10-1) = 364.34$ with sample standard deviation $s = \sqrt{364.34} = 19.1$. Notice that the sample mean and variance (standard deviation) for this sample are not the same as the mean and variance (standard deviation) for the population of students but differ due to **sampling error**.

## 2.3 Sampling Distribution of the Sample Mean

A sample mean $\overline{y}$ is unknown before the sample is selected and is a random variable with its own probability distribution, mean, and standard deviation.

In Example 2.1 the sample mean for the particular sample selected was 162.2 pounds. If another sample is selected from the same population the sample mean would be a different value. The probability distribution of the sample mean $\overline{y}$ regarded as a random variable is called the **sampling distribution of the mean**. Properties of the sampling distribution of the mean are reviewed below.

1. The mean of the sampling distribution of $\overline{y}$ is $\mu_{\overline{y}} = E[\overline{y}] = \mu$. Note this says that the average of sample means under repeated sampling is equal to $\mu$. In practice repeated sampling is NOT done. A researcher will have only one sample mean and that one sample mean will NOT be the same as $\mu$.

2. The standard deviation of the sampling distribution of $\overline{y}$ is $\sigma_{\overline{y}} = \sqrt{E[(\overline{y} - \mu_{\overline{y}})^2]} = \sqrt{E[(\overline{y} - \mu)^2]} = \sigma/\sqrt{n}$. The quantity $\sigma_{\overline{y}}$ gives a crude measure of how far "off" an observed sample mean is away from the unknown population mean $\mu$. Note that this number is not very useful in practice, however, since $\sigma$ is unknown. However we could estimate $\sigma$ with $s$, the sample standard deviation. See Example 2.2.

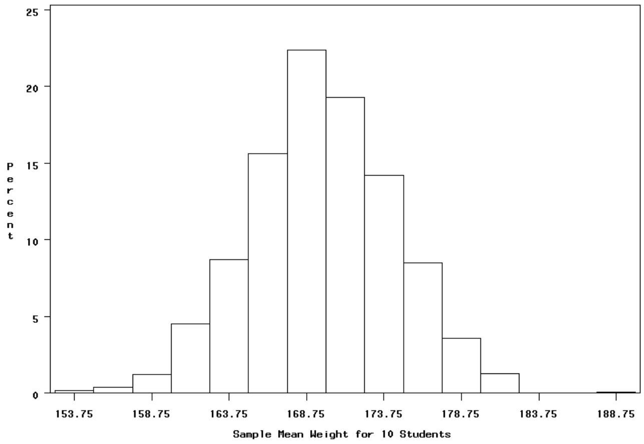
3. If the population is normally distributed then the sampling distribution of $\overline{y}$ is exactly normally distributed. If the population is not normally distributed but the sample size is sufficiently large, then by the **Central Limit Theorem** the sampling distribution of $\overline{y}$ is approximately normally distributed.

The properties listed above apply to repeated sampling from the same population. A computer can be programmed to illustrate the properties. The following example illustrates.

**Example 2.2** *A SAS program was written to simulate the random sampling of 1000 samples of male university students, each of size* $n = 10$ *from the same population used in Example 2.1. For each sample of 10 weights obtained the sample mean was calculated. Figure 2.2 gives a histogram for the 1000 sample means. Note the bell shaped appearance of the histogram. Also the mean of the 1000 sample means is* 169.8 *which closely approximates the theoretical value of* $\mu_{\overline{y}} = \mu = 170$*, in this example. The standard deviation of the 1000 sample means is* 4.71 *which is close to the theoretical value of* $\sigma_{\overline{y}} = \sigma/\sqrt{n} = 15/\sqrt{10} = 4.74$*.*

In practice usually only one sample is selected. That one sample will give a particular sample mean which provides an estimate of the unknown population mean $\mu$. The sample will also provide an observed sample standard deviation, $s$. This observed sample standard deviation is used to estimate $\sigma$, which then provides an estimate of $\sigma_{\overline{y}}$. This estimate of $\sigma_{\overline{y}}$, called the **standard error of the mean**, is $s_{\overline{y}} = s/\sqrt{n}$. The standard error of the mean provides a rough idea of the error associated with the one observed sample mean as an estimate of the unknown population mean.

Figure 2.2: Histogram of 1000 Sample Mean Weights



By the properties, if the population is normally distributed or if the sample size is large then the sample mean $\overline{y}$ is normally distributed (or approximately normal) with mean $\mu_{\overline{y}} = \mu$ and $\sigma_{\overline{y}} = \sigma/\sqrt{n}$. Hence by Fact 2.1 the standardized sample mean,

$$\frac{(\overline{y} - \mu)}{\sigma/\sqrt{n}}$$

has a standard normal sampling distribution. So for example if a normal variable/population has a mean $\mu = 200$ and standard deviation $\sigma = 24$ then if we repeatedly sample from this population samples of size $n = 9$ then $\overline{y}$ has a normal distribution with mean $\mu_{\overline{y}} = 200$ and standard deviation $\sigma_{\overline{y}} = 24/\sqrt{9} = 8$. Also

$$\frac{(\overline{y} - 200)}{8}$$

has a standard normal sampling distribution.

If in the standardized sample mean we replace $\sigma$ in $\frac{(\overline{y}-\mu)}{\sigma/\sqrt{n}}$ with the sample standard deviation, $s$, then the probability distribution of the resulting standardized sample mean $\frac{(\overline{y}-\mu)}{s/\sqrt{n}}$no longer has the standard normal distribution. It has what is called the **Student's t** or simply the **t** probability/sampling distribution. The $t$ distribution will arise within the context of a confidence interval for a single unknown population mean in the next section and in many other contexts in this book. It is an important probability distribution in analysis of variance.

## 2.4 Confidence Interval for a Normal Population Mean

In practice a population mean $\mu$ is unknown and must be estimated based on a sample. A **confidence interval** estimate for a population mean is an interval of plausible values for $\mu$. Associated with the interval is a "**confidence level**" which indicates how confident we are that the interval actually contains $\mu$. Typical confidence levels are 90%, 95%, and 99%. The **one sample "t" interval** for a population mean is based on the Student's $t$ distribution.

**Fact 2.2** *Suppose a random sample $y_1, y_2, ..., y_n$ of size $n$ is selected from a normal population with mean $\mu$ and standard deviation $\sigma$. Let $\overline{y}$ and $s$ be the sample mean and sample standard deviation, respectively. Then the standardized sample mean*

$$\frac{(\overline{y} - \mu)}{s/\sqrt{n}}$$

*has a probability distribution called the* **t** *distribution with* **degrees of freedom (df)** $\nu = n - 1$.

The $t$ distribution is symmetric, bell shaped with a mean of 0, that is

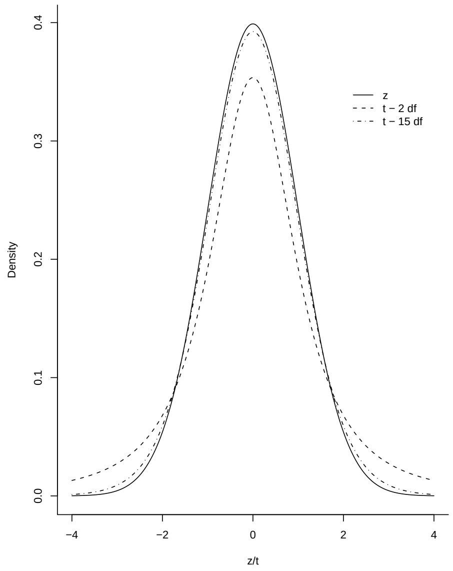
$$E[\frac{(\overline{y} - \mu)}{s/\sqrt{n}}] = 0$$

and standard deviation of $\frac{\nu}{\nu-2}$. The $t$ distribution is a family of distributions indexed by the parameter $\nu$. The distributions are all bell shaped, symmetric, centered at 0, which makes them similar to the standard normal distribution. Unlike the standard normal distribution which has a standard deviation of 1, the standard deviation of the $t$ distribution depends upon the parameter $\nu$, which depends on sample size. Figure 2.3 gives a picture of two $t$ distributions compared to the standard normal distribution.

Table A.2 gives the upper $\alpha$ probability points, denoted by $t_{\alpha;\nu}$, for certain values of $\alpha$ and degrees of freedom (df), $\nu$. The area under the $t$ distribution to the right of $t_{\alpha;\nu}$ is $\alpha$. Thus for example, the upper $\alpha = 0.05$ probability point from a $t$ distribution with $\nu = 2$ degrees of freedom is 2.920. Note that the area under this t curve to the left of 2.920 would be 0.95. The area under the t-curve between $-2.920$ and 2.920 would be 0.90.

Suppose a random sample of size $n = 15$ is to be selected from a normal population with unknown population mean $\mu$ and population standard deviation $\sigma$. The sample mean $\overline{y}$ and sample standard deviation $s$ will be used to summarize the sample. The goal is to estimate the unknown population mean $\mu$. A 95% **confidence interval** for $\mu$ is an interval of plausible values for $\mu$. The 95% "confidence level" refers to how confident we are that the value of $\mu$ takes on one of the values in the interval. A brief derivation of such an interval follows.

Figure 2.3: Example of t distributions; $\nu = 2, 15$



By Fact 2.2, $\frac{(\overline{Y}-\mu)}{s/\sqrt{n}}$ has the $t$ distribution with $\nu = 15-1$ degrees of freedom. Thus using Appendix Table A.2,

$$P[-2.145 < \frac{(\overline{y}-\mu)}{s/\sqrt{n}} < 2.145] = 0.95,$$

Note that 0.95 is a middle area. The area to the right of 2.145 under the t-curve is 0.025. So the appropriate probability point from Table A.2 is the upper 0.025 probability point, 2.145, not the upper 0.05 probability point.

After some algebra the probability statement can be written as

$$P[\overline{y} - 2.145(s/\sqrt{n}) < \mu < \overline{y} + 2.145(s/\sqrt{n})] = 0.95$$

The interval within the brackets is a random interval because it has random endpoints. The statement says that, before sampling, there is a 95% chance of this random interval containing $\mu$. After the sampling has occurred, the values of $\overline{y}$ and $s$ are known. They can then be substituted into the formula to obtain an actual interval. This calculated interval is then called a 95% confidence interval for $\mu$.

**Example 2.3** *Suppose that amount of money spent by students on textbooks in a given semester are normally distributed with some (unknown) mean $\mu$ and standard deviation $\sigma$. Suppose a random sample of $n = 15$ students is selected. The sample mean amount spent by those 15 students was $\overline{y} = \$375.32$ with* *sample standard deviation* $s = \$27.18$. *The standard error of the sample mean,* \$375.32*, as an estimate of* $\mu$ *is thus*

$$\frac{s}{\sqrt{n}} = \frac{27.18}{\sqrt{15}} = \$7.02$$

*The standard error of* \$7.02 *gives a crude idea of how far away the sample mean of* \$375.32 *is from the unknown population mean amount spent. The* 95% **"margin of error"** *for the estimate of* \$375.32 *is* 2.145(\$7.02) = \$15.06*, where* 2.145 *is the upper* 0.025 *probability point from a t distribution with* 14 *degrees of freedom. The* 95% *confidence interval for the unknown mean amount of money spent on textbooks is*

$$375.32 - 15.06 < \mu < 375.32 + 15.06$$

*or*

$$\$360.26 < \mu < \$390.38$$

The general form of the endpoints of a **one sample t confidence interval** for a normal population mean with confidence level of $100(1-\alpha)\%$ is

$$\overline{y} \pm t_{\alpha/2;n-1}(s/\sqrt{n})$$

where $t_{\alpha/2;n-1}$ is the upper $\alpha/2$ probability point from the $t$ distribution with $\nu = n - 1$ degrees of freedom. Table A.2 is entered with $\alpha/2$ to obtain the correct probability point.

A more general form of confidence intervals that will be seen in this text is

$$point\ estimate \pm margin\ of\ error$$

or

$$point\ estimate \pm multiplier\ *\ standard\ error\ of\ point\ estimate$$

The *point estimate* in the interval just considered was the sample mean $\overline{y}$. The *multiplier* was the upper $\alpha/2$ probability point from the $t$ distribution. The *standard error of the point estimate* in the interval was the standard error of $\overline{y}$, $\frac{s}{\sqrt{n}}$.

## 2.5 Hypothesis Testing about a Normal Population Mean

The previous section described a statistical technique for estimating a normal population mean with an interval of plausible values and providing a measure of the reliability of the interval. Another statistical technique that is used in practice is hypothesis testing. In hypothesis testing a researcher wishes to provide evidence in favor of a conjecture involving an unknown population mean. Data is collected and based on the data the conjecture is either supported or not. The basic ideas are illustrated with an example.

**Example 2.4** *Suppose that a standard method of treating a disease in the past has resulted in a (population) mean survival time of* 5 *years or* 60 *months. The actual survival time for particular individuals has varied from* 60 *months due to extraneous variables. A new treatment is being proposed which is believed to increase the (population) mean. A sample of* 15 *patients with the disease is given the new treatment and their survival times (in months) are given below.*

*61 55 68 62 65 54 70 63 56 51 72*
*63 76 53 71*

*The sample mean* $\overline{y} = 62.7$ *months, sample standard deviation* $s = 7.7$ *months, and standard error of the mean is* $s_{\overline{y}} = 2.0$ *months. Is this enough evidence to conclude that the new treatment results in higher (population) average survival time? Use a significance level of* $\alpha = 0.05$.

Certainly the sample mean of 62.7 months is greater than 60 months, but this mean is based on a sample, not the entire "population" of individuals that could be treated. Is the difference between 62.7 months and 60 months "real", that is an indication that the true population mean with the new treatment is greater than 60 months? Or could we obtain a sample mean of 62.7 months simply due to sampling variability and the fact that survival times will vary naturally, even if the true population mean with the new treatment is no different than 60 months. That is, is the result due solely to chance (sampling) or is the new treatment really better?

Let $\mu$ be the true population mean survival time with the new treatment. The claim that the researcher hopes to provide evidence for is $\mu > 60$, which is called the **alternative** claim or alternative hypothesis and denoted by $H_a : \mu > 60$. Of course the opposite or the null hypothesis could be true, which is denoted by $H_o : \mu \leq 60$.

The general approach to decision making in hypothesis testing is as follows:

- Assume initially that $H_o$ is true
- Assuming $H_o$ is true, calculate a summary of the data called the **test statistic**. The probability distribution of the test statistic is known.
- Calculate the probability of obtaining a value of the test statistic like the observed value or more extreme in the direction of the alternative hypothesis if in fact $H_o$ is true. This probability is called the **P-value**.
- If the *P-value* is less than or equal to ($\leq$) some prescribed probability then reject $H_o$ as true and conclude that $H_a$ is true. If the *P-value* is greater than ($>$) the prescribed probability then $H_o$ is not rejected - the null hypothesis could be true. This prescribed probability is called the **significance level** of the test and denoted by $\alpha$. A common value used for $\alpha$ is 0.05.

For this example suppose that $H_0$ is true and for the moment suppose that $\mu = 60$ months, that is, there is no difference in population mean survival time between the new treatment and the standard treatment. Under this assumption and normality of the population it is known from the previous section that

$$\frac{(\overline{y} - 60)}{s/\sqrt{15}}$$

treated as a variable, has the Student's $t$ distribution with degrees of freedom $\nu = 15 - 1 = 14$. The variable $\frac{(\overline{y}-60)}{s/\sqrt{15}}$ is the *test statistic* in this example. Now the *observed value* of the test statistic is

$$\frac{(62.7 - 60)}{2.0} = 1.35$$

That is, the observed sample mean of 62.7, is 1.35 standard errors above the null hypothesized value for $\mu$ of 60. The P-value is the probability of obtaining a value of the test statistic like the observed value of 1.35 or more extreme in the direction of the alternative hypothesis, here greater than 1.35 since the alternative hypothesis points to value of $\mu > 60$. Symbolically the P-value for this example is

$$P[\frac{(\overline{y} - 60)}{s/\sqrt{15}} \geq 1.35]$$

which is the area of the shaded region under the t-curve in Figure 2.4.

Using Table A.2 with $\nu = 15 - 1 = 14$ degrees of freedom, the P-value (area) is approximately 0.10. A statistical program such as SAS or SPSS will give P-value = 0.1010. Thus there is about a 10% chance of obtaining a value of the test statistic, $\frac{(\overline{y}-60)}{s/\sqrt{15}}$, like the observed value of 1.35 or greater if in fact the null hypothesis is true, that is $\mu = 60$.

Using a significance level of $\alpha = 0.05$, since the P-value of $0.1010 > 0.05$ there is not enough evidence to reject the population mean being equal to 60 months and thus not enough evidence to support the researcher's claim that the new treatment extends the survival times of these patients. This conclusion of not enough evidence to support the new treatment extending survival time is based on survival times from 15 patients. A larger study with more patients may have reached a different conclusion.

## 2.6 The General Form of the One Sample t test

The example in the previous section was an example of a one-sided single sample $t$ test. The term one-sided comes from the form of the alternative hypothesis and the fact that the alternative is supported if the observed value of the test statistic is on one side, the upper side, of the appropriate $t$ distribution. The general form of null and alternative hypotheses for the three versions of the $t$ test are given in Table 2.2, where $\mu_o$ some reference or standard value.

Figure 2.4: P-value for Example 2.4

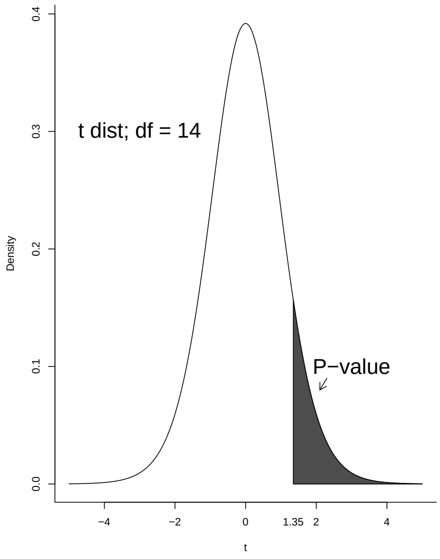


Table 2.2: General Forms of Null and Alternative Hypotheses for One Sample t test

| (1) | (2) | (3) |
|---|---|---|
| $H_o : \mu \leq \mu_o$<br>$H_a : \mu > \mu_o$ | $H_o : \mu \geq \mu_o$<br>$H_a : \mu < \mu_o$ | $H_o : \mu = \mu_o$<br>$H_a : \mu \neq \mu_o$ |

Table 2.2 is a generalization of Example 2.4. The **test statistic** for all three tests is the $t$ statistic,

$$t = \frac{\overline{y} - \mu_o}{s/\sqrt{n}}$$

which has a "$t$" sampling distribution if the population is normally distributed and an approximate $t$ distribution as long as the sample size is "large."

Let $t^*$ be the observed value of the $t$ statistic based on the data. Then the P-values for the alternatives (1), (2), and (3) in Table 2.2 are, respectively, $P[t \geq t^*]$, $P[t \leq t^*]$, and $P[|t| \geq |t^*|]$. The alternative hypotheses in (1) and (2) of Table 2.2 are called one-sided alternatives and the tests are called one-sided tests. The alternative hypothesis in (3) of Table 2.2 is called a two-sided alternative and the test is called a two-sided test.

## 2.7 Errors and Probabilities of Errors in Hypothesis Testing

In the decision making process of hypothesis testing one of two possible errors may result. The null hypothesis is really true yet the data and the test indicate that the null hypothesis should be rejected and the alternative accepted. This is called a **Type I error**. The other possible error is incurred if the alternative hypothesis is true but the null hypothesis is retained or the alternative hypothesis is not accepted. This is called a **Type II error**.

In Example 2.4 concluding that the new drug increases the survival time as compared to the standard treatment when in fact the true mean survival time is 60 months (or less) would be a Type I error. A Type II error would be to not conclude that the new drug increases survival time (fail to reject the null hypothesis) when in fact the new drug does increase mean survival time.

Researchers cannot guarantee that either error is not made but they can ensure that the probabilities of making these errors are low. Let's consider the probability of making the Type I error first.

Recall in the one sample t test that we reject the null hypothesis if the P-value, calculated assuming the null is true, is smaller than some prescribed probability, called the significance level, or the $\alpha$ level. Typical values used here are $\alpha = 0.01$ or $\alpha = 0.05$. Symbolically the null hypothesis is rejected if $P - value \leq \alpha$. Now it is possible to obtain a P-value this low even if the null hypothesis is true and thus mistakenly reject a true null hypothesis, a Type I error. In fact the probability is exactly $\alpha$ of obtaining a $P - value \leq \alpha$ when in fact the null hypothesis is true. Thus the probability of making a Type I error is $\alpha$, a value prescribed by the researcher. Thus it is relatively easy to control the probability of a Type I error. If the Type I error is a very serious error then the researcher or some regulatory authority would request that the significance level or $\alpha$ level be perhaps set at 0.01 rather than 0.05. Note that regardless of sample size setting the $\alpha$ level to some prescribed value ensures that the probability of a Type I error is fixed at that level. Sample size however does influence the probability of a Type II error.

The probability of a Type II error, $\beta$, is not as easily controlled as the probability of a Type I error. Discussion of the probability of a Type II error usually focuses on $1-\beta$, called the **power** of the test. Thus power is the probability of a correct decision, that of concluding that the alternative hypothesis is true based on the test, when in fact the alternative hypothesis is true. Thus researchers want $\beta$ to be small, such as 0.20 or 0.10, or power to be high such as 0.80 or 0.90.

There are many different values for $\beta$ or power $(1-\beta)$ depending upon various characteristics of the study. For example in the one sample t test $\beta$, or power depends upon:

1. Sample size. For a particular population standard deviation and particular $\alpha$ level, increasing sample size will decrease $\beta$ (increase power).

2. The $\alpha$ level. For a particular population standard deviation and fixed sample size, increasing $\alpha$ decreases $\beta$ (increases power) and decreasing $\alpha$ increases $\beta$ (decreases power).

3. Population standard deviation. For fixed sample size and $\alpha$ level, $\beta$ will be larger (power smaller) for populations with larger standard deviations.

4. The difference between the null hypothesis value of $\mu$ and the true value of $\mu$ under the alternative hypothesis. In our example the null hypothesis value of $\mu$ was $\mu_o = 60$ months. If the new drug is more effective (alternative true) and the true mean is $\mu_a = 62$ months then the difference or effect of the drug is an increase of 2 months. If the new drug is more effective and the true mean is $\mu_a = 70$ months then the difference or effect of the drug is an increase of 10 months. The $\beta$ level will depend upon the difference or "effect" in this example. The greater the effect the smaller the level of $\beta$ or the higher the power. Thus in this example it is more likely that the new drug is correctly concluded as being effective if it is a lot more effective as compared with being minimally effective.

The probability of correctly concluding a true alternative hypothesis that is, $1-\beta$, is called the power of a test. Table 2.3 gives the power of a one-sided single sample t test using a significance level of $\alpha = 0.05$ in terms of the standardized effect E,

$$E = \frac{|\mu_a - \mu_o|}{\sigma}$$

and sample size $n$. The values $\mu_o$ and $\mu_a$ are null and alternative values of $\mu$, respectively.

Table 2.3: Power of One-Sided One Sample t test, $\alpha = 0.05$

| | Standardized Effect E | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample Size | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| 5 | 0.073 | 0.102 | 0.140 | 0.185 | 0.239 | 0.300 | 0.366 | 0.436 | 0.508 | 0.580 |
| 10 | 0.088 | 0.145 | 0.222 | 0.317 | 0.427 | 0.543 | 0.655 | 0.754 | 0.836 | 0.898 |
| 15 | 0.101 | 0.182 | 0.295 | 0.432 | 0.578 | 0.714 | 0.824 | 0.903 | 0.952 | 0.979 |
| 20 | 0.112 | 0.217 | 0.363 | 0.532 | 0.695 | 0.827 | 0.915 | 0.964 | 0.987 | 0.996 |
| 25 | 0.123 | 0.250 | 0.426 | 0.617 | 0.783 | 0.898 | 0.960 | 0.987 | 0.997 | 0.999 |
| 30 | 0.134 | 0.283 | 0.484 | 0.690 | 0.848 | 0.941 | 0.982 | 0.996 | 0.999 | 1.000 |
| 35 | 0.143 | 0.314 | 0.538 | 0.750 | 0.895 | 0.967 | 0.998 | 0.999 | 1.000 | 1.000 |
| 40 | 0.153 | 0.344 | 0.587 | 0.800 | 0.928 | 0.981 | 0.992 | 1.000 | 1.000 | 1.000 |
| 45 | 0.162 | 0.373 | 0.632 | 0.841 | 0.951 | 0.990 | 0.997 | 1.000 | 1.000 | 1.000 |
| 50 | 0.172 | 0.401 | 0.673 | 0.874 | 0.967 | 0.994 | 0.999 | 1.000 | 1.000 | 1.000 |
| 55 | 0.181 | 0.428 | 0.710 | 0.900 | 0.978 | 0.997 | 0.999 | 1.000 | 1.000 | 1.000 |
| 60 | 0.190 | 0.455 | 0.743 | 0.922 | 0.985 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 |
| 65 | 0.198 | 0.480 | 0.773 | 0.939 | 0.990 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 |
| 70 | 0.207 | 0.505 | 0.800 | 0.952 | 0.994 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 75 | 0.216 | 0.528 | 0.824 | 0.963 | 0.996 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 80 | 0.224 | 0.551 | 0.845 | 0.971 | 0.997 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 85 | 0.233 | 0.573 | 0.864 | 0.978 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 90 | 0.241 | 0.594 | 0.881 | 0.983 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 95 | 0.249 | 0.614 | 0.896 | 0.987 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 100 | 0.257 | 0.634 | 0.909 | 0.990 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 150 | 0.335 | 0.786 | 0.978 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 200 | 0.407 | 0.880 | 0.995 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 500 | 0.722 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

# SAS Code for Chapter 2

Example 3.2

```
* Input survival times;
Data SURVIVAL;
  Input Survival_Time @@;
datalines;
61  55  68  62  65  54  70  63  56  51  72
63  76  53  71
;
run;

* Use proc ttest to obtain results of one sample t test;
Proc ttest ho=60 data = SURVIVAL;
var Survival_time;
run;
```

# Problems for Chapter 2

2.1* Rebecca Aaron and Rebecca Redman in 2014 conducted an experiment to study the effects of brand of rubber band and temperature conditioning of the rubberband on how far a rubber band would stretch until breaking. The two brands of rubberbands were Staples and CLI (Douglas Stewart Company). Both brands were approximately 9 cm in length and 0.32 cm in width. Prior to stretching rubberbands were exposed to either cold (freezer for 24 hours), heat (microwave for 1 minute), or room temperature. Rubberbands were tested/stretched one a time. The procedure was as follows. A combination of brand and temperature was randomly selected. A rubberband of the randomly selected brand and temperature conditioning was selected. An apparatus was used to stretch the rubberband until it broke with a tape measure below the rubberband to measure the stretched length. The stretching was videotaped to aid in the measurement. This process was repeated for a total of 30 rubberbands, 5 for each of the combinations of brand and temperature conditioning. Given below are the stretched lengths (cm) of the Staples heated rubberbands.

72.6 76.2 79.0 66.6 72.5

Find the mean and standard deviation of the lengths. Interpret the standard deviation within the context of this study.

2.2* A group of 48 pigs receiving a new medicine for treating a bacterial intestinal disease gained during the study period on average 1.25 pounds per day with a standard deviation of 0.2 pound. What is the standard error of the sample mean of 1.25 pounds? Explain within the context of this example the difference between the sample standard deviation and the standard error of the mean?

2.3* Suppose that a random sample of size n = 16 is selected from a normal population with $\mu = 50$ and standard deviation $\sigma = 5$. Let the random variables $\overline{y}$ and $s$ refer to the sample mean and sample standard deviation, respectively.

a. What is the form of the probability/sampling distribution of the sample mean $\overline{y}$?

b. What is the form of the probility/sampling distribution of $\frac{\overline{y}-50}{5/\sqrt{16}}$?

c. What is the form of the probability/sampling distribution of $\frac{\overline{y}-50}{s/\sqrt{16}}$?

2.4* Suppose a random sample of size n = 25 is selected from a normal population with $\mu = 100$. Let the random variables $\overline{y}$ and $s$ represent the sample mean and sample standard deviation of the sample.

a. What is the upper 0.05 probability point (or 95$th$ percentile) of the sampling distribution of $\frac{\overline{y}-100}{s/\sqrt{25}}$

b. Find $P[-1.318 < \frac{\overline{y}-100}{s/\sqrt{25}} < 1.318]$.

2.5* One question asked of randomly selected students at a university was how many hours the student typically spent studying during the week. Based on the data for $n = 30$ responses the 95% confidence interval for the mean amount of time spent studying was $(28.7, 36.5)$.

a. What is the sample mean amount of time for the 30 students?

b. What is the 95% error margin for the sample mean from part (a)?

c. What is the standard error associated with the sample mean from part (a)?

d. Interpret the interval $(28.7, 36.5)$ within the context of this example.

e. Suppose a different set of $n = 30$ students were random selected from the same population of students. Would we get a different interval? Explain.

2.6* This example is taken from Devore and Peck ([8], page 555). Calorie contents for each of $n = 12$ frozen dinners was taken from the production line during a particular period and are reported in the table below:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 255 | 244 | 239 | 242 | 265 | 245 | 259 | 248 |
| 225 | 226 | 251 | 233 | | | | |

The calorie content given on the box is 240. Do the data give any reason to believe that the true mean calorie of the population of frozen dinners is different than stated on the box? Carry out a hypothesis test using a significance level of 0.05. Use a statistical program to obtain a $P - value$. Use this $P - value$ to make your decision. Interpret the P-value within the context of this example.

2.7* In the article "Quality of Life and Functional Health Status of Long-Term Meditators"(*Evidence-Based Complementary and Alternative Medicine vol. 2012, Article ID 350674, 9 pages 2012. doi:10.1155/2012/350674*) researchers compared the quality of life and functional health of a sample of 343 long-term meditators to that of population norms for Australia. Participants completed the Medical Outcomes Study Short Form 36 (MOS SF-36). One of the domains of health evaluated in the SF-36 is mental health. The population norm score for Australia is 75.75. The mean and standard deviation of the 337 respondents with valid mental health scores was 85.31 and 12.31, respectively. The authors used a one sample t test to compare the sample mean to the population norm of 75.75.

a. Give the null and alternative hypothesis for the test. Assume that a two-sided test was used. Be sure to define the relevant population mean of interest in symbols and in words.

b. Calculate the value of the test statistic.

c. Suppose the P-value based on computer software is $P < 0.0001$. Using a significance level of 0.05, what is the conclusion regarding mental health scores of long term meditators as compared to the population?

d. Calculate a 95 percent confidence interval for the population mean mental health score and interpret within the context of this problem.

e. It is possible that an error was made in your conclusion in part(c). What is the name of this type of error?

2.8* In one part of a research study ("Glass Shape Influences Consumption Rate for Alcoholic Beverages," PloS ONE 7 (2012) e43007)), each of 160 participants (50% male) viewed computer images of straight and curved 12 fl. oz glasses, the activity resulting in a numerical value reflecting what the participant perceived to represent half full. The true value of a half full glass was 30 with the perceived values by the participants deviating from this value. We will only consider the data reported on the curved glass. The article reported that the mean and standard deviation of perceived midpoints of the curved glass by the 160 participants were 21 and 3, respectively. The authors conducted a one-sample t test to determine if participants' perceived midpoint differed on average from the true value of 30. The value of the one sample t statistic reported was $-37.97$, $P < 0.001$.

a. Give the null and alternative hypothesis for the test. Be sure to define the relevant population mean of interest in symbols and in words.

b. What is the conclusion based on the results of the test using a significant level of 0.05?

2.9* Suppose that you are repeating the perception study of Exercise 2.7 to see if you get similar results. You have limited resources and can only employ 20 subjects. You will also use a curved glass but of a different type of curvature than that used in the study of Exercise 2.7. You will conduct a one-sided one-sample t test to determine if the (mean) perceived midpoint is less than 30. You don't believe that the drop in perceived midpoint will be as low as that observed in the article and might be only about a 2 unit drop. Using Table 2.3, what is the (approximate) power of your test? Explain what this number means. Assume that population standard deviation is 3 units. A significance level of 0.05 will be used for the t test.

2.10* The mean drying time of a spray paint is known to be 90 seconds. The research division of the company that produces this paint contemplates that adding a new chemical ingredient to the paint will accelerate the drying process. To investigate this conjecture, the paint with the chemical additive will be sprayed on a number of surfaces and the drying times are recorded. For similar experiments drying time has been approximately normally distributed with a standard deviation of drying times of about 8 seconds. A drop of 5 seconds (on average) is deemed to be of practical importance. The one-sample $t$ test with a significance level of 0.05 will be used. The research division wants the power of the test to be about 90%. What sample size would you recommend to the research division?

# Chapter 3

# The Two Sample Problem

In this chapter it is assumed that there are two samples of quantitative normally distributed data corresponding to the treatments in an experiment or the two populations in an observational study. The two scenarios where the two samples are independent or where they are dependent will be examined. Section 3.1 will examine hypothesis testing and confidence interval estimation in the two independent samples situation. Section 3.2 will examine inferences within the dependent samples situation.

## 3.1 Two Independent Samples/Completely Randomized Design

In this section it is assumed two samples of quantitative data have been gathered by one of two designs:

- An experiment has been conducted whereby two treatments have been assigned completely at random to two groups of experimental units, that is a **completely randomized design**.
- A survey or observational study has been conducted whereby two samples have been randomly and independently selected from two different populations.

As an example of a survey, a group of students doing a project wanted to compare GPAs of male and female undergraduates at their universities. They obtained a list of all undergraduate male and all undergraduate female students and randomly selected 100 students from each list.

Let $y_{11}, y_{12}, ..., y_{1n_1}$ represent the values of a random sample of size $n_1$ from a normal population with mean $\mu_1$ and variance $\sigma_1^2$. Let $\overline{y}_{1.}$ and $s_1$ represent the sample mean and standard deviation of the sample. Then from Chapter 2, $E[\overline{y}_{1.}] = \mu_1$ and the standard deviation of $\overline{y}_{1.}$ is $\sigma_{\overline{y}_{1.}} = \sigma_1/\sqrt{n_1}$ . Similarly, let $y_{21}, y_{22}, ..., y_{2n_2}$ represent the values of a random sample of size $n_2$ from another normal population with mean $\mu_2$ and variance $\sigma_2^2$. Let $\overline{y}_{2\cdot}$ and $s_2$ represent the sample mean and standard deviation of the sample. Then from Chapter 2, $E[\overline{y}_{2\cdot}] = \mu_2$ and the standard deviation of $\overline{y}_{2\cdot}$ is $\sigma_{\overline{y}_{2\cdot}} = \sigma_2/\sqrt{n_2}$.

Suppose the purpose of the two sample study is to compare the two unknown population means of y, $\mu_1$ and $\mu_2$. The comparison is typically carried out by drawing inferences on the unknown difference $\mu_1 - \mu_2$. Similar to Chapter 2, we need an estimator of $\mu_1 - \mu_2$ and the standard deviation of this estimator. We also need to know the probability or sampling distribution of the estimator. The usual estimator of $\mu_1 - \mu_2$ is the analogous difference in sample means, $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$.

### 3.1.1 Sampling Distribution of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$

Properties of the sampling distribution of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ are given below.

- The mean of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ is $\mu_{\overline{y}_{1\cdot}-\overline{y}_{2\cdot}} = E[\overline{y}_{1\cdot} - \overline{y}_{2\cdot}] = \mu_1 - \mu_2$. Thus while differences in sample means $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ will vary from pair of samples to pair of samples the average of these differences is equal to the difference in population means $\mu_1 - \mu_2$.

- The variance of $(\overline{y}_{1\cdot} - \overline{y}_{2\cdot})$ is $\sigma^2_{\overline{y}_{1\cdot}-\overline{y}_{2\cdot}} = E[\{(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - (\mu_1 - \mu_2)\}^2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. The variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ measures the average squared distance between differences in sample means $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ and the difference in the population mean $\mu_1 - \mu_2$.

- The sampling distribution of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ is normal if the two populations are normal and approximately normal if the two sample sizes are “large.”

The standard deviation of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ as an estimate of $\mu_1 - \mu_2$ is the square root of the variance, $\sigma_{\overline{y}_{1\cdot}-\overline{y}_{2\cdot}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. The standard deviation gives the average distance differences in sample means are from the difference in population means $\mu_1 - \mu_2$. The population variances $\sigma_1^2$ and $\sigma_2^2$ are unknown. So in order to be of practical value these two population variances need to be estimated. We will first consider hypothesis testing and confidence interval estimation when the two population variances are estimated separately with the sample variances $s_1^2$ and $s_2^2$.

### 3.1.2 Two sample $t$ test and Confidence Interval

Since $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ then the standardized version

$$\frac{(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - \mu_{\overline{y}_{1\cdot}-\overline{y}_{2\cdot}}}{\sigma_{\overline{y}_{1\cdot}-\overline{y}_{2\cdot}}} = \frac{(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution. Suppose that we substitute the sample variances $s_1^2$ and $s_2^2$ for the population variances in the standard deviation in the denominator to obtain the **standard error** of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$, $s_{\overline{y}_{1\cdot}-\overline{y}_{2\cdot}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Then the resulting ratio

$$\frac{\overline{y}_{1\cdot} - \overline{y}_{2\cdot} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.1}$$

has an approximate Student's $t$ distribution if the populations are normally distributed or if the sample sizes are sufficiently large. The degrees of freedom, $\nu$, used for this approximate $t$ distribution is called the Satterthwaite approximation and is data based:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Fortunately we can use computer software to obtain the degrees of freedom and P-values.

The general form of the two-sided $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) \pm t_{\alpha/2;\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}} \tag{3.2}$$

where $t_{\alpha/2;\nu}$ is the upper $\alpha/2$ probability point from a $t$ distribution with degrees of freedom, $\nu$, the Satterthwaite approximation. The Satterthwaite degrees of freedom is not usually integer. If using Table A.2 then round down to the nearest integer to obtain the probability point. This will result in a wider or more conservative interval. Computer software will give more precise probability points and intervals. Note that the interval of possible values for $\mu_1 - \mu_2$ is formed by taking the estimate $(\overline{y}_{1\cdot} - \overline{y}_{2\cdot})$ and adding and subtracting a margin of error, here $t_{\alpha/2;\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$. The margin of error is the product of a probability point from a $t$ distribution and the standard error of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$. This is the same general form as the confidence interval for a single population mean given in Chapter 2.

The general forms of the null and alternative hypotheses for a test involving the difference between $\mu_1$ and $\mu_2$ are given in Table 3.1. Note that the alternative hypotheses in (1), (2) and (3) reflect differences (one or two directional) in the two means and thus the test is used to determine if there is sufficient evidence of a difference of some specified type.

The **test statistic** for the two independent samples $t$ test is given in (3.3) and is just the ratio (3.1) assuming that the null hypothesis is true, in particular that $\mu_1 - \mu_2 = 0$. Thus the numerator is a measure of how far away the observed difference in sample means is away from the null hypothesized value of 0 for the difference in population means. If the observed difference in sample means is sufficiently far from 0 then the null hypothesis of no difference in population

Table 3.1: General Forms of $H_o$ and $H_a$ for Two Sample t test

| (1) | (2) | (3) |
|---|---|---|
| $H_o : \mu_1 - \mu_2 \leq 0$ | $H_o : \mu_1 - \mu_2 \geq 0$ | $H_o : \mu_1 - \mu_2 = 0$ |
| $H_a : \mu_1 - \mu_2 > 0$ | $H_a : \mu_1 - \mu_2 < 0$ | $H_a : \mu_1 - \mu_2 \neq 0$ |

means is rejected in favor of the alternative of a difference. Sufficiently far away can be defined in terms of P-values like in Chapter 2.

$$t = \frac{\overline{y}_{1\cdot} - \overline{y}_{2\cdot} - (0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.3}$$

The P-value is the probability, assuming equal population means, of getting a value of the test statistic (3.3) like the observed value or more extreme in the direction of the alternative hypothesis. The P-value is an area under a $t$ curve with degrees of freedom $\nu$ equalling the Satterthwaite approximation. For the one-sided alternative hypothesis (1) in Table 3.1 the P-value is the area under the t-curve to the right of the observed value of (3.3). For the one-sided alternative hypothesis (2) in Table 3.1 the P-value is the area under the t-curve to the left of the observed value of (3.3). For the two-sided alternative hypothesis (3) in Table 3.1 the P-value is twice the area to the right of the observed value if the observed value is positive or twice the area to the left of the observed value if the observed value is negative. As with the confidence interval, the Satterthwaite degrees of freedom is generally not integer. If using Table A.2 to obtain an approximate P-value then round down to the nearest integer for a conservative value. Computer software will calculate more precise P-values based on the Satterthwaite degrees of freedom.

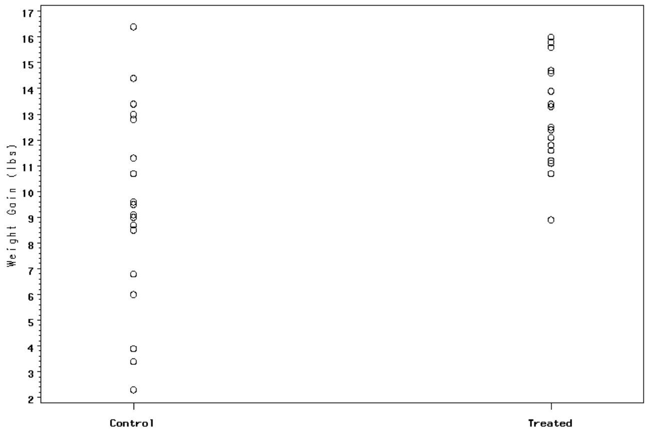
**Example 3.1** *Animal health researchers develop drugs to treat diseases of animals. Suppose that in one study $n_1 = 22$ pigs are treated with a medication to control an intestinal disease while $n_2 = 18$ other pigs served as a control and were not treated. Weight gain (lbs.) is measured over the study period and is reported in the table below.*

| | |
|---|---|
| *Control(2)* | *16.4,12.8,13.0,10.7,3.9,9.1,8.7,9.5,8.5,6.0* |
| | *9.0,13.4,3.4,9.6,14.4,11.3,6.8,2.3* |
| *Treated(1)* | *11.6,8.9,14.6,12.4,13.3,16.0,11.1,15.8,15.6,10.7* |
| | *12.4,14.6,11.2,10.7,11.6,14.7,13.9,11.8,13.4,12.1* |
| | *13.4,12.5* |

*Is there sufficient evidence that the medication improves weight gain for pigs?*

Figure 3.1 gives a plot of the weight gains versus treatment. Weight gain does appear to be improved in the treated group. There also appears to be less variability in weight gains in the treated group.

Figure 3.1: Plot of Weight Gain (lbs) versus Treatment



Let $n_1$ and $n_2$ represent the numbers of pigs in the treated and control groups respectively. The sample mean weight gains for the treated and control groups, respectively, are $\overline{y}_{1.} = 12.83$ lbs. and $\overline{y}_{2.} = 9.38$ lbs. The standard deviation of weight gains for the treated and control groups are, respectively, $s_1 = 1.87$ lbs. and $s_2 = 3.88$ lbs. Let $\mu_1$ and $\mu_2$ represent the "true" mean weight gains for the treated and control pigs, respectively. Then the null and alternative hypotheses are of the form (1) in Table 3.1. The observed value of the test statistic is

$$t = \frac{\overline{y}_{1.} - \overline{y}_{2.}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{12.83 - 9.38}{\sqrt{\frac{1.87^2}{22} + \frac{3.88^2}{18}}} = \frac{3.48}{0.99} = 3.46$$

Degrees of freedom would be

$$\nu = \frac{\left(\frac{1.87^2}{22} + \frac{3.88^2}{18}\right)^2}{\frac{1}{22-1}\left(\frac{1.87^2}{22}\right)^2 + \frac{1}{18-1}\left(\frac{3.88^2}{18}\right)^2} = 23.4$$

The P-value is the probability of getting a value of the test statistic like the observed value, 3.46, or more extreme (greater than) if in fact there is no difference in population mean weight gains for the treated and control groups. The P-value can only be approximated using Appendix Table A.2. Rounding down and using $\nu = 23$ from Table A.2 we see that

$$0.0005 \leq P - value \leq 0.005$$

Thus at $\alpha = 0.05$ there is evidence that weight gain is improved with the medication.

The mean improvement in weight gain could be estimated with a 95% confidence interval using Formula 3.2. The appropriate upper 0.025 probability point, using $\nu = 23$ and Appendix Table A.2 would be 2.069. Thus the confidence interval would be

$$(12.83 - 9.38) - 2.069(0.99) < \mu_1 - \mu_2 < (12.83 - 9.38) + 2.069(0.99)$$

or

$$3.45 - 2.05 < \mu_1 - \mu_2 < 3.45 + 2.05$$

or

$$1.4 < \mu_1 - \mu_2 < 5.5$$

Thus it is estimated with 95% confidence that the true effect of the treatment when compared to control is to increase average weight gain by somewhere from 1.4 to 5.5 pounds.

### 3.1.3 Two Independent Samples "Pooled" $t$ test and Confidence Interval

In some experimental and survey situations it is reasonable to assume that the two population variances are equal either based on the data or on theoretical considerations, that is, it is assumed that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Thus the standard deviation of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$ can be written as $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$.

**Example 3.2** *An example from McClave and Sincich [18], page 329, will illustrate. A new method of teaching reading to children who are slow learners is compared to a current standard method. The comparison is based on a reading test score given at the end of the learning period. Ten subjects are taught by the new method and 12 are taught by the standard method. The results of the reading scores are given in the table. Is there statistical evidence that the new method results in higher scores? Use a significance level of* 0.05.

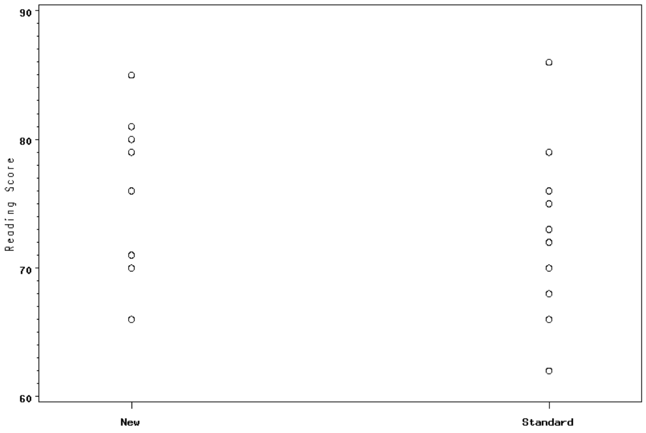
| | |
|---|---|
| *New Method (1)* | *80,76,70,80,66,85,79,71,81,76* |
| *Standard Method (2)* | *79,73,72,62,76,68,70,86,75,68,73,66* |

A plot of the scores versus the method is given in Figure 3.2. Note that average and variation in scores are similar for the two methods.

Let $n_1 = 10$ and $n_2 = 12$ represent the numbers of children receiving the new and standard methods respectively. The sample mean reading scores for the new and standard method groups are $\overline{y}_{1\cdot} = 76.40$ and $\overline{y}_{2\cdot} = 72.33$. The standard deviation of reading scores for the new and standard method groups are, respectively, $s_1 = 5.83$ and $s_2 = 6.34$. Let $\mu_1$ and $\mu_2$ represent the "true" mean reading scores for the new and standard methods. Then the null and alternative hypotheses are of the form (1) in Table 3.1.

If the assumption of equal population variances is reasonable and since then there is only one unknown population variance, it makes sense to combine the two sample variances into one "pooled" sample variance, $s_p^2$, where

Figure 3.2: Plot of Reading Score versus Method



$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

Note that the numerator is just the sum of squared deviations of the scores from their respective means. The denominator is the sum of the degrees of freedom associated with the two sample variances. Estimating the common population variance, $\sigma^2$ with $s_p^2$ we have the standard error of $\overline{y}_{1.} - \overline{y}_{2.}$, $s_{\overline{y}_{1.}-\overline{y}_{2.}}$ to be

$$s_{\overline{y}_{1.}-\overline{y}_{2.}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

or

$$s_{\overline{y}_{1.}-\overline{y}_{2.}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In order to construct confidence intervals and perform hypothesis testing we need to have a sampling distribution to calculate P-values and to obtain probability points for error margins. From earlier it is known that $\overline{y}_{1.} - \overline{y}_{2.}$ is normally distributed with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$, assuming equal population variances. Thus $\frac{(\overline{y}_{1.}-\overline{y}_{2.})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ has the standard normal probability distribution and P-values and error margins could be based on this distribution. However, again, the standard deviation in the denominator is unknown. If we replace the standard deviation with its estimate, standard error, then the ratio

$$\frac{(\overline{y}_{1\cdot}-\overline{y}_{2\cdot})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

has a $t$ distribution with degrees of freedom $\nu = n_1 + n_2 - 2$ if the two populations are normally distribution and an approximate $t$ distribution if sample sizes are large. This ratio serves as the test statistic for a test comparing the two population means.

Using the general form of a confidence interval from Chapter 2, we have that the general two sided $100(1-\alpha)\%$ **independent pooled samples confidence interval** for $\mu_1 - \mu_2$ is

$$(\overline{y}_{1\cdot}-\overline{y}_{2\cdot}) \pm t_{\alpha/2;(n_1+n_2-2)} s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Continuing with our example, we have that

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}} = \sqrt{\frac{(10-1)5.83^2+(12-1)6.34^2}{(10-1)+(12-1)}} = 6.12$$

The standard error of $\overline{y}_{1\cdot}-\overline{y}_{2\cdot}$ is therefore

$$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 6.12\sqrt{\frac{1}{10}+\frac{1}{12}} = 2.62$$

The observed value of the test statistic for the **independent samples pooled samples t test** under the null hypothesis is thus

$$\frac{(\overline{y}_{1\cdot}-\overline{y}_{2\cdot})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(72.23-68.30)-(0)}{2.62} = 1.55$$

The appropriate degrees of freedom is $\nu = n_1 + n_2 - 2 = 20$. The P-value can only be approximated using Appendix Table A.2 with

$$0.05 < P - value < 0.10$$

Thus there is not enough evidence that reading scores are improved with the new method at the 0.05 level of significance.

The appropriate upper 0.025 probability point for a 95% confidence interval with $\nu = 20$ from Appendix Table A.2 would be 2.086. Thus the pooled samples t confidence interval would be

$$(76.40-72.33)-2.086(2.62) < \mu_1-\mu_2 < (76.40-72.33)+2.086(2.62)$$

or

$$4.07-5.47 < \mu_1-\mu_2 < 4.07+5.47$$

or

$$-1.40 < \mu_1 - \mu_2 < 9.54$$

The results based on the interval are inconclusive. With 95% confidence, the new method may result in greater reading scores by as much as 9.54; there could be no difference in mean reading scores between the two methods; or the standard method may result in greater reading scores by as much as 1.40.

### 3.1.4 Which independent samples $t$ test to use?

As previously described there are two possible tests, the two sample $t$ test and the pooled two sample $t$ test, for comparing means of normal populations when the samples are independent. When the sample sizes are the same the test statistics for the two procedures take on identical values; however P-values will be different since the degrees of freedom will generally be different. When the sample sizes are different the two procedures will generally result in different values of the test statistics, degrees of freedom, and P-values. The two sample $t$ test does not make any assumptions about population variances whereas the pooled $t$ test assumes population variances are equal.

If the two population variances are equal both tests are valid. The pooled $t$ test does have slightly higher power. But how does one know if the population variances are equal. These are unknown population characteristics. There are statistical tests for comparing population variances but the test are extremely sensitive to the assumption of normality and significant results may indicate a difference in standard deviations or non-normality. Also the hypothesis tests do not address the magnitude of the difference in the population variances. The pooled two sample $t$ test is still approximately valid in some circumstances when the population variances, while not equal, are approximately the same. Instead of tests, some authors recommend rules of thumb regarding sample standard deviations (or sample variances) to decide if the assumption is reasonable. Cobb [4] recommends that if the ratio of the largest to smallest standard deviation is greater than 3 then do not assume that the population standard deviations are equal. Agresti and Franklin [1] comment that in practice equality of population standard deviations is not relied upon if the ratio of the largest to smallest standard deviation is greater than two.

When the population variances are not equal, the two sample $t$ test is valid. The pooled samples $t$ test may be invalid depending upon the degree of difference between the population variances and the distribution of the sample sizes across the two groups.

The recommendation of this text for two sample comparisons is the two sample $t$ test. It is approximately valid regardless of the population variances. Some elementary statistics texts discuss the pooled $t$ test but also recommend using the two sample $t$ test (See DeVeaux, Velleman, and Bock [7]; Peck and Devore [12]).

So why is the pooled samples $t$ test even discussed in this text? As is shown in later chapters the assumption of equal population variances is a basic assumption of ANOVA, a main topic of this text. In fact the ANOVA for a one factor study with only two groups is equivalent to the pooled samples $t$ test.

An alternative approach to handling unequal variances besides the two sample t test would be to transform the data to a new scale where the variances would be equal or approximately equal. A larger discussion on statistical assumptions of ANOVA, including equal population variances, will be conducted in Chapter 8.

## 3.2 Two Dependent/Paired Samples

In this section we consider two sample designs where the two samples are dependent or paired. Listed below are types of pairing (See Cobb [4]) and examples. These are all examples of blocking as discussed in Chapter 1.

### Types of Pairing/Blocking

- **Re-Using**: Each person or object is measured at two different time slots or occasions. There may be only two treatments in an experiment. Each individual receives one treatment on one occasion and the other treatment on another occasion. Or the individual may be measured before some treatment or intervention and then measured again at a later time. In an observational study, blood pressures of women in late pregnancy are compared while at work and while at home. Here each women is measured twice under two conditions: while at work and while at home. Each individual serves as a block or the pair of time slots/occasions.

- **Sorting/Pairing**. Subjects/objects are paired according to some extraneous variable related to the response variable. The two persons in each pair are randomly assigned to the two treatments in the study. This is repeated for several pairs. In an experiment to compare two methods for learning difficult material subjects are paired according to academic ability and IQ. Each person within the pair is assigned at random to one of the methods. A score is obtained indicating the degree of learning. Each pair of individuals serves as a block.

- **Splitting**. Some experimental material, such as a volume of liquid or piece of cloth, is physically split into two portions. The two halves are randomly assigned to the two treatments and the response variable measured resulting in two samples of data. The two halves form a pair/block. In an experiment a batch of bread dough is prepared and then split into two halves. Both halves are put into the oven with one piece being baked for 30 minutes and the other for 40 minutes. The assignment of the amount of time of baking to the two halves is random. This process is repeated for several batches with different oven runs for the different batches. Each batch of dough serves as a block.

Let the $n$ random pairs or blocks of observations from a population of pairs be denoted by $(y_{11}, y_{21}), ..., (y_{1n}, y_{2n})$. As before it is assumed that the $y_1$ ' s are a random sample from a population with mean $\mu_1$ and variance $\sigma_1^2$. The $y_2$ ' s are a random sample from a population with mean $\mu_2$ and variance $\sigma_2^2$. Since the observations on the response are paired there may be a relationship between the two values within a pair. The degree of relationship between any two $y$'s in a pair is quantified in the population by the population **covariance**, $\sigma_{12}$ and population **correlation coefficient**, $\rho$. Readers may remember covariance and correlation from their elementary statistics class. The covariance between any two observations in a pair is a measure of the degree of linear relationship between the two observations. The correlation coefficient is a scaled version of the covariance which takes on values between -1 and 1 with values close to -1 or 1 indicating a strong linear relationship between the two variables.

The sample of $y_1$ ' s may be summarized with the sample mean and sample standard deviation $\overline{y}_{1.}$ and $s_1$. Similarly the sample of $y_2$ ' s may be summarized with $\overline{y}_{2.}$ and $s_2$. The degree of linear relationship between the two samples may be summarized with the **sample correlation coefficient,** $r$. The sample correlation coefficient has properties similar to the population correlation coefficient, $\rho$ described above.

The analysis for paired data is based on the differences

$$\begin{aligned} d_1 &= y_{11} - y_{21} \\ d_2 &= y_{12} - y_{22} \\ .. &= .. \\ .. &= .. \\ d_n &= y_{1n} - y_{2n} \end{aligned}$$

In theory the difference between the $y's$ is a comparison of the two treatments for each individual uninfluenced by the pairing or blocking variable since the pairing variable is roughly constant within the pair. Thus the analysis of the $d's$ should give a more precise comparison of the treatments than the comparison of treatments in a completely randomized design.

The sample of $d's$ is summarized by the sample mean

$$\overline{d} = (\sum_{i=1}^{n} d_i)/n = \overline{y}_{1.} - \overline{y}_{2.}$$

and standard deviation

$$s_d = \sqrt{\sum_{i=1}^{n}(d_i - \overline{d})^2/(n-1)}$$

The standard error of $\overline{d}$ is $s_{\overline{d}} = s_d/\sqrt{n}$.

Inferences are about the unknown "population" mean of differences, $\mu_d = \mu_1 - \mu_2$, the true effect or difference of the treatments.

If it is assumed that the $d's$ constitute a random sample from a normal population with mean $\mu_d$ and standard deviation $\sigma_d$ then we can use the one sample $t$ confidence interval and $t$ test from Chapter 2 to draw inferences about $\mu_d$ and thus conclusions about the differences in treatments or conditions.

The general form of the **paired/dependent samples confidence interval** for $\mu_d$ with a confidence level of $100(1-\alpha)\%$ is thus

$$\overline{d} \pm t_{\alpha/2;n-1} s_d/\sqrt{n}$$

The hypotheses for a two-sided test regarding $\mu_d$ are $H_o : \mu_d = \delta$ versus $H_a : \mu_d \neq \delta$. The hypotheses for the upper one-sided test are $H_o : \mu_d \leq \delta$ versus $H_a : \mu_d > \delta$. The hypotheses for the lower one-sided test are $H_o : \mu_d \geq \delta$ versus $H_a : \mu_d < \delta$. The value $\delta$ is taken to be 0 if the objective is to determine if there are any differences in the treatments.

The test statistic for a hypothesis test of $\mu_d$ is

$$t = \frac{\overline{d} - \delta}{s_{\overline{d}}} = \frac{\overline{d} - \delta}{s_d/\sqrt{n}}$$

where $\delta$ is a null hypothesized value for $\mu_d$, typically 0. P-values are based on the $t$ distribution with degrees of freedom $\nu = n - 1$, that is number of differences minus 1.

The paired samples procedures assume normality of the differences. The procedures do not assume that the standard deviations, $\sigma_1^2$ and $\sigma_2^2$, of $y$ in the populations are equal.

It can be shown that the variance of the $d's$ can be written as

$$s_d^2 = s_1^2 + s_2^2 - 2rs_1s_2$$

where $r$ is the sample correlation coefficient.

Thus the test statistic can be written as

$$\begin{aligned} t &= \frac{\overline{d} - \delta}{s_d/\sqrt{n}} \\ &= \frac{\overline{d} - \delta}{\sqrt{\frac{s_d^2}{n}}} \\ &= \frac{(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - \delta}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n} - \frac{(2rs_1s_2)}{n}}} \end{aligned}$$

Assuming equal group sizes and $\delta = 0$, the paired samples $t$ test statistic is the same as the test statistic for the independent samples $t$ test statistic of Equation 3.3,

$$t = \frac{\overline{y}_{1\cdot} - \overline{y}_{2\cdot} - (0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.4}$$

except for the adjustment of the standard error by $\frac{(2rs_1s_2)}{n}$, in the denominator. If the correlation coefficient, $r$, is positive, then the adjustment results in a reduction of the standard error. The reduction however may be offset by the loss in degrees of freedom for the paired samples $t$ test, $n-1$ compared to degrees of freedom for the independent samples $t$ test.

**Example 3.3** *One semester the author conducted an experiment in his 3 elementary statistics classes to determine if the ability to recall words was dependent on the type of word, concrete or abstract. Two lists of words, each of size 25, were constructed. List A had 25 concrete words, such as Bridge, Supermarket, Television; List B had more abstract words, such as Happiness, Government, Beauty. The entire set of words is given in Table 3.2. The two lists were constructed so that length of the words and familiarity were not much different. Each student studied both lists, in a random order, for two minutes, and then immediately wrote down the number of words that he or she recalled. The number of words recalled from each list by each student is given in Table 3.3 along with the difference in the numbers of words. Is there sufficient evidence that recall depends upon the type of word? Use a significance level of 0.05.*

In this example $\mu_d$ equals the "true" mean difference of numbers of words recalled (List A - List B) over a population of students. The value $\delta = 0$ so that $H_o : \mu_d = 0$ and the alternative is two sided with $H_a : \mu_d \neq 0$. The sample mean of the differences $\overline{d} = 0.05$ with standard deviation $s_d = 3.45$. Thus the observed value of the test statistic is

$$t = \frac{0.05 - 0}{3.45/\sqrt{60}} = 0.11$$

The P-value is determined from a computer program to be $P[|t| \geq |0.11|] = 0.9110$ based on a t distribution with $60 - 1 = 59$ degrees of freedom. Thus at the 0.05 level of significance there is no evidence of a difference in recall for the two types of words. Note that since sample size is large then normality of the population of differences is not necessary for the validity of the test result.

## 3.3 Connection between Two-Sided Tests and Confidence Intervals

In the two-sided tests described in this chapter using a significance level of $\alpha$ the null hypothesis of equality of two population means is rejected and the alternative of a difference in means is concluded if the $P-value \leq \alpha$. It can be shown in this case that a $100(1-\alpha)\%$ confidence interval for the difference $\mu_1-\mu_2$ will not contain zero, indicating that the two means are different, consistent

Table 3.2: Words Lists for Student Experiment

| List A: Concrete | List B: Abstract |
|---|---|
| Bridge | Happiness |
| Supermarket | Government |
| Bathroom | Reputation |
| Refrigerator | Beauty |
| Chocolate | Music |
| Screwdriver | Christmas |
| Lightning | Health |
| Bicycle | Time |
| Candle | Marriage |
| Sister | Magic |
| Baseball | Power |
| Spoon | Love |
| Apartment | Foolishness |
| Piano | Excitement |
| Underwear | Honesty |
| Microphone | Internet |
| Water | Religion |
| Chimpanzee | Fairness |
| Newspaper | Friendship |
| Television | Wealth |
| Mountain | Motivation |
| Honeybee | Inflation |
| Highway | Jealousy |
| Rainbow | Anger |
| Eyeglasses | Competition |

Table 3.3: Number of Words Recalled out of 25

| Student | List A | List B | Difference |
|---|---|---|---|
| 1 | 18 | 17 | 1 |
| 2 | 20 | 19 | 1 |
| 3 | 20 | 16 | 4 |
| 4 | 15 | 14 | 1 |
| 5 | 17 | 11 | 6 |
| 6 | 16 | 19 | -3 |
| 7 | 13 | 14 | -1 |
| 8 | 22 | 21 | 1 |
| 9 | 18 | 17 | 1 |
| 10 | 16 | 14 | 2 |
| 11 | 15 | 19 | -4 |
| 12 | 13 | 14 | -1 |
| 13 | 12 | 15 | -3 |
| 14 | 21 | 16 | 5 |
| 15 | 13 | 12 | 1 |
| 16 | 20 | 14 | 6 |
| 17 | 18 | 15 | 3 |
| 18 | 7 | 10 | -3 |
| 19 | 16 | 23 | -7 |
| 20 | 13 | 14 | -1 |
| 21 | 19 | 22 | -3 |
| 22 | 10 | 19 | -9 |
| 23 | 18 | 15 | 3 |
| 24 | 11 | 13 | -2 |
| 25 | 14 | 13 | 1 |
| 26 | 24 | 21 | 3 |
| 27 | 16 | 16 | 0 |
| 28 | 12 | 13 | -1 |
| 29 | 12 | 12 | 0 |
| 30 | 17 | 15 | 2 |
| 31 | 17 | 22 | -5 |
| 32 | 15 | 16 | -1 |
| 33 | 20 | 19 | 1 |
| 34 | 21 | 22 | -1 |
| 35 | 19 | 17 | 2 |
| 36 | 19 | 21 | -2 |
| 37 | 15 | 18 | -3 |
| 38 | 12 | 10 | 2 |
| 39 | 20 | 12 | 8 |
| 40 | 17 | 19 | -2 |
| 41 | 17 | 16 | 1 |
| 42 | 21 | 19 | 2 |
| 43 | 16 | 15 | 1 |
| 44 | 14 | 14 | 0 |
| 45 | 16 | 16 | 0 |
| 46 | 16 | 18 | -2 |
| 47 | 20 | 13 | 7 |
| 48 | 17 | 15 | 2 |
| 49 | 17 | 17 | 0 |
| 50 | 13 | 12 | 1 |
| 51 | 18 | 12 | 6 |
| 52 | 16 | 20 | -4 |
| 53 | 19 | 17 | 2 |
| 54 | 11 | 17 | -6 |
| 55 | 15 | 18 | -3 |
| 56 | 22 | 25 | -3 |
| 57 | 17 | 13 | 4 |
| 58 | 12 | 11 | 1 |
| 59 | 18 | 19 | -1 |
| 60 | 12 | 19 | -7 |

Table 3.4: Power of One-Sided Two Sample t test

| | Common Sample Size n | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $E$ | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| 0.5 | 0.179 | 0.285 | 0.379 | 0.463 | 0.539 | 0.606 | 0.665 | 0.716 | 0.761 | 0.799 |
| 0.6 | 0.219 | 0.362 | 0.483 | 0.587 | 0.672 | 0.743 | 0.780 | 0.845 | 0.881 | 0.909 |
| 0.7 | 0.264 | 0.445 | 0.589 | 0.702 | 0.787 | 0.850 | 0.895 | 0.928 | 0.951 | 0.966 |
| 0.8 | 0.313 | 0.530 | 0.689 | 0.799 | 0.874 | 0.922 | 0.952 | 0.971 | 0.983 | 0.990 |
| 0.9 | 0.366 | 0.615 | 0.776 | 0.875 | 0.932 | 0.964 | 0.981 | 0.990 | 0.995 | 0.998 |
| 1.0 | 0.421 | 0.694 | 0.848 | 0.928 | 0.967 | 0.985 | 0.994 | 0.997 | 0.999 | 1.000 |
| 1.1 | 0.478 | 0.764 | 0.902 | 0.962 | 0.986 | 0.995 | 0.998 | 0.999 | 1.000 | 1.000 |
| 1.2 | 0.536 | 0.825 | 0.941 | 0.981 | 0.994 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 |
| 1.3 | 0.592 | 0.875 | 0.966 | 0.992 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 1.4 | 0.647 | 0.914 | 0.914 | 0.982 | 0.997 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 |
| 1.5 | 0.698 | 0.943 | 0.991 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

with the results of the test. Similarly if the null hypothesis is not rejected ($P-value > \alpha$), implying that the two means could possibly be the same, then the $100(1-\alpha)\%$ confidence interval will contain 0, again consistent with the test. Thus the confidence interval could be used to perform a two-sided test. Additionally the confidence interval provides information about the magnitude of differences between the means.

## 3.4 Power of the One-Sided Pooled Two Sample $t$ test

In this section power calculations will be given for the one-sided two independent pooled samples t test (assuming equal population variances) under certain alternatives. Consider the test of the null hypothesis $H_o : \mu_1 - \mu_2 \leq 0$ versus the alternative $H_a : \mu_1 - \mu_2 > 0$. Suppose the alternative hypothesis is true. Let $E$ be defined as the absolute difference $|\mu_1 - \mu_2|$ in numbers of the common standard deviation $\sigma$, i.e.

$$E = |\mu_1 - \mu_2|/\sigma$$

Table 3.4 gives power for various values of $E$, common sample size $n$, and significance level $\alpha = 0.05$.

Suppose that an educational researcher believes that a new method of teaching reading will increase reading scores by as much as 10 points compared to a standard method. Variability of reading scores for the standard method has been about 15 points. The researcher will conduct an experiment comparing the two methods with two equal sized groups of students. The researcher believes that variability will be about the same in the two groups and will use the pooled two sample $t$ test to compare reading scores for the two groups. The researcher would like at least a 90% chance of concluding that the new method is better assuming the parameters above are correct.

Thus $E = 10/15 = 0.7$. From Table 3.4 it is concluded that the researcher needs 40 students in each group to achieved a power of 92.8%.

## 3.5 SAS Code

### 3.5.1 Example 3.1

```
*  The following data step inputs the weights for each
   of the treated and control animals;
data WEIGHTS;
  input Treatment $ WeightGain;
datalines;
Control 16.4
Control 12.8
Control 13.0
Control 10.7
Control 3.9
Control 9.1
Control 8.7
Control 9.5
Control 8.5
Control 6.0
Control 9.0
Control 13.4
Control 3.4
Control 9.6
Control 14.4
Control 11.3
Control 6.8
Control 2.3
Treated 11.6
Treated 8.9
Treated 14.6
Treated 12.4
Treated 13.3
Treated 16.0
Treated 11.1
Treated 15.8
Treated 15.6
Treated 10.7
Treated 12.4
Treated 14.6
Treated 11.2
Treated 10.7
Treated 11.6
Treated 14.7
Treated 13.9
Treated 11.8
Treated 13.4
Treated 12.1
Treated 13.4
Treated 12.5
;
run;
* The proc means calculates descriptive statistics
  associated with the two groups;
proc means data = WEIGHTS;
  class treatment;
  var weightgain;
run;
* The proc ttest does the necessary calculations necessary
  for an independent samples t test and a confidence interval;
proc ttest data = WEIGHTS;
  class treatment;
  var weightgain;
run;
```

### 3.5.2 Example 3.2

```
*  The following data step inputs the scores for each
   of the children getting the new and standard methods;
data READING;
  input Method $ Score;
datalines;
New 80
New 76
New 70
New 80
New 66
New 85
New 79
New 71
New 81
New 76
Standard 79
Standard 73
Standard 72
Standard 62
Standard 76
Standard 68
Standard 70
Standard 86
Standard 75
Standard 68
Standard 73
Standard 66
;
run;
* The proc means calculates descriptive statistics
  associated with the two methods;
proc means data = READING;
  class Method;
  var Score;
run;
* The proc ttest does the necessary calculations
  for an independent samples t test and a confidence interval;
proc ttest data = READING;
  class Method;
  var Score;
run;
```

### 3.5.3 Example 3.3

```
*  Input the number of words recalled by each student
   from List A and List B;
data WORDLIST;
   input Student NumWordsA  NumWordsB;
datalines;
1       18      17
2       20      19
3       20      16
4       15      14
5       17      11
6       16      19
7       13      14
8       22      21
9       18      17
10      16      14
11      15      19
12      13      14
13      12      15
14      21      16
15      13      12
16      20      14
17      18      15
18      7       10
19      16      23
20      13      14
21      19      22
22      10      19
23      18      15
```

```
24      11      13
25      14      13
26      24      21
27      16      16
28      12      13
29      12      12
30      17      15
31      17      22
32      15      16
33      20      19
34      21      22
35      19      17
36      19      21
37      15      18
38      12      10
39      20      12
40      17      19
41      17      16
42      21      19
43      16      15
44      14      14
45      16      16
46      16      18
47      20      13
48      17      15
49      17      17
50      13      12
51      18      12
52      16      20
53      19      17
54      11      17
55      15      18
56      22      25
57      17      13
58      12      11
59      18      19
60      12      19
;
run;
*  Use proc ttest to do the necessary calculations to
   perform a paired samples t test and to obtain a
   confidence interval for the difference in means;
proc ttest data = WORDLIST;
  paired NumWordsA * NumWordsB;
run;
```

# Problems for Chapter 3

3.1* A researcher tested two new fertilizers for growing tomatoes. One fertilizer, A, was a fertilizer that was used for two years and the other, B, was a new fertilizer being tested for the first time. Sixteen tomato plants of the same variety and about the same size were planted in a garden in a 4x4 rectangular fashion with the plants being about 6 feet apart. The sixteen plants were randomly assigned to their plots and the two fertilizers were randomly assigned to the plant/plot combination with eight plants receiving each of the two fertilizers. The total amount of tomatoes in pounds from each plant for the two different fertilizers was measured.

a. What is the factor of interest? What is the response variable?

b. What are the experimental units?

c. Is this a completely randomized design or a paired design? Explain.

d. What are some extraneous variables? How are these controlled?

3.2* A student in an experimental design class wanted to see if there was a difference in the amount of time (in minutes) that scented candles burned as compared with non-scented candles. She bought twenty candles which appeared to be the same except ten were scented and ten were unscented. She could not burn all candles in the same day so she decided to burn a pair of candles, one scented and one unscented per day at roughly the same time of the day, for ten days. The same two locations in a room were used for all days. On each day she randomly selected one scented and one unscented candle. These two were then randomly assigned to location and initial lighting. The data are given in the table below.

| Test | Scented | Unscented |
|---|---|---|
| 1 | 680 | 696 |
| 2 | 752 | 697 |
| 3 | 818 | 750 |
| 4 | 793 | 774 |
| 5 | 771 | 672 |
| 6 | 744 | 676 |
| 7 | 798 | 782 |
| 8 | 678 | 777 |
| 9 | 742 | 762 |
| 10 | 763 | 703 |

a. This is a paired design. Explain.

b. What is the factor? What is the response variable?

c. Is there evidence that scented candles of this kind have different mean burning times than unscented candles. Use a significance level of 0.05. Use statistical software to obtain a P-value and look at a histogram of differences to check the normality assumption.

3.3* Identify the experimental design in the following studies as either being completely randomized or paired/blocked. If the design is a paired design, then identify the type of pairing (re-using, sorting/grouping or splitting).

a. In a study of the effect of a diet for reducing weight, the weights of ten subjects are measured both before and after being put on the diet for five weeks.

b. In a study of flirtatious behavior, sixty male students were given false information about female job applicants. Thirty of the male students selected at random were falsely told that the female applicant was attracted to her interviewer and the other thirty were not told such false information. The men in both groups were asked if the female exhibited any flirtatious behavior on the phone.

c. In order to determine whether the zipcode+4 gets a letter faster to its destination than just the zipcode, a student data project mailed two letters to each of twenty-six cities. The letters/envelopes were the same except that one had the 5 digit zipcode on it while the other letter had the zipcode+5 digits on it.

3.4* Approximately 200 patients with Alzheimer's disease were measured for mental ability before and after being given 120 mg to 240 mg of ginkgo biloba, a plant extract, daily for three to six months.

a. What are the conditions of interest to be compared?

b. What are the "experimental" units?

c. Are the conditions assigned to the units? Explain

d. What are the consequences of your response to part (c) on the interpretation of the results of the study?

3.5* A trucking firm wishes to choose between two alternate routes for transporting merchandize from one depot to another. One major concern is the travel time. In a study, 5 drivers are randomly assigned to route A, the other 5 were assigned to route B. Data was obtained from each driver on travel time (hours) and given below.

Route A: 18 24 30 21 32

Route B: 22 29 34 25 35

a. What is the factor in this experiment?

b. What is the response variable?

c. Give one extraneous variable whose effects are potentially balanced out by the randomization.

d. Based on the data, is there evidence of a difference in driving time between the two routes? Use the independent samples t test that does not assume anything about the population variances. Use computer software to obtain a P-value. Use a significance level of 0.05.

e. Describe an alternate design for this experiment that would use pairing/blocking.

3.6* In the article "Feeding Preferences of Captive Tassel-Eared Squirrels (*Sciurus Aberti*) for Ponderosa Pine Twigs"(*Journal of Mammalogy [1980]: 734-737*) researchers wanted to determine in a laboratory setting if squirrels could distinguish between twigs from known feeding trees (FT) and nonfeeding trees (FT). The feeding trees and nonfeeding trees were determined in the field by the extent of defoliation and amounts of clipped needles. Squirrels in the field presumably eat from certain Ponderosa pines depending on nutritional quality, the occurrence of certain plant compounds in the tree, pheromonal cures and other contextual factors. Each of five squirrels was tested for preference on 6 different days at two-week intervals. A testing consisting of providing a squirrel with one FT twig and one NFT twig and then measuring the amount of the twig eaten after 24 hours. The data below provide the mean amounts eaten by each of the six squirrels over the 6 day treatment (The data were approximated from a bar graph in the article).

| Squirrel | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| FT | 5.5 | 4.4 | 6.0 | 4.8 | 8.4 |
| NFT | 3.2 | 2.4 | 3.2 | 4.4 | 1.7 |

a. Is this an independent samples or paired samples design? Explain.

b. Based on the data, is there evidence that squirrels are attracted to the twigs that come from the FTs. Use statistical software to obtain a P-value. Use a significance level of 0.05.

3.7* The article "Operational Plantations of Improved Slash Pine: Age 15 results" (http://www.rngr.net/Publications/sftic/1983/)compared “improved” and “unimproved” slash pines in terms of volume production and fusiform rust infection after 15 years of planting. The “improved” trees were grown from seeds taken from parents selected for volume production, crown and bole characteristics, and disease resistance. The data below are based on two stands of slash pines, one with improved (I) and one with unimproved (U) trees at each of the 10 locations.

| Location | Seed Source | Vol/Acre $(ft^3)$o.b. | Vol/Acre $(ft^3)$i.b. | Fusiform % |
|---|---|---|---|---|
| Appling Co | I | 1677 | 1115 | 27.3 |
| | U | 1792 | 1181 | 15.4 |
| Atkinson Co | I | 2248 | 1535 | 16.8 |
| | U | 2041 | 1355 | 13.9 |
| Ben Hill Co | I | 849 | 529 | 28.9 |
| | U | 937 | 570 | 12.9 |
| Camden Co | I | 2252 | 1534 | 12.0 |
| | U | 2102 | 1402 | 8.2 |
| Laurens Co | I | 905 | 592 | 60.4 |
| | U | 1243 | 794 | 73.8 |
| Long Co | I | 849 | 534 | 16.1 |
| | U | 994 | 625 | 15.2 |
| Toombs Co | I | 1076 | 707 | 45.2 |
| | U | 874 | 546 | 41.3 |
| Ware Co | I | 1760 | 1171 | 5.8 |
| | U | 1734 | 1135 | 6.7 |
| Wheeler Co | I | 447 | 282 | 35.9 |
| | U | 577 | 358 | 33.0 |
| Wayne Co | I | 1466 | 959 | 13.2 |
| | U | 1350 | 862 | 10.6 |

a. This is a paired samples design. Explain.

b. Based on the data, is there evidence that "improved" trees have higher inner bark (i.b.) volume per acre. Use a significance level of 0.05.

3.8* Researchers studied the the maximum voluntary closing forces (in newtons, N) of the upper and lower lips for 15 young male and 15 young female subjects ("Maximum Voluntary Closing Forces in the Upper and Lower Lips of Humans",*Journal of Speech and Hearing Research,* Volume 28, 373-376, 1985). Each subject was measured 5 times on the upper lip and 5 times on the lower lip. The table below gives means of the 5 upper and lower lip measures approximated from a scatterplot of the data given in the article.

| Males | | | Females | | |
|---|---|---|---|---|---|
| Subject | Lower Lip | Upper Lip | Subject | Lower Lip | Upper Lip |
| 1 | 19.5 | 3.0 | 1 | 7.0 | 2.5 |
| 2 | 7.5 | 5.5 | 2 | 13.0 | 3.0 |
| 3 | 17.0 | 6.0 | 3 | 11.0 | 4.5 |
| 4 | 11.5 | 2.0 | 4 | 8.0 | 4.0 |
| 5 | 13.0 | 4.0 | 5 | 4.5 | 2.0 |
| 6 | 16.0 | 3.5 | 6 | 12.0 | 4.5 |
| 7 | 15.0 | 5.0 | 7 | 9.0 | 4.0 |
| 8 | 11.0 | 4.0 | 8 | 5.5 | 3.5 |
| 9 | 10.0 | 5.5 | 9 | 8.5 | 3.0 |
| 10 | 19.0 | 4.0 | 10 | 5.5 | 1.0 |
| 11 | 13.5 | 5.0 | 11 | 13.0 | 3.0 |
| 12 | 9.0 | 5.0 | 12 | 7.0 | 3.0 |
| 13 | 17.0 | 5.5 | 13 | 11.0 | 4.0 |
| 14 | 10.0 | 4.5 | 14 | 6.0 | 3.0 |
| 15 | 21.0 | 3.5 | 15 | 11.0 | 4.0 |

a. i. Construct a plot of lower lip force versus gender. Comment on any difference in average and spread.

ii. Is there evidence of a difference in mean lower lip force between young males and females? Use statistical software to perform an independent samples t test (do not assume anything about population variances). Use a significance level of 0.05.

iii. Why is the independent samples t test more appropriate than the paired samples t test?

b. Give an another example of a comparison involving the data for which the independent samples t test would be appropriate. Give an example of a comparison for which the paired samples t test would be appropriate.

3.9* In one part of a research study ("Glass Shape Influences Consumption Rate for Alcoholic Beverages," PloS ONE 7 (2012) e43007)). each of 160 participants (50 % male) viewed computer images of both straight and curved 12 fl oz glasses, the activity resulting in a numerical value reflecting what the participant perceived to represent half full for either type of glass. The true value of a half full glass (straight or curved) was 30 with the perceived midpoints in volume given by the participants varying from the true value. The authors gave the following summary;

"One-sample t-tests against a test value of 30 indicated that for both straight (M = 28, SD = 2, t[159] = -16.91, $p < 0.001$) and curved (M=21, SD = 3, t[159] = -37.97, $p < 0.001$) glasses the half-way point was perceived to be below the true half-way point. A paired-sample t-test indicated a significant difference between the two glass conditions (t[159] = 30.89, $p < 0.001$).

In the above paragraph $M$ refers to a sample mean.

a. Give null and alternative hypotheses corresponding to the results of the paired samples t-test. Assume a two-sided test. Be sure to give an appropriate symbol and describe in words what that symbol means.

b. Is it possible to do the calculation of the t test statistic by hand given the information provided? Explain.

c. Using a significance level of 0.05 what is the conclusion that can be drawn regarding the two glasses and perceived midpoints in volume?

3.10* An experiment is conducted to test the effectiveness of a diet supplement to increase the weight gain of chicks as compared to a control of no supplement. A number of chicks will be fed with the supplement in their diet and the same number of chicks will be fed the regular diet. An improvement in weight gain of on average 50 grams during the test period is considered practically important. Based on other studies similar to the current one, the standard deviation of weight gains is believed to be about 100 grams for both the diet supplement and the control groups and the weight gains should be normally distributed. The two independent samples t test, assuming equal population standard deviations, will be used to formally compare the two diets, using a significance level of 0.05. The power for detecting an increase in weight gain of 50 grams should be about 80%. What group sizes would you recommend?

3.11* The American Statistical Association holds an annual poster and project competition for students from grades K-12. Winners receive a monetary award and a plaque. One of the winners in the 2013 competition conducted an experiment to answer the question: Do dryer ball reduce drying time? The student conducted the experiment in response to an ad that claimed that balls "reduce drying time by up to 25%. The student randomly assigned the next 40 of his family's wash loads to either be dried with dryer balls added to the dryer or not. Only one washer and dryer was used. The student weighed each load prior to drying and recorded how long it took the load to dry (in minutes) using a stop watch. The dryer has a sensor that detects when the clothes are dry. The drying times are provided in the table below.

| Dryer Balls | No Dryer Balls |
|---|---|
| 34.5 | 23.5 |
| 28.0 | 24.5 |
| 28.5 | 22.5 |
| 24.0 | 21.0 |
| 24.0 | 29.0 |
| 40.0 | 21.0 |
| 21.5 | 34.0 |
| 29.0 | 36.0 |
| 34.0 | 34.0 |
| 30.0 | 24.0 |
| 33.0 | 31.0 |
| 22.0 | 26.0 |
| 31.0 | 21.0 |
| 22.0 | 41.0 |
| 22.0 | 32.0 |
| 29.0 | 36.0 |
| 21.0 | 39.0 |
| 33.0 | 29.0 |
| 35.0 | 22.0 |
| 24.5 | 21.0 |

Use statistical software where appropriate.

a. Construct a dotplot of the drying times versus dryer ball or not. Compare the two groups in terms of average level and variation.

b. Determine the sample mean and standard deviation of the drying times for the two groups.

c. Compare the two groups of drying times using an appropriate one-sided t test. Give null and alternative hypotheses. Give the value of the test statistic with a P-value and draw a conclusion. Use a significance level $\alpha = 0.05$.

3.12 In the article "Sound Level of Environmental Music and Drinking Behavior: A Field Experiment with Beer Drinkers"( *Alcoholism: Clinical and Experimental Research*, Vol 32, No. 10, 20082: 1-4 ) researchers compared number of drinks (beer) ordered, time spent to drink a glass (minutes), and number of gulps per drink for two groups of patrons at a bar under two experimental conditions. Unbeknownst to the patrons twenty of them were assigned at random to listen to background music at the usual sound level. Another twenty listened to the higher sound level. Means and standard deviations are reported in the table below.

| Level of environmental music | Mean (standard deviation)<br>Number of drinks ordered | Time spent to drink a glass (in minutes) | Number of gulps per drink |
|---|---|---|---|
| Usual level | 2.6 (1.14) | 14.51 (4.88) | 7.02 (1.26) |
| High level | 3.4 (0.99) | 11.45 (2.89) | 7.18 (1.29) |

The authors reported that "the difference between the 2 experimental conditions was tested by the help of an unpaired t-test." They reported the following result for the comparison of the two groups on time spent to drink a glass: $t(38, two - dailed) = 2.36, P < 0.03$. The value of 38 refers to degrees of freedom. The value of 2.36 is the value of the test statistic for the unpaired t-test. The article did not indicate how the variable time spent to drink a glass was recorded for data analysis for the t test if a patron ordered more than one drink.

i. What are the treatments in this study?

ii. What are the experimental units? Can we tell how many experimental units there are based on the information reported above?

iii. What are the measurement units in the study? How many are there? Explain.

iv. Suppose that the authors used the pooled two sample t test as their unpaired t test. Confirm their calculation by calculating the test statistic using the summaries provided above.

v. State null and alternative hypotheses for the test conducted. What is the conclusion based on the reported P-value?

vi. The article stated that subjects were observed sitting at tables in the bar and only tables with 2 patrons were used for data collection with at least one of the patrons ordering a beer. Some tables provided data on one patron and other tables provided data on 2 patrons. Why does the data collection procedure raise questions about the use of the unpaired t test?

# Chapter 4

# Analysis for the One Factor Completely Randomized Design

## 4.1 Decomposing Data

A medical researcher on aging is studying the effects of diet on the longevity of mice. Twelve mice were randomly assigned to one of three different diets with four mice assigned to each diet. Thus the design is completely randomized. There is only one factor of interest. The diets along with the lifelengths (in months) of the mice are given below. A dotplot of the data is given in Figure 4.1.

| | | | | |
|---|---|---|---|---|
| Diet 1 (High Calorie) | 22 | 18 | 21 | 22 |
| Diet 2 (Medium Calorie) | 20 | 19 | 23 | 21 |
| Diet 3 (Low Calorie) | 23 | 24 | 20 | 25 |

In general suppose there are $t$ treatments and $n_i$ observations on the response $y$ for the $i^{th}$ treatment where $i = 1, ...t$. In the example above $t = 3$, where $i = 1$ corresponds to Diet 1, $i = 2$ corresponds to Diet 2, and $i = 3$ corresponds to Diet 3. Also $n_1 = 4$, $n_2 = 4$, and $n_3 = 4$. Often treatment group sizes are the same or similar. Let $N = \sum_{i=1}^{t} n_i$ be the total number of observations on $y$. In this example $N = \sum_{i=1}^{3} n_i = n_1 + n_2 + n_3 = 4 + 4 + 4 = 12$. For each treatment let $y_{ij}$ denote the $j^{th}$ observation on treatment $i$. In this example, $y_{11} = 22$, $y_{32} = 24$, etc.

The goal in this chapter is to develop a hypothesis test to determine if the observed differences in longevity among the three diets are "real" or could be explained by random extraneous variables, that is error, such as genetics, stress factors, etc. The hypothesis test will be based on a **decomposition of the data** into parts that reflect the contributions of those parts to the variation in the lifelengths.

Figure 4.1: Plot of Lifelength versus Diet

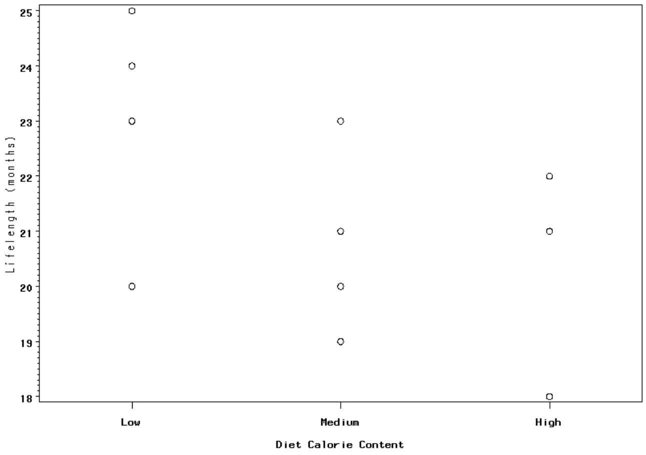


Table 4.1: Descriptives for Mice Data

| Group | Mean | St.Dev. |
|---|---|---|
| Diet 1 (High Calorie) | $\overline{y}_{1.} = 20.75$ | $s_1 = 1.89$ |
| Diet 2 (Medium Calorie) | $\overline{y}_{2.} = 20.75$ | $s_2 = 1.71$ |
| Diet 3 (Low Calorie) | $\overline{y}_{3.} = 23.00$ | $s_3 = 2.16$ |
| All | $\overline{y}_{..} = 21.50$ | $s = 2.07$ |

Table 4.2: Errors for Mice Receiving Diet 3

$$\begin{aligned} e_{31} &= y_{31} - \overline{y}_{3\cdot} = 23 - 23 = 0 \\ e_{32} &= y_{32} - \overline{y}_{3\cdot} = 24 - 23 = 1 \\ e_{33} &= y_{33} - \overline{y}_{3\cdot} = 20 - 23 = -3 \\ e_{34} &= y_{34} - \overline{y}_{3\cdot} = 25 - 23 = 2 \end{aligned}$$

The parts will be based on means and deviations from means. Table 4.1 provides the means and standard deviations for each diet group and the combined groups. The mean $\overline{y}_{\cdot\cdot} = 21.50$ of all 12 lifespans is called the **grand mean**.

Now consider a particular diet, say Diet 3, the low calorie diet. While all mice received Diet 3 the observed lifelengths differed, presumably because of extraneous variables such as genetics, weight, etc. The effects of these extraneous variables, called **"errors"** and denoted by $e$ for these mice are measured by the differences between lifelengths for Diet 3 mice and the mean for Diet 3, 23.00. The errors for all 4 mice receiving Diet 3 are given in Table 4.2.

In a first course in statistics errors were likely called **deviations from the mean** and form the basis for a sample variance/standard deviation of a set of data. In fact the sample variance (see Chapter 2) for the four lifelengths for the mice receiving Diet 3 would be $s_3^2 = \sum_{j=1}^{n_3}(y_{3j} - \overline{y}_{3\cdot})^2/(n_3 - 1)$, based on the what are now being called errors.

The term "error" does not mean that something is wrong with a mouse; it is simply a reflection of the fact that all animals getting the same Diet will still vary in lifelength due to other uncontrolled extraneous variables. For example, the error for the second animal receiving Diet 3 is 1 month. The particular extraneous variables associated with this animal resulted in a lifelength which was 1 month higher than the average lifelength of all animals receiving the same Diet 3. Note that if all lifelengths for mice on Diet 3 had been the same then the errors would all be 0. Larger errors in magnitude reflect greater effects of extraneous variables as compared to smaller errors. Note that the sum of the errors for all four mice receiving Diet 3 is $0 + 1 + -3 + 2 = 0$. In general the differences between a group of values and their mean will equal to 0.

The errors for the other groups are calculated similarly by subtracting from the lifelengths of mice the mean of the group to which the mice belong. The errors for groups 1 and 2 are given below.

Table 4.3: Decomposition of Lifelength Data: Group Mean + Error

| | $y_{ij}$ | $=$ | $\overline{y}_{i\cdot}$ | $+$ | $e_{ij}$ |
|---|---|---|---|---|---|
| Diet 1 | 22 | $=$ | 20.75 | $+$ | 1.25 |
| | 18 | $=$ | 20.75 | $+$ | $-2.75$ |
| | 21 | $=$ | 20.75 | $+$ | 0.25 |
| | 22 | $=$ | 20.75 | $+$ | 1.25 |
| Diet 2 | 20 | $=$ | 20.75 | $+$ | $-0.75$ |
| | 19 | $=$ | 20.75 | $+$ | $-1.75$ |
| | 23 | $=$ | 20.75 | $+$ | 2.25 |
| | 21 | $=$ | 20.75 | $+$ | 0.25 |
| Diet 3 | 23 | $=$ | 23.00 | $+$ | 0.00 |
| | 24 | $=$ | 23.00 | $+$ | 1.00 |
| | 20 | $=$ | 23.00 | $+$ | $-3.00$ |
| | 25 | $=$ | 23.00 | $+$ | 2.00 |

Diet 1

$$\begin{aligned}
e_{11} &= y_{11} - \overline{y}_{1\cdot} = 22 - 20.75 = 1.25 \\
e_{12} &= y_{12} - \overline{y}_{1\cdot} = 18 - 20.75 = -2.75 \\
e_{13} &= y_{13} - \overline{y}_{1\cdot} = 21 - 20.75 = 0.25 \\
e_{14} &= y_{14} - \overline{y}_{1\cdot} = 22 - 20.75 = 1.25
\end{aligned}$$

Diet 2

$$\begin{aligned}
e_{21} &= y_{31} - \overline{y}_{2\cdot} = 20 - 20.75 = -0.75 \\
e_{22} &= y_{32} - \overline{y}_{2\cdot} = 19 - 20.75 = -1.75 \\
e_{23} &= y_{33} - \overline{y}_{2\cdot} = 23 - 20.75 = 2.25 \\
e_{24} &= y_{34} - \overline{y}_{2\cdot} = 21 - 20.75 = 0.25
\end{aligned}$$

Note that by solving for lifelength in the definition of error we have a decomposition of lifelength. Consider Diet 3 again.

$$\begin{array}{rcc}
y_{31} = \overline{y}_{3\cdot} + e_{31} & \textit{or} & 23 = 23 + 0 \\
y_{32} = \overline{y}_{3\cdot} + e_{32} & \textit{or} & 24 = 23 + 1 \\
y_{33} = \overline{y}_{3\cdot} + e_{33} & \textit{or} & 20 = 23 + (-3) \\
y_{34} = \overline{y}_{3\cdot} + e_{34} & \textit{or} & 25 = 23 + (2)
\end{array}$$

Thus each observed value of lifelength can be expressed as the Diet 3 mean lifelength plus the “effect” due to the other uncontrollable variables.

Table 4.3 gives the entire decomposition of the 12 lifelengths in terms of group mean and error. The set of 12 equations is referred to as the **sample means model** for the life lengths data.

Table 4.4: Decomposition of Lifelength Data

| | $y_{ij}$ | = | $\overline{y}_{..}$ | + | $A_i$ | + | $e_{ij}$ |
|---|---|---|---|---|---|---|---|
| Diet 1 | 22 | = | 21.50 | + | -0.75 | + | 1.25 |
| | 18 | = | 21.50 | + | -0.75 | + | -2.75 |
| | 21 | = | 21.50 | + | -0.75 | + | 0.25 |
| | 22 | = | 21.50 | + | -0.75 | + | 1.25 |
| Diet 2 | 20 | = | 21.50 | + | -0.75 | + | -0.75 |
| | 19 | = | 21.50 | + | -0.75 | + | -1.75 |
| | 23 | = | 21.50 | + | -0.75 | + | 2.25 |
| | 21 | = | 21.50 | + | -0.75 | + | 0.25 |
| Diet 3 | 23 | = | 21.50 | + | 1.50 | + | 0.00 |
| | 24 | = | 21.50 | + | 1.50 | + | 1.00 |
| | 20 | = | 21.50 | + | 1.50 | + | -3.00 |
| | 25 | = | 21.50 | + | 1.50 | + | 2.00 |

There is one more step in the decomposition process. The effect of diet, denoted with the letter $A$, will be measured by comparing the mean lifelength for a diet to the grand mean of all lifelengths. The effects for the three diets, $A_1, A_2, A_3$, are calculated as follows:

$$\begin{aligned} A_1 &= \overline{y}_{1.} - \overline{y}_{..} = 20.75 - 21.50 = -0.75 \\ A_2 &= \overline{y}_{2.} - \overline{y}_{..} = 20.75 - 21.50 = -0.75 \\ A_3 &= \overline{y}_{3.} - \overline{y}_{..} = 23.00 - 21.50 = 1.50 \end{aligned}$$

Thus diet 3 has the "effect" of raising lifelength by 1.50 months compared to the grand mean of 21.50 months. Diets 1 and 2 have the same effects, that is of lowering lifelength compared to the grand mean. In general effects can all be different, but note that the effects add to 0.

Using the effect definition the three diet group means can be decomposed as follows:

$$\begin{array}{lcl} \overline{y}_{1.} = \overline{y}_{..} + A_1 & \textit{or} & 20.75 = 21.50 + (-0.75) \\ \overline{y}_{2.} = \overline{y}_{..} + A_2 & \textit{or} & 20.75 = 21.50 + (-0.75) \\ \overline{y}_{3.} = \overline{y}_{..} + A_3 & \textit{or} & 23.00 = 21.50 + (1.50) \end{array}$$

Replacing each diet group mean in Table 4.3 with its decomposition results in Table 4.4 where each lifelength is now written in terms of a sum of grand mean, diet effect, and error (effect of extraneous variables). This completes the decomposition of the lifelength data.

The 12 equations in Table 4.4 can be expressed symbolically as

$$y_{ij} = \overline{y}_{..} + A_i + e_{ij} \tag{4.1}$$

where $i = 1, 2, 3$; $j = 1, 2, 3, 4$; $A_i = \overline{y}_{i.} - \overline{y}_{..}$; and $e_{ij} = y_{ij} - \overline{y}_{i.}$.

## 4.2 Degrees of Freedom

A concept associated with the decomposition in the last section and the analysis in subsequent sections is **degrees of freedom**. A set of numbers or as we shall see shortly a sum of squared numbers is said to have a certain "number of degrees of freedom" associated with them. For example, the 12 values for lifelength in the decomposition Table 4.4 have "12 degrees of freedom" because the 12 values can be almost anything, that is all 12 are free to vary–there are no mathematical restrictions on them. The 12 grand means have only "1 degree of freedom" since they all have to be the same number, because of the way they were calculated. The 12 diet/treatment effects in the decomposition table have 2 degrees of freedom because of the repetitiveness within each diet and the fact that they all add to 0. Only two of the 12 treatment effects are free to vary. The other 10 can be determined by repeating and the restriction that they add to zero. The 12 "errors" in the decomposition table have 9 degrees of freedom. This is so because within each diet only 3 of the errors are free to vary–once we know 3, we can get the $4^{th}$ since the errors for a particular diet have to add to zero.

The degrees of freedom (df) are additive, that is

$$df\ for\ data = df\ for\ grand\ mean + df\ for\ treatment\ effects\ + df for\ errors$$

or

$$N = 1 + (t-1) + (N-t)$$

In our example,

$$12 = 1 + 2 + 9$$

## 4.3 Population Models

The set of equations

$$y_{ij} = \overline{y}_{i\cdot} + e_{ij}$$

for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$ given in Section 4.1 is referred to as the **sample means model**. The set of equations that uses the sample effects $A_i$,

$$y_{ij} = \overline{y}_{\cdot\cdot} + A_i + e_{ij}$$

is referred to as the **sample effects model**. Both models are sample based, that is based on the observed samples of data.

The **population means model** for the lifelength data refers to the set of 12 equations:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$.

The **population effects model** refers to the set of 12 equations:

$$y_{ij} = \overline{\mu}. + \alpha_i + \epsilon_{ij}$$

for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$.
In the above equations,

- $\mu_i$ = the population or true mean longevity for the $i^{th}$ diet
- $\overline{\mu}.$ = the population or true grand mean = $(\sum_{i=1}^{t} \mu_i)/t$
- $\alpha_i = \mu_i - \overline{\mu}.$ = population or true effect of the $i^{th}$ diet on longevity
- $\epsilon_{ij} = y_{ij} - \mu_i$ = population or true experimental error , that is true or population effect of extraneous variables associated with the experimental unit for the $j^{th}$ observation on the $i^{th}$ diet.

Assumptions of the population model are that the experimental errors, $\epsilon_{ij}'s$, are values of independent normal random variables each with mean or expected value of 0 and unknown variance $\sigma^2$. That is,

- $\epsilon_{ij}'s$ are statistically independent
- $\epsilon_{ij}'s$ each have mean of 0, that is $E[\epsilon_{ij}] = 0$
- $\epsilon_{ij}'s$ each have variance, that is $E[(\epsilon_{ij} - E[\epsilon_{ij}])^2] = E[\epsilon_{ij}^2] = \sigma^2$
- $\epsilon_{ij}'s$ are normally distributed

Thus $\sigma^2$ is the expected square of an error. As will be seen shortly, $\sigma^2$ will be estimated by an averaging of the squares of the estimated errors, that is the $e_{ij}'s$.

It is important to realize that $\overline{\mu}., \mu_i, \alpha_i, \epsilon_{ij}$ in the population effects model are NOT the same as $\overline{y}.., \overline{y}_i., A_i$ and $e_{ij}$ in the sample effects model. The latter are based on sample data; the former are the "true" values obtained if populations or very large numbers of mice were observed for each diet. The distinction is analogous to the distinction between a sample mean and a population mean in a first course in statistics. Inferences, such as confidence interval estimation and hypothesis testing, concern the "true" values.

Returning to the lifelength study, recall that for Diet 3,

$$\begin{aligned} y_{31} &= \overline{y}.. + A_3 + e_{31} \qquad (4.2) \\ y_{32} &= \overline{y}.. + A_3 + e_{32} \\ y_{33} &= \overline{y}.. + A_3 + e_{33} \\ y_{34} &= \overline{y}.. + A_3 + e_{34} \end{aligned}$$

Substituting actual values we have

$$\begin{aligned} 23 &= 21.50 + 1.50 + 0 \qquad (4.3)\\ 24 &= 21.50 + 1.50 + 1\\ 20 &= 21.50 + 1.50 + (-3)\\ 25 &= 21.50 + 1.50 + 2 \end{aligned}$$

With the population effects model we would have, for example, for $y_{31}$,

$$y_{31} = 23 = \overline{\mu}. + \alpha_3 + \epsilon_{31}$$

We cannot fill in values for $\overline{\mu}., \alpha_3, \epsilon_{31}$ because we don't know what the true values are. The value of $\overline{y}_{..} = 21.50$ is an estimate of $\overline{\mu}.$; $A_3 = 1.50$ is an estimate of $\alpha_3$; $e_{31} = 0$ is an estimate of $\epsilon_{31}$.

## 4.4 Testing for Overall Differences

### 4.4.1 Logic of the Test

In this section a hypothesis test is developed to test the null hypothesis that all true or population treatment means are equal versus the alternative hypothesis that the true means are not all the same. We will use the lifelength data to illustrate. In that example the null hypothesis is

$$H_o : \mu_1 = \mu_2 = \mu_3$$

versus the alternative hypothesis,

$$H_a : \textit{not all } \mu_i{}'s \textit{ are equal}$$

Note that $\mu_1 = \mu_2 = \mu_3$ is equivalent to $\alpha_1 = \alpha_2 = \alpha_3 = 0$ so an equivalent set of hypotheses is

$$H_o : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

versus

$$H_a : \textit{not all } \alpha_i{}'s = 0$$

Intuitively if the null hypothesis of no difference in true diet means or equivalently 0 diet effects holds, then the **sample** mean lifelengths $\overline{y}_{i.}$ would all be about the same or the **sample** diet effects $A_i$ would all be about 0. If the alternative hypothesis is really true, then the sample means should be different looking or the sample diet effects should not be close to 0.

Just because the sample diet means look different or the sample diet effects are not close to 0 does not necessarily prove that the diets truly have differential effects. One can obtain different sample means $\overline{y}_{i.}$ even if the $\mu_i$ are the same or obtain sample effects $A_i$ different from 0 even if the true effects $\alpha_i$ are all 0, simply because of the effects of extraneous factors. To see this consider the means model for Diet 1 and Diet 2, which are, respectively:

$$y_{1j} = \mu_1 + \epsilon_{1j}$$

and

$$y_{2j} = \mu_2 + \epsilon_{2j}$$

Remember that the $\epsilon$'s represent the effects of extraneous variables. Now averaging $y_{11}, y_{12}, y_{13}, y_{14}$ and $y_{21}, y_{22}, y_{23}, y_{24}$ according to the two models we have

$$\begin{aligned} \overline{y}_{1\cdot} &= \mu_1 + \overline{\epsilon}_{1\cdot} && (4.4) \\ \overline{y}_{2\cdot} &= \mu_2 + \overline{\epsilon}_{2\cdot} && (4.5) \end{aligned}$$

Thus according to the models,

$$\overline{y}_{1\cdot} - \overline{y}_{2\cdot} = (\mu_1 - \mu_2) + (\overline{\epsilon}_{1\cdot} - \overline{\epsilon}_{2\cdot})$$

The difference in sample means is a function of the difference in true means AND the difference of (average) errors. So even if the true means for diet 1 and diet 2 are the same ($\mu_1 - \mu_2 = 0$), it is still possible to obtain two sample means that are different simply due to the effects of extraneous variables. Thus what is needed to assess whether or not differences in sample means are "real" is some idea of what to expect for a difference in sample means solely from the effects of extraneous variables.

Consider the decomposition table again in Table 4.4. The calculated errors $e_{ij}$ measure solely the effects of extraneous variables. The calculated diet effects, $A_i$, however, contain the effects of extraneous variables and also the effects of diets if there truly are diet effects. So intuitively if the calculated diet effects, $A_i$ are "larger" than the calculated errors or extraneous variable effects, $e_{ij}$, then that is evidence that diet truly has an effect on lifelength. If the calculated diet effects are of about the same magnitude as the extraneous variable effects, then there is not enough evidence of true Diet effects.

In order to develop a test statistic which compares the "Diet" effects $A_i$ with the extraneous variable effects $e_{ij}$, we shall summarize the two sets of values. We are going to summarize the two sets not by averaging the effects, but by averaging the squares of the effects.

The sum of squared effects for the Diet treatment, SSTR, is defined as the sum of the squared diet effects for all observations in the decomposition table, and since there are repeats,

$$SSTR = n_1 A_1^2 + n_2 A_2^2 + n_3 A_3^2 \qquad (4.6)$$

$$\begin{aligned} &= 4(-0.75)^2 + 4(-0.75)^2 + 4(1.50)^2 \\ &= 2.25 + 2.25 + 9.00 \\ &= 13.50 \end{aligned}$$

The mean square of effects for the Diet treatment, MSTR, is obtained by dividing $SSTR$ by its degrees of freedom, $t - 1 = 3 - 1 = 2$.

$$\begin{aligned} MSTR &= SSTR/(t-1) \qquad (4.7) \\ &= 13.50/2 \\ &= 6.75 \end{aligned}$$

Now we will "average" the squared errors $e_{ij}$ by summing the squares of these values and then dividing by degrees of freedom. The sum of squared errors, denoted by SSE, for the lifelength data is

$$\begin{aligned} SSE &= (1.25)^2 + (-2.75)^2 + (0.25)^2 + (1.25)^2 \qquad Diet1 \\ &\quad +(-0.75)^2 + (-1.75)^2 + (2.25)^2 + (0.25)^2 \qquad Diet2 \\ &\quad +(0.00)^2 + (1.00)^2 + (-3.00)^2 + (2.00)^2 \qquad Diet3 \\ &= 10.75 + 8.75 + 14.00 \\ &= 33.5 \qquad (4.8) \end{aligned}$$

The mean squared error, denoted by MSE, is defined as the sum of squared errors divided by the degrees of freedom associated with the errors, which is $N - t = 12 - 3 = 9$. Thus

$$MSE = SSE/(N - t) = 33.5/9 = 3.72$$

It would be expected that mean square diet effects, $MSTR$, would be about the same as $MSE$ if diet truly has no effect, or equivalently it would be expected that the ratio $\frac{MSTR}{MSE}$ would be about 1. If diet does have an effect, then we would expect $\frac{MSTR}{MSE}$ to be somewhat larger than 1. Expectations can be quantified. It can be shown (see Kuehl [14], page 62) that

$$\begin{aligned} E[MSTR] &= \sigma^2 + \frac{1}{t-1}\sum_{i=1}^{t} n_i \alpha_i^2 \\ E[MSE] &= \sigma^2 \end{aligned}$$

Recall that $\sigma^2$ is the common variance of the errors $\epsilon_{ij}'s$.

Thus on average MSE is equal to the error variance $\sigma^2$ regardless of whether or not treatments have an effect. The actual observed value of MSE will be our estimate of the unknown error variance. If diet truly has an effect on lifelength (not all $\alpha_i$ are zero) then the expected value or average of $MSTR$ is greater than the expected value or average of $MSE$, equal to $\sigma^2$. If diet does not have an effect ($\alpha_i$ are all zero), then the expected values of $MSTR$ and $MSE$ are both equal to $\sigma^2$, in which case the observed values of $MSTR$ and $MSE$ should be similar.

In the lifelengths example, the estimate of the error variance $\sigma^2$, regardless of whether treatment effects exist, is the observed value of $MSE = 3.72$. The observed value of $MSTR = 6.75$. Thus the ratio

$$\frac{MSTR}{MSE} = 6.75/3.72 = 1.81$$

Thus the ratio is larger than 1, but is it "large enough" to provide convincing evidence that the diets truly have an effect. In order to answer this question we need to consider the probability or sampling distribution of the ratio $\frac{MSTR}{MSE}$ under the null hypothesis that Diet has no effect. That is, what are the possible values of the ratio simply due to error (effects of extraneous variables) when diet has no real effect. This sampling distribution is considered in the next section.

### 4.4.2 The F Sampling Distribution

In this section we describe the sampling distribution of the ratio $\frac{MSTR}{MSE}$.

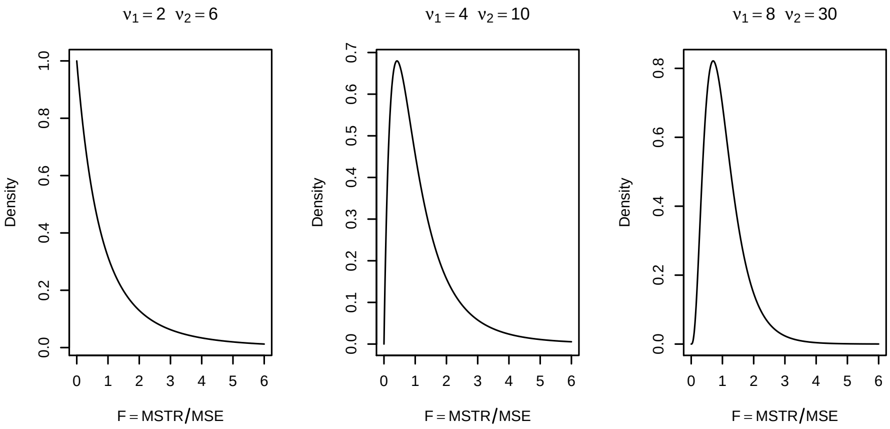
**Fact 4.1** *From statistical theory it is known that if the t populations corresponding to the t different treatments are normally distributed with identical population variances, the observations from the populations are independent, and the null hypothesis* $H_o : \alpha_1 = \alpha_2 = ... = \alpha_t = 0$ *is true, then the ratio* $MSTR/MSE$ *has the Fisher's "F" probability distribution with "numerator degrees of freedom",* $\nu_1 = t-1$ *and "denominator degrees of freedom"* $\nu_2 = N-t$*. The numerator degrees of freedom* $\nu_1 = t-1$ *is the degrees of freedom associated with* $MSTR$ *in the numerator of the ratio. The denominator degrees of freedom* $\nu_2 = N-t$ *is the degrees of freedom associated with* $MSE$ *in the denominator of the ratio.*

Properties of the F probability distribution:

- There are an infinite number of F distributions, depending on two parameters, the numerator degrees of freedom, $\nu_1$, and denominator degrees of freedom, $\nu_2$.
- The F distribution represents the probability distribution of a statistic which is non-negative, such as $MSTR/MSE$.
- The F distributions are positively skewed.

The density curves for three different F distributions are given in Figure 4.2. Note the skewness of the distributions. The upper $\alpha$ probability points, denoted

Figure 4.2: Examples of F distributions



by $F_{\alpha;\nu_1,\nu_2}$, are given in the Appendix, Tables A.7 and A.8, for $\alpha = 0.05$ and $\alpha = 0.01$, respectively, for various values of $\nu_1$ and $\nu_2$. For example, the upper 0.05 probability point for the F distribution in Figure 4.2 with $\nu_1 = 2$ and $\nu_2 = 6$ is from Table A.7, $F_{0.05;2,6} = 5.14$.

The right tail of the appropriate F distribution for the Diet example with $\nu_1 = t - 1 = 3 - 1 = 2$ and $\nu_2 = N - t = 12 - 3 = 9$ is graphed in Figure 4.3. The upper 0.05 probability point, 4.26, and the observed value of the ratio $MSTR/MSE$ are plotted along the horizontal axis. The P-value is shaded.

The observed value of the F ratio for the lifelength data is $F = 1.81$. The P-value associated with this value is

$$P - value = P(F \geq 1.81)$$

This P-value can only be approximated from Table A.7 as $P > 0.05$. A computer program will show that P-value = 0.218. Assuming a significance level $\alpha = 0.05$ then there is not enough evidence that Diet has an effect on lifelength.

### 4.4.3 Summary of the F test for Treatment Effects

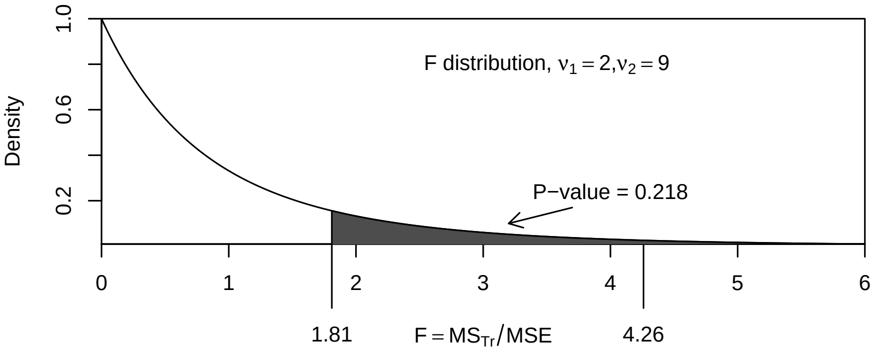
The null and alternative hypotheses for the test of treatment effects for $t$ treatments are

$$H_o : \alpha_1 = \alpha_2 = \ldots = \alpha_t = 0$$

or equivalently,

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_t$$

Figure 4.3: P-value for Diet Example



The alternative hypothesis is

$$H_a : \text{ not all } \alpha_i{}'s = 0$$

or equivalently,

$$H_a : \text{ not all } \mu_i{}'s \text{ are equal}$$

The test statistic is

$$F = \frac{MSTR}{MSE} = \frac{SSTR/(t-1)}{SSE/(N-t)}$$

where

$$SSTR = \sum_{i=1}^{t} n_i A_i^2 = \sum_{i=1}^{t} n_i (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$$

and

$$\begin{aligned} SSE &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (e_{ij})^2 \\ &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\cdot})^2 \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_t - 1)s_t^2 \qquad (4.9) \end{aligned}$$

where $s_1^2, s_2^2, ..., s_t^2$ are the sample variances of the responses in treatment groups $1, 2, .., t$, respectively. Thus SSE can be calculated with only a knowledge of the treatment group sizes and variances (or standard deviations).

Table 4.5: Decomposition of Lifelength Data

| | $y_{ij}$ | = | $\overline{y}_{..}$ | + | $A_i$ | + | $e_{ij}$ |
|---|---|---|---|---|---|---|---|
| Diet 1 | 22 | = | 21.50 | + | -0.75 | + | 1.25 |
| | 18 | = | 21.50 | + | -0.75 | + | -2.75 |
| | 21 | = | 21.50 | + | -0.75 | + | 0.25 |
| | 22 | = | 21.50 | + | -0.75 | + | 1.25 |
| Diet 2 | 20 | = | 21.50 | + | -0.75 | + | -0.75 |
| | 19 | = | 21.50 | + | -0.75 | + | -1.75 |
| | 23 | = | 21.50 | + | -0.75 | + | 2.25 |
| | 21 | = | 21.50 | + | -0.75 | + | 0.25 |
| Diet 3 | 23 | = | 21.50 | + | 1.50 | + | 0.00 |
| | 24 | = | 21.50 | + | 1.50 | + | 1.00 |
| | 20 | = | 21.50 | + | 1.50 | + | -3.00 |
| | 25 | = | 21.50 | + | 1.50 | + | 2.00 |

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common variance holds, the $F$ ratio above has the "F" distribution with $\nu_1 = (t-1)$ numerator degrees of freedom and $\nu_2 = (N-t)$ denominator degrees of freedom.

At a significance level of $\alpha$ the null hypothesis would be rejected if the observed value of the test statistic, $F_o$, is larger than $F_{\alpha;(t-1),N-t}$ the upper $\alpha$ probability point from the appropriate F distribution. Equivalently the null hypothesis is rejected if $P - value \leq \alpha$, where $P - value = P[F \geq F_o]$. Upper $\alpha$ probability points for $\alpha = 0.05$ and $\alpha = 0.01$ are given in Tables A.7 and A.8, respectively. P-values can only be approximated using Table $A$.7 or $A$.8. More precise P-values can be obtained using statistical computing software such as SAS or SPSS.

## 4.5 The Analysis of Variance (ANOVA) Table

The decomposition of the lifelength data is reproduced in Table 4.5.

Recall that we used the sum of squared diet effects and the sum of squared errors to develop a test for true diet effects, where $SSTR = 13.50$ and SSE = 33.50.

In this section we will also sum the squares of the lifelengths, which we shall call total sum of squares and denote by $SSTOT$:

$$\begin{aligned} SSTOT &= (22)^2 + (18)^2 + (21)^2 + (22)^2 && Diet1 \\ &\quad +(20)^2 + (19)^2 + (23)^2 + (21)^2 && Diet2 \\ &\quad +(23)^2 + (24)^2 + (20)^2 + (25)^2 && Diet3 \\ &= 1733 + 1731 + 2130 \\ &= 5594 && (4.10) \end{aligned}$$

Table 4.6: ANOVA Table for Lifelength Data

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Grand Mean | 1 | 5547 | | | |
| Treatments | 2 | 13.50 | 6.75 | 1.81 | 0.218 |
| Error | 9 | 33.50 | 3.72 | | |
| Total | 12 | 5594 | | | |

We will also consider the sum of the squares of the grand mean, $SSGM$ associated with each lifelength. This is easier since all values are the same.

$$SSGM = 12(21.50)^2 = 5547$$

Note that

$$SSTOT = SSGM + SSTR + SSE$$

or

$$5594 = 5547 + 13.50 + 33.50$$

From a conceptual standpoint $SSTOT$ can be regarded as a summary measure of variability in the lifelengths and we are partitioning this variability into components, that due to some common value, the grand mean, that due to the treatments (diets here), and that due to error. We are doing an ANalysis Of the VAriation (ANOVA) in lifelengths by breaking it up into parts.

An ANOVA table is a summary of the components of the total variation in the response variable with also the F ratio for testing for treatment effects (and a P-value if you are using a computer). An ANOVA table in our example is given in Table 4.6.

The general form of the ANOVA table for a one factor completely randomized design when there are $t$ treatments and $N$ total observations is given in Table 4.7. Note that $MSE$, the estimate of the variance of the error terms in the population model, appears in the denominator of the F test statistic and thus is used to determine if there are treatment effects. $MSE$ will also be used in the next chapter to help determine which means differ if we conclude from the F test that there are differences somewhere.

An alternative partitioning of the lifelength data is given in Table 4.8.

The grand mean of 21.50 months was subtracted from each lifelength (this is called lifelength corrected for the grand mean). The interpretation now is that each deviation of a lifelength from the grand mean of 21.50 is partly due to diet effect and partly due to error. For example, the difference between the lifelength of 22 months and the grand mean of 21.50 months (= 0.50) is part diet effect of $-0.75$ and part error of 1.25.

Table 4.7: General ANOVA Table - One Factor CRD

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Grand Mean | 1 | $SSGM$ | | | |
| Treatments | $t-1$ | $SSTR$ | $MSTR$ | $MSTR/MSE$ | *** |
| Error | $N-t$ | $SSE$ | $MSE$ | | |
| Total | $N$ | $SSTOT$ | | | |

Table 4.8: Decomposition of Lifelength Corrected for Grand Mean

| | $y_{ij} - \overline{y}_{..}$ | $=$ | $A_i$ | $+$ | $e_{ij}$ |
|---|---|---|---|---|---|
| Diet 1 | 0.50 | = | -0.75 | + | 1.25 |
| | -1.50 | = | -0.75 | + | -2.75 |
| | -0.50 | = | -0.75 | + | 0.25 |
| | 0.50 | = | -0.75 | + | 1.25 |
| Diet 2 | -1.50 | = | -0.75 | + | -0.75 |
| | -2.50 | = | -0.75 | + | -1.75 |
| | 1.50 | = | -0.75 | + | 2.25 |
| | -0.50 | = | -0.75 | + | 0.25 |
| Diet 3 | 1.50 | = | 1.50 | + | 0.00 |
| | 2.50 | = | 1.50 | + | 1.00 |
| | -0.50 | = | 1.50 | + | -3.00 |
| | 3.50 | = | 1.50 | + | 2.00 |

Table 4.9: ANOVA Table for Lifelength Data: Correction for Grand Mean

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Treatments | 2 | 13.50 | 6.75 | 1.81 | 0.218 |
| Error | 9 | 33.50 | 3.72 | | |
| Total (Corrected) | 11 | 47 | | | |

Table 4.10: General ANOVA Table with Correction for Grand Mean - One Factor CRD

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Treatments | $t-1$ | $SSTR$ | $MSTR$ | $MSTR/MSE$ | *** |
| Error | $N-t$ | $SSE$ | $MSE$ | | |
| Total(Corrected) | $N-1$ | $SSTOT_C$ | | | |

To obtain the modified ANOVA table we sum the squares of the corrected lifelengths and obtain the total sum of squares corrected for the mean, denoted by $SSTOT_C$

$$SSTOT_C = (0.50)^2 + (-1.50)^2 + \ldots + (3.50)^2 = 47$$

The sums of squares partitioning is

$$SSTOT_C = SSTR + SSE$$

or in the lifelength example,

$$47 = 13.50 + 33.50$$

The modified ANOVA table would then look as in Table 4.9. Note that rather than summing squares of lifelengths corrected for the grand mean, $SSTOT_C$ can be calculated by $SSTOT_C = SSTOT - SSGM$.

The general form of the ANOVA table with the correction for the grand mean is given in Table 4.10.

## 4.6 Independent Samples t Test Revisited

Consider the two-sided pooled independent samples $t$ test introduced in Chapter 3 that assumed equal population variances. This special testing situation can be shown to be equivalent to ANOVA and F test of this chapter with $t = 2$ treatments.

First we can model the independent samples t test situation as follows. Let $y_{11}, y_{12}, ..., y_{1,n_1}$ be independent measurements from a population with mean $\mu_1$. Let $y_{21}, y_{22}, ..., y_{2,n_2}$ be independent measuresment from a population with mean $\mu_2$. Let $\sigma^2$ be the common population variance.

We can decompose the $y_{ij}$ as in this chapter:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where the $\epsilon_{ij}$ are independent, normally distributed random variables with mean of 0 and variance of $\sigma^2$. These are the conditions that are assumed in this chapter. Note that without the assumption of equal population variances the variances of the errors would not be equal. Thus an F test for the equality of the $t = 2$ means should be equivalent to the two-sided $t$ test of Chapter 3. An example follows.

**Example 4.1** *This example is from the Chapter* 3 *exercises. A trucking firm wishes to choose between two alternate routes for transporting merchandise from one depot to another. One major concern is the travel time. In a study, 5 drivers are randomly assigned to route A, the other 5 were assigned to route B. Data was obtained from each driver on travel time (hours) and given below.*

| *Route* | *Travel Time (hours)* |
|---|---|
| *A* | *18,24,30,21,32* |
| *B* | *22,29,34,25,35* |

*Is there evidence of a difference in driving time between the two routes?*

The following table gives the means and standard deviations of the two groups of travel times:

| Route | n | Mean | Standard Deviation |
|---|---|---|---|
| A | 5 | 25.0 | 5.92 |
| B | 5 | 29.0 | 5.61 |

The pooled sample variance $s_p^2$, the estimate of the common population variance, is $s_p^2 = 33.25$. The observed $t$ ratio is $t = -1.10$. The two-sided P-value is $P[|t| \geq |-1.10|] = 0.3046$ with $df = 5 + 5 - 2 = 8$. The ANOVA table for the route times data is given below.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Routes | 1 | 40 | 40 | 1.20 | 0.3046 |
| Error | 8 | 266 | 33.25 | | |
| Total (Corrected) | 9 | 306 | | | |

Note that the P-value for the observed F ratio of 1.20 is the same as the P-value for the observed t ratio of $-1.10$. It can also be shown that the square of the $t$ ratio is equal to the $F$ ratio. Note here that the square of the $t$ ratio, $(-1.10)^2 = 1.21$ differs slightly from the observed F ratio, 1.20, because of rounding. The equivalence is only between the two-sided independent samples t test, assuming equal population variances, and the $F$ test. It should also be noted that the estimate of the common population variance, $s_p^2 = 33.25$ from the $t$ procedure, is the same as $MSE = 33.25$, the estimate of the common variance of the error terms in the one factor population model. Thus one can cast the pooled independent samples $t$ test (with equal population variances) within the context of analysis of variance.

## 4.7 SAS Code

### 4.7.1 Lifelength Example

```
* Lifelength Example;

* Input diet and lifelength;
data DIET;
  input Diet Lifelength;
datalines;
1  22
1  18
1  21
1  22
2  20
2  19
2  23
2  21
3  23
3  24
3  20
3  25
;
run;

*  Use proc glm to obtain ANOVA table;
proc glm data = DIET;
  class Diet;
  model Lifelength = Diet;
run;
```

### 4.7.2 Example 4.1

```
* Example 4.1;

* Input;
data TruckRoute;
  input Route $ TravelTime;
datalines;
A  18
A  24
A  30
A  21
A  32
B  22
B  29
```

```
B  34
B  25
B  35
;

* proc ttest for obtaining results of independent
  samples t test;
proc ttest data = TruckRoute;
  class Route;
  var TravelTime;

* proc glm for obtaining ANOVA table;
proc glm data = TruckRoute;
  class Route;
  model TravelTime = Route;
run;
```

# Problems for Chapter 4

4.1* A psychologist was interested in the effects of three different kinds of drugs on the mean time to complete a certain task. The psychologist used 15 subjects and randomly assigned 5 of them to each drug A, B, and C. The data represent the time in minutes to complete the task.

| | | | | | |
|---|---|---|---|---|---|
| A | 20 | 22 | 25 | 24 | 19 |
| B | 21 | 26 | 26 | 27 | 25 |
| C | 30 | 24 | 26 | 25 | 30 |

a. Construct a decomposition table (one without the correction for the grand mean and one with the correction for the grand mean).

b. Construct the ANOVA table (not corrected for the grand mean and corrected for the grand mean).

c. At the 5% significance level, is there evidence of a difference in true mean time (or a difference in true effects from 0) for the drugs? Use the upper 0.05 probability point from the F-table in the Appendix rather than a P-value to make your decision.

4.2* Consider the following incomplete ANOVA table for a one factor completely randomized design.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Grand Mean | 1 | 1728 | | | |
| Treatments | 4 | ___ | ___ | ___ | ___ |
| Error | ___ | 105 | ___ | | |
| Total | 30 | 1918 | | | |

a. Fill in the blanks of the table. Using Table A.7 or A.8 give an inequality expressing the approximate $P - value$.

b. How many treatments are in the study upon which this ANOVA table is based?

c. Assuming equal replication of treatments how many replications are there per treatment?

d. At a significance level of $\alpha = 0.01$ would the null hypothesis of equal true treatment means be rejected?

4.3* A former statistics student investigated the effects of plant food and music on growth of pansy plants. She investigated two levels of plant food (yes, no) and three levels of music (techno, classical, reggae). The data for plant growth (in inches) for those replications where plant food was supplied only is given below.

| | | | | |
|---|---|---|---|---|
| Classical | 2.3 | 2.6 | 2.9 | 3.0 |
| Reggae | 2.5 | 2.1 | 2.3 | 2.4 |
| Techno | 1.7 | 2.3 | 2.7 | 2.9 |

Is there evidence that the different types of music result in different mean growth for pansy plants? Use $\alpha = 0.05$.

4.4* Two students, Cheryl Butterworth and Josh Hiller, performed an experiment to study the effect of beverage type on the amount of time for ice cubes to melt. Types of beverage were coca-cola, orange juice, and water. The beverages were left out over night to set them at a constant temperature. Fifteen ice cubes of approximately the same size were randomly assigned to fifteen identical cups. Equal amounts of beverage, five of each kind, were randomly assigned to the cups. The amount of time (minutes) for the ice cubes to melt was recorded and given below.

| | | | | | |
|---|---|---|---|---|---|
| Coca cola | 19 | 17 | 15 | 14 | 18 |
| Orange Juice | 27 | 28 | 30 | 26 | 27 |
| Water | 10 | 11 | 13 | 7 | 9 |

a. What is the factor of interest?

b. What are the treatments?

c. What are the experimental units?

d. Give some extraneous variables that are part of experimental error?

e. Give the (true) effects model for the data and describe the parameters of the model within the context of this study.

f. Is there evidence of a difference in melting times for the three treatments?

   i. Give the null and alternative hypotheses in terms of the effects parameters from part (e).

   ii. Use a statistical program to calculate the F ratio and associated P-value. Answer the question at the 0.05 level of significance.

   iii. What assumptions about the true errors are necessary in order to ensure that your conclusions in part(ii) are valid?

4.5* This data comes from an example in Kutner, Nachtsheim, Neter, and Li [15]. Four brands of rust inhibitors (A,B,C,D) were compared. The four brands were assigned to 40 experimental units, 10 for each brand in a completely randomized design. The rust inhibition measurements are given in the table below with the higher the value, the more effective is the brand.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 43.9 | 39.0 | 46.7 | 43.8 | 44.2 | 47.7 | 43.6 | 38.9 | 43.6 | 40.0 |
| B | 89.8 | 87.1 | 92.7 | 90.6 | 87.7 | 92.4 | 86.1 | 88.1 | 90.8 | 89.1 |
| C | 68.4 | 69.3 | 68.5 | 66.4 | 70.0 | 68.1 | 70.6 | 65.2 | 63.8 | 69.2 |
| D | 36.2 | 45.2 | 40.7 | 40.5 | 39.3 | 40.3 | 43.2 | 38.7 | 40.9 | 39.7 |

With the help of statistical software answer the following:

a. Give the sample means and standard deviations. Does there appear to be a different in brands?

b. Obtain an ANOVA table for this data. What is the estimate of the variance of the errors?

c. Is there evidence of a difference in the degree of inhibition? Use $\alpha = 0.05$. Use the P-value obtained from your software to answer the question.

4.6* In the article "A prospective study of patients with chronic back pain randomised to group exercise, physiotherapy or osteopathy" (*Physiotherapy* [2008]: 94, 21-28) researchers investigated the difference on a disability index between patients treated with group exercise, physiotherapy or osteopathy. At baseline and at a 6-week followup patients were measured on the Oswestry Diability Index (ODI), which measures aspects of pain and functional ability on a scale from 0 (no disability) to 100 (extreme disability). The response variable was the difference in the ODI (baseline - followup). The treatment group sizes, means, and standard deviations are given in the following table.

| Treatment Group | n | Mean | Standard Deviation |
|---|---|---|---|
| Group exercise | 24 | 4.5 | 8.4 |
| Physiotherapy | 35 | 4.1 | 8.0 |
| Osteopathy | 39 | 5.0 | 10.5 |

Suppose an ANOVA was conducted on the data. What is SSE? What is MSE?

4.7* Suppose that an experiment is to be conducted with 3 treatments labelled 1,2,3 and with group sizes of $n_1 = n_2 = n_3 = 10$. Suppose that the true or population means of the response variable corresponding to the treatments are $\mu_1 = 4$, $\mu_2 = 4$, and $\mu_3 = 7$. Suppose that the true error variance is $\sigma^2 = 2$.

a. What are the true or population effects of the treatments?

b. Suppose that an observation on the response variable under treatment 2 is $Y = 6$. Decompose this observation of $Y = 6$ according to the population effects model.

c. If this experiment is repeated a large number of times then MSE will on average be equal to what value?

d. If this experiment is repeated a large number of times then MSRT will on average be equal to what value?

e. Suppose the experiment is performed once and data is collected. Is it possible that the difference in sample means for treatments 1 and 2, $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$, will be different than 0 even though there is no difference in the corresponding population means ($\mu_1 = 4$, $\mu_2 = 4$)? Explain.

4.8* In the article "Sex differences in viewing sexual stimuli: An eye-tracking study in men and women" (*Hormones and Behavior* [2007]: Vol. 51, pgs 524-533 ) researchers compared three groups of heterosexuals: males, normal (menstrual) cycling females (NC), and oral contracepting females (OC), on various responses to viewing sexual stimuli. Stimuli were sexually explicit photos of heterosexual couples engaged in oral sex or intercourse. Because researchers thought that any group differences found in the current study could be due to differences in participants' previous experience viewing sexually explicit stimuli, sexual attitudes, sexual motivation, or comfort with visual sexual stimuli, they compared the three groups on these variables using a one factor ANOVA. Only the results of the comparison of the three groups on sexual motivation as measured by the frequency of sexual thoughts and desire to engage in sexual activity in the previous month is given here. The table below gives the means and standard deviations of the sexual motivation variable for the three groups.

| Group | n | Mean | Standard Deviation |
|---|---|---|---|
| Men | 15 | 5.2 | 0.71 |
| NC Women | 15 | 3.8 | 1.18 |
| OC Women | 14 | 4.64 | 0.73 |

a. Give the population effects model for this study and describe within context the various terms in the model. Be sure to give the assumptions associated with the error terms.

b. Give an appropriate null and alternative hypothesis in terms of effects of group.

c. The F ratio was 8.72 with P-value = 0.001.

  i. What are the numerator and denominator degrees of freedom for the F ratio?

  ii. What does the P-value mean as a probability within the context of this problem?

  iii. What is your conclusion in context? Use a significance level of 0.05.

4.9* The American Statistical Association holds an annual poster and project competition for students from grades K-12. Winners receive a monetary award and a plaque. One of the winners in the 2013 competition conducted an experiment to answer the question: Do dryer ball reduce drying time? The student conducted the experiment in response to an ad that claimed that balls "reduce drying time by up to 25%. The student randomly assigned the next 40 of his family's wash loads to either be dried with dryer balls added to the dryer or not. Only one washer and dryer was used. The student weighed each load prior to drying and recorded how long it took the load to dry (in minutes) using a stop watch. The dryer has a sensor that detects when the clothes are dry. The drying times are provided in the table below.

| Dryer Balls | No Dryer Balls |
|---|---|
| 34.5 | 23.5 |
| 28.0 | 24.5 |
| 28.5 | 22.5 |
| 24.0 | 21.0 |
| 24.0 | 29.0 |
| 40.0 | 21.0 |
| 21.5 | 34.0 |
| 29.0 | 36.0 |
| 34.0 | 34.0 |
| 30.0 | 24.0 |
| 33.0 | 31.0 |
| 22.0 | 26.0 |
| 31.0 | 21.0 |
| 22.0 | 41.0 |
| 22.0 | 32.0 |
| 29.0 | 36.0 |
| 21.0 | 39.0 |
| 33.0 | 29.0 |
| 35.0 | 22.0 |
| 24.5 | 21.0 |

Analyze the data in two ways. Use the two-sided independent samples t test first. Give the value of the $t$ test statistic, degrees of freedom, and P-value. Next analyze the data using a one factor analysis of variance. Give the value of the F statistic, degrees of freedom, and P-value. Show the equivalence between the two sets of results in terms of test statistics, degrees of freedom, and P-values and thus the same conclusion is reached.

4.10 This example is based on an experiment described in the article "Acid rain and the survival of fresh water guppies" (www.all-science-fair-projects.com) Fifty freshwater guppies were placed in 5 small fish tanks, 10 guppies per tank. The pH level of the water in the 5 tanks were 4.5, 5.0, 5.5, 6.0, and 6.5. The number of fish surviving was recorded 2, 4, 6, 8, and 10 hours after the start of the experiment. The data are given below.

| Tank No | pH Level | Number of Fish Surviving | | | | | |
|---|---|---|---|---|---|---|---|
| | | Start | 2 hrs | 4 hrs | 6 hrs | 8 hrs | 10 hrs |
| 1 | 6.5 | 10 | 10 | 10 | 10 | 10 | 10 |
| 2 | 6.0 | 10 | 10 | 9 | 9 | 8 | 8 |
| 3 | 5.5 | 10 | 9 | 9 | 8 | 8 | 7 |
| 4 | 5.0 | 10 | 3 | 1 | 0 | 0 | 0 |
| 5 | 4.5 | 10 | 1 | 0 | 0 | 0 | 0 |

i. The treatments are the 5 different pH levels of water. How many replicates are there of each treatment?

ii. Is the one factor analysis for the completely randomized design (6 observations per pH level) discussed in this chapter appropriate for the analysis of this data? Explain.

4.11 This example is based on data reported in Oehlert ([24], page 61) on leaflet angle (degrees) from plants in the genus *Albizzia* after exposure to red light. Certain plants from this genus have the ability to fold and unfold their leaves under various light conditions. The researcher selected 15 leaves and subjected them to red light for 3 minutes. The leaves were then divided at random into three groups with 5 leaves per group. The groups were defined by the length of time, 30, 45, or 60 minutes after exposure to the red light when leaflet angle was measured for the leaves. The data are given below.

| Delay (minutes) | Angle (degrees) | | | | |
|---|---|---|---|---|---|
| 30 | 140 | 138 | 140 | 138 | 142 |
| 45 | 140 | 150 | 120 | 128 | 130 |
| 60 | 118 | 130 | 128 | 118 | 118 |

a. What is the factor in this study? What are the treatments? What are the experimental units?

b. Construct a decomposition table (one without the correction for the mean and one with the correction for the mean).

c. Construct the ANOVA table (not corrected for mean and corrected for mean).

d. At the 5% significance level, is there evidence of a difference in true mean angle (or a difference in true effects from 0) among the delay times? Use the upper 0.05 probability point from the F-table in the Appendix rather than a P-value to make your decision.

# Chapter 5

# Multiple Comparisons

## 5.1 Introduction

If it has been concluded from the F test from ANOVA that there are some differences in the means of a response variable, then a researcher typically would want to know which means differ. Depending upon the study objectives the researcher may wish to make pairwise comparisons of all possible means and then rank the treatments. Or the researcher may only be interested in comparing treatment means with a control mean. In another scenario the researcher may wish to compare means of subsets of treatments. The researcher will in general make **multiple comparisons** of the means to satisfy the objectives of the study.

## 5.2 Types of Multiple Comparisons

### 5.2.1 All Pairwise Comparisons

If there are $t$ treatments in a study then it is easily shown that there are

$$m = \frac{t!}{2!(t-2)!}$$

possible pairwise comparisons, where in general the symbol $x!$ stands for the product $(x)(x-1)\dots(1)$. For example, if $t = 3$ there are 3 possible comparisons; if $t = 4$ there are 6 possible comparisons, and so on.

Suppose that a one factor experiment is conducted using a completely randomized design as in Chapter 4. A significant F test is obtained and the researcher is interested in making all possible pairwise comparisons of the $t$ means. One approach to making the $m$ comparisons is the do $m$ t-tests according to methods of Chapter 3. We shall use the pooled independent samples $t$ test that assumes equal population variances, consistent with the assumption of equal population variances in the ANOVA. The pooled standard deviation $s_p$ in the test statistic from Chapter 3 will be replaced by $\sqrt{MSE}$ from the ANOVA. Suppose that population means $\mu_i$ and $\mu_j$ are to be compared with $i$ and $j$ referring to two of the possible $t$ means. Then for an $\alpha$ level of significance and a two-sided test the null hypothesis $H_o : \mu_i = \mu_j$ is rejected if P-value $\leq \alpha$, or equivalently,

$$\left| \frac{\overline{y}_{i\cdot} - \overline{y}_{j\cdot}}{\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \right| \geq t_{\frac{\alpha}{2};\nu} \tag{5.1}$$

or

$$|\overline{y}_i - \overline{y}_j| \geq t_{\frac{\alpha}{2};\nu}\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{5.2}$$

Here $t_{\frac{\alpha}{2};\nu}$ refers to the upper $\frac{\alpha}{2}$ probability point from a $t$ distribution with degrees of freedom $\nu = N - t$ associated with MSE. $\sqrt{MSE}$ is the estimate from ANOVA of the common population standard deviation $\sigma$. The product $\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ is the standard error of the difference in sample means $\overline{y}_{i\cdot} - \overline{y}_{j\cdot}$. Note that this procedure differs slightly from the $t$ test considered in Chapter 3. First the estimate of the population standard deviation, $\sqrt{MSE}$, is based on all $t$ samples, not just the two samples being compared. The appropriate degrees of freedom, $\nu$, is the degrees of freedom associated with MSE. For other designs discussed later in this text, degrees of freedom associated with the estimate of the population standard deviation will differ from that of the one factor completely randomized design.

A $100(1 - \alpha)\%$ confidence interval for the difference $(\mu_i - \mu_j)$ is

$$(\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) \pm (t_{\frac{\alpha}{2};\nu})\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{5.3}$$

If the confidence interval does not include zero then the null hypothesis is rejected and we conclude that the two population means $\mu_i$ and $\mu_j$ are different. If the interval includes zero the null hypothesis is not rejected, i.e. there is not enough evidence that the two means are different.

**Example 5.1** *Source: Weber and Skilling, p. 241. A company is considering three different covers for boxes of a brand of cereal. Box cover 1 has a picture of a sports hero eating the cereal, cover 2 has a picture of a child eating the cereal, and cover 3 has a picture of a bowl of the cereal. The company wants to determine which cereal box type provides for the most sales. Eighteen test markets were selected by the company and each box type was randomly assigned to six markets. The number of boxes of this cereal sold per 10,000 population in a specified period is recorded for each test market. The sales are as follows:*

| | | | | | | |
|---|---|---|---|---|---|---|
| *Cover 1: Sports Hero* | *52.4* | *47.8* | *52.4* | *51.3* | *50.0* | *52.1* |
| *Cover 2: Child* | *50.1* | *45.2* | *46.0* | *46.5* | *47.4* | *46.2* |
| *Cover 3: Cereal Bowl* | *49.2* | *48.3* | *49.0* | *47.2* | *48.6* | *48.2* |

Table 5.1: Descriptives for Sales Data

| Group | Mean | St.Dev. |
|---|---|---|
| Cover 1 (Sports Hero) | $\overline{y}_{1.} = 51.00$ | $s_1 = 1.81$ |
| Cover 2 (Child) | $\overline{y}_{2.} = 46.90$ | $s_2 = 1.72$ |
| Cover 3 (Cereal Bowl) | $\overline{y}_{3.} = 48.42$ | $s_3 = 0.71$ |

Table 5.2: ANOVA Table for Box Cover Sales

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Covers | 2 | 51.57 | 25.78 | 11.43 | 0.0010 |
| Error | 15 | 33.83 | 2.26 | | |
| Total (Corrected) | 17 | 85.40 | | | |

*Is there evidence of a difference in population mean sales among the three types of covers? Use a significance level of* 0.05*. If the F test for overall differences is significant then use* 95% $t$ *confidence intervals to determine which means differ.*

Table 5.1 provides the means and standard deviations of the sales data for the three cover types. The ANOVA table is given in Table 5.2.

The differences in the mean sales for the three cover types are significant with $F = 11.43$ and P-value= 0.0010. The endpoints for the confidence intervals for the differences in population means $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$, respectively, are

$$\begin{array}{rcl} (\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) & \pm & (t_{\frac{0.05}{2};15})\sqrt{MSE}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ (\overline{y}_{1\cdot} - \overline{y}_{3\cdot}) & \pm & (t_{\frac{0.05}{2};15})\sqrt{MSE}\sqrt{\frac{1}{n_1} + \frac{1}{n_3}} \\ (\overline{y}_{2\cdot} - \overline{y}_{3\cdot}) & \pm & (t_{\frac{0.05}{2};15})\sqrt{MSE}\sqrt{\frac{1}{n_2} + \frac{1}{n_3}} \end{array}$$

The upper 0.025 percentile from the $t$ distribution, $t_{\frac{0.05}{2};15}$, with $\nu = 15$ degrees of freedom is, from Table A.2, 2.131. Thus the endpoints for the 3 confidence intervals are:

$$\begin{array}{rcl} (51.00 - 46.90) & \pm & (2.131)\sqrt{2.26}\sqrt{\frac{1}{6} + \frac{1}{6}} \\ (51.00 - 48.42) & \pm & (2.131)\sqrt{2.26}\sqrt{\frac{1}{6} + \frac{1}{6}} \\ (46.90 - 48.42) & \pm & (2.131)\sqrt{2.26}\sqrt{\frac{1}{6} + \frac{1}{6}} \end{array}$$

or

$$\begin{array}{rcl} 4.10 & \pm & 1.85 \\ 2.58 & \pm & 1.85 \\ -1.52 & \pm & 1.85 \end{array}$$

Thus the three intervals are:

$$\begin{array}{rcccl} 2.25 & \leq & \mu_1 - \mu_2 & \leq & 5.95 \\ 0.73 & \leq & \mu_1 - \mu_3 & \leq & 4.43 \\ -3.37 & \leq & \mu_2 - \mu_3 & \leq & 0.33 \end{array}$$

It can be concluded that the box cover with the sports hero results in the highest means sales. There is not enough evidence of a difference in mean sales between the box cover with the child and the box cover with the bowl of cereal. It is estimated with 95% confidence that the mean sales with the sports hero as the box cover is between $2.25(/10,000)$ and $5.95(/10,000)$ boxes higher than when the cover has a child. It is estimated with 95% confidence that the mean sales with the sports hero as the box cover is between $0.73(/10,000)$ and $4.43(/10,000)$ boxes higher than when the cover has a bowl of cereal.

### 5.2.2 Contrasts - Generalization of Pairwise Comparisons

The difference between two means, $\mu_i - \mu_j$, is an example of a more general comparison of the means called a **contrast**. In some studies there is some kind of structure to the treatments and interest is not in all possible pairwise comparisons but in certain pre-planned comparisons of subgroups of the means.

Let $\mu_1, ..., \mu_t$ be the population treatment means. Then we define a **contrast** of the means to be a **linear combination** of the means, $C$:

$$C = c_1\mu_1 + c_2\mu_2 + ... + c_t\mu_t \tag{5.4}$$

where the $c's$ are constants defined so that $c_1 + c_2 + ... + c_t = 0$.

Suppose in a study with one factor there are $t = 4$ treatments with means $\mu_1, ..., \mu_4$. An example of a contrast would be a pairwise comparison such as $\mu_3 - \mu_4$ because we can write this difference as $C_1$ where

$$C_1 = 0\mu_1 + 0\mu_2 + (1)\mu_3 + (-1)\mu_4$$

where $c_1 = 0$, $c_2 = 0$, $c_3 = -1$, and $c_4 = 1$ with the sum of the $c's$ being 0.

However another example of a contrast would be $C_2$ defined as

$$C_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$$

This contrast represents a comparison of the average of $\mu_1$ and $\mu_2$ with the average of the $\mu_3$ and $\mu_4$. Here $c_1 = \frac{1}{2}$, $c_2 = \frac{1}{2}$, $c_3 = -\frac{1}{2}$, and $c_4 = -\frac{1}{2}$ with the sum of the $c's$ equalling 0.

We will now consider estimation and hypothesis testing regarding an arbitrary contrast defined as in Equation 5.4. We first need an estimate of $C$.

The point estimate of the contrast $C$ in Equation 5.4 is

$$\hat{C} = c_1\overline{y}_{1\cdot} + c_2\overline{y}_{2\cdot} + ... + c_t\overline{y}_{t\cdot} \tag{5.5}$$

This estimate is normally distributed with mean or expected value

$$E[\hat{C}] = C \tag{5.6}$$

and variance, denoted by $\sigma^2\{\hat{C}\}$, can be shown to be

$$\begin{aligned} \sigma^2\{\hat{C}\} &= c_1^2\sigma^2\{\overline{y}_{1\cdot}\} + c_2^2\sigma^2\{\overline{y}_{2\cdot}\} + ... + c_t^2\sigma^2\{\overline{y}_{t\cdot}\} \\ &= c_1^2\frac{\sigma^2}{n_1} + c_2^2\frac{\sigma^2}{n_2} + ... + c_t^2\frac{\sigma^2}{n_t} \\ &= \sigma^2(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + ... + \frac{c_t^2}{n_t}) \end{aligned} \tag{5.7}$$

In practice the error variance, like in Chapter 4, is unknown and is estimated with $MSE$ from the analysis of variance. Thus the estimate of the variance of $\hat{C}$, denoted by $s^2\{\hat{C}\}$ is

$$s^2\{\hat{C}\} = \text{MSE}(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + ... + \frac{c_t^2}{n_t}) \tag{5.8}$$

The estimated standard error, $s\{\hat{C}\}$ of the estimated contrast $\hat{C}$ is the square root of the estimated variance, that is,

$$s\{\hat{C}\} = \sqrt{s^2\{\hat{C}\}} \tag{5.9}$$

Since $\hat{C}$ is normally distributed with mean $C$ and variance $\sigma^2\{\hat{C}\}$ then the ratio

$$\frac{\hat{C} - C}{\sqrt{\sigma^2\{\hat{C}\}}} \tag{5.10}$$

has a standard normal distribution. If we replace the denominator with the estimate we have the ratio

$$\frac{\hat{C} - C}{\sqrt{s^2\{\hat{C}\}}} \tag{5.11}$$

which has a $t$ distribution with degrees of freedom equal to $N - t$ for the one factor model in a completely randomized design.

Thus the endpoints for the $100(1 - \alpha)\%$ confidence interval for C is:

$$\hat{C} \pm t_{\alpha/2;N-t}\sqrt{s^2\{\hat{C}\}} \tag{5.12}$$

Testing hypotheses about $C$ usually involves testing hypotheses of the form:

$$\begin{aligned} H_0 : C &= 0 && (5.13) \\ H_a : C &\neq 0 && (5.14) \end{aligned}$$

The test statistic is

$$t = \frac{\hat{C}}{\sqrt{s^2\{\hat{C}\}}} \tag{5.15}$$

The null hypothesis is rejected and the alternative accepted at the $\alpha$ level of significance if $|t| \geq t_{\alpha/2;N-t}$ or P-value $\leq \alpha$. An example follows.

**Example 5.2** *A study was conducted at a large university to compare different methods of teaching the non-calculus based elementary statistics course. Five different methods were used:*

*Method 1: Lecture method of instruction, large class*

*Method 2: Lecture method of instruction, large class with smaller problem sessions once a week*

*Method 3: Lecture method of instruction, small class*

*Method 4: Half lecture, half group work, small class*

*Method 5: All group work, small class*

*The five methods of instruction were assigned completely at random to* 30 *sections, with six sections per method. At the end of the session students rated their satisfaction with the course on a scale from* 1 *to* 15*, with larger values indicating greater satisfaction. The response variable is the class mean satisfactory score.*

*Four comparisons of the methods were formulated prior to the conduct of the study:*

1. *Large classes versus small classes (1,2 vs 3,4,5)*
2. *Comparison of large classes, with and without problem sessions (1 vs. 2)*
3. *Comparison of small classes, all group work versus other (3,4 vs 5)*
4. *Comparison of small classes, lecture versus mix of lecture and group (3 vs 4)*

*The researchers used a significance level of* 0.05 *to test each comparison.*

The values of the class mean satisfaction score for the different methods of instruction are given in Table 5.3.

Table 5.4 provides the means and standard deviations of the satisfaction data for the five methods of instruction.

Table 5.3: Satisfaction Data

| | | | | | | |
|---|---|---|---|---|---|---|
| Method 1 | 8.0 | 9.3 | 8.3 | 6.6 | 10.7 | 7.8 |
| Method 2 | 7.3 | 7.7 | 8.2 | 10.0 | 8.7 | 8.6 |
| Method 3 | 8.7 | 10.6 | 10.7 | 10.4 | 8.1 | 7.5 |
| Method 4 | 11.5 | 10.2 | 9.3 | 9.3 | 12.1 | 11.7 |
| Method 5 | 9.4 | 10.9 | 8.2 | 8.7 | 9.3 | 9.2 |

Table 5.4: Descriptives for Satisfaction Data

| Group | Mean | St.Dev. |
|---|---|---|
| Method 1 | $\overline{y}_{1\cdot} = 8.45$ | $s_1 = 1.40$ |
| Method 2 | $\overline{y}_{2\cdot} = 8.42$ | $s_2 = 0.94$ |
| Method 3 | $\overline{y}_{3\cdot} = 9.33$ | $s_3 = 1.41$ |
| Method 4 | $\overline{y}_{4\cdot} = 10.68$ | $s_4 = 1.25$ |
| Method 5 | $\overline{y}_{5\cdot} = 9.28$ | $s_5 = 0.91$ |

Table 5.5: ANOVA Table for Instruction Method Satisfaction

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Methods | 4 | 20.37 | 5.09 | 3.53 | 0.0205 |
| Error | 25 | 36.09 | 1.44 | | |
| Total (Corrected) | 29 | 56.47 | | | |

The ANOVA table is given in Table 5.5. The differences in the sample mean class satisfaction scores are significant at the 0.05 level with $F = 3.53, P = 0.0205$.

The contrasts of the instruction method population means corresponding to the four comparisons of interest are:

$$\begin{aligned}
C_1 &= (\frac{1}{2})\mu_1 + (\frac{1}{2})\mu_2 + (-\frac{1}{3})\mu_3 + (-\frac{1}{3})\mu_4 + (-\frac{1}{3})\mu_5 \\
C_2 &= (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 + (0)\mu_5 \\
C_3 &= (0)\mu_1 + (0)\mu_2 + (\frac{1}{2})\mu_3 + (\frac{1}{2})\mu_4 + (-1)\mu_5 \\
C_4 &= (0)\mu_1 + (0)\mu_2 + (1)\mu_3 + (-1)\mu_4 + (0)\mu_5 \qquad (5.16)
\end{aligned}$$

Note that the coefficients add to 0 for each of the linear combinations of means.

The estimated contrasts are:

$$\begin{aligned}
\hat{C}_1 &= (\frac{1}{2})\overline{y}_{1\cdot} + (\frac{1}{2})\overline{y}_{2\cdot} + (-\frac{1}{3})\overline{y}_{3\cdot} + (-\frac{1}{3})\overline{y}_{4\cdot} + (-\frac{1}{3})\overline{y}_{5\cdot} \\
&= (\frac{1}{2})8.45 + (\frac{1}{2})8.42 + (-\frac{1}{3})9.33 + (-\frac{1}{3})10.68 + (-\frac{1}{3})9.28 \\
&= 8.43 - 9.76 \\
&= -1.33 \qquad (5.17) \\
\hat{C}_2 &= (1)\overline{y}_{1\cdot} + (-1)\overline{y}_{2\cdot} + (0)\overline{y}_{3\cdot} + (0)\overline{y}_{4\cdot} + (0)\overline{y}_{5\cdot} \\
&= (1)8.45 + (-1)8.42 + (0)9.33 + (0)10.68 + (0)9.28 \\
&= 0.03 \qquad (5.18) \\
\hat{C}_3 &= (0)\overline{y}_{1\cdot} + (0)\overline{y}_{2\cdot} + (\frac{1}{2})\overline{y}_{3\cdot} + (\frac{1}{2})\overline{y}_{4\cdot} + (-1)\overline{y}_{5\cdot} \\
&= (0)8.45 + (0)8.42 + (\frac{1}{2})9.33 + (\frac{1}{2})10.68 + (-1)9.28 \\
&= 10.00 - 9.28 \\
&= 0.72 \qquad (5.19) \\
\hat{C}_4 &= (0)\overline{y}_{1\cdot} + (0)\overline{y}_{2\cdot} + (1)\overline{y}_{3\cdot} + (-1)\overline{y}_{4\cdot} + (0)\overline{y}_{5\cdot} \\
&= (0)8.45 + (0)8.42 + (1)9.33 + (-1)10.68 + (0)9.28 \\
&= -1.35 \qquad (5.20)
\end{aligned}$$

The estimated variances of the estimated contrasts are from 5.11:

$$\begin{aligned}
s^2\{\hat{C}_1\} &= (1.44)(\frac{(1/2)^2}{6} + \frac{(1/2)^2}{6} + \frac{(-1/3)^2}{6} + \frac{(-1/3)^2}{6} + \frac{(-1/3)^2}{6}) \\
&= 0.20 \\
s^2\{\hat{C}_2\} &= (1.44)(\frac{(1)^2}{6} + \frac{(-1)^2}{6} + \frac{(0)^2}{6} + \frac{(0)^2}{6} + \frac{(0)^2}{6})
\end{aligned}$$

$$
\begin{aligned}
&= 0.48 \\
s^2\{\hat{C}_3\} &= (1.44)(\frac{(0)^2}{6} + \frac{(0)^2}{6} + \frac{(1/2)^2}{6} + \frac{(1/2)^2}{6} + \frac{(-1)^2}{6}) \\
&= 0.36 \\
s^2\{\hat{C}_4\} &= (1.44)(\frac{(0)^2}{6} + \frac{(0)^2}{6} + \frac{(1)^2}{6} + \frac{(-1)^2}{6} + \frac{(0)^2}{6}) \\
&= 0.48
\end{aligned}
$$

The small and large classes will be compared first. The null and alternative hypotheses are

$$
\begin{aligned}
H_0 : C_1 &= 0 \\
H_a : C_1 &\neq 0
\end{aligned}
$$

The observed value of the test statistic is $t = \frac{\hat{C}_1}{s\{\hat{C}_1\}} = \frac{-1.33}{\sqrt{0.20}} = -2.97$. For a significance level of 0.05 the upper 0.05/2 probability point is $t_{\alpha/2;N-t} = t_{0.05/2;30-5} = 2.060$. Since $|-2.97| \geq 2.060$, the alternative hypothesis is accepted and it is concluded that there is a difference in mean satisfaction between the small and large classes. The 95 percent confidence interval for $C_1$ is $-1.33 \pm (2.060)(\sqrt{0.20})$ or $-1.33 \pm 0.92$ or

$$-2.25 \leq C_1 \leq -0.41$$

Thus we are 95% confident that small classes result on average anywhere between 0.41 and 2.25 points higher on the satisfaction scale compared to large classes.

The second comparison is a comparison between the large classes for problem session effect. The null and alternative hypotheses are:

$$
\begin{aligned}
H_0 : C_2 &= 0 \\
H_a : C_2 &\neq 0
\end{aligned}
$$

The observed value of the test statistic is $t = \frac{\hat{C}_2}{s\{\hat{C}_2\}} = \frac{0.03}{\sqrt{0.48}} = 0.04$. Since $|0.04| < 2.060$ there is not enough evidence of an effect of problem session on student satisfaction among the large class methods. The 95% confidence interval for $C_2$ is $0.03 \pm (2.060)(\sqrt{0.48})$ or $0.03 \pm 1.43$ or

$$-1.40 \leq C_2 \leq 1.46$$

The interval includes 0, indicating a possibility of no difference in mean satisfaction between the large classes with and without the problem sessions.

The contrast $C_3$ is a comparison of satisfaction among the small classes, those with all group work versus those with some group work and no group work.

$$\begin{aligned} H_0 : C_3 &= 0 \\ H_a : C_3 &\neq 0 \end{aligned}$$

The observed value of the test statistic is $t = \frac{\hat{C}_3}{s\{\hat{C}_3\}} = \frac{0.72}{\sqrt{0.36}} = 1.20$. Since $|1.20| < 2.060$ there is not enough evidence of a difference in satisfaction between those small classes doing all group work and those small classes doing some or no group work. The 95% confidence interval for $C_3$ is $0.72 \pm (2.060)(\sqrt{0.36})$ or $0.72 \pm 1.24$ or

$$-0.52 \leq C_3 \leq 1.96$$

The interval includes 0, indicating a possibility of no difference in mean satisfaction between the small classes with all group work and those with some or no group work.

The contrast $C_4$ is a comparison of satisfaction among the small classes, those with all lecture versus those with some lecture or no lecture.

$$\begin{aligned} H_0 : C_4 &= 0 \\ H_a : C_4 &\neq 0 \end{aligned}$$

The observed value of the test statistic is $t = \frac{\hat{C}_4}{s\{\hat{C}\}} = \frac{-1.35}{\sqrt{0.48}} = -1.95$. Since $|-1.95| < 2.060$ there is not quite enough evidence of a difference in satisfaction between those small classes doing all lecture and those doing some lecture or none. The 95% confidence interval for $C_4$ is from $-1.35 \pm (2.060)(\sqrt{0.48})$ or $-1.35 \pm 1.43$ or

$$-2.78 \leq C_4 \leq 0.08$$

The interval includes 0, indicating a possibility of no difference in mean satisfaction between the small classes with all lecture versus those with some lecture and none.

## 5.3 Effect of Multiple Testing on Type I Error Rate and Confidence Levels

In the procedures described in the last section $\alpha$ represents the pre-assigned probability of making a Type I error for a particular test, called the significance level of the test. Recall that a Type I error is made by concluding that the alternative hypothesis is true when in fact the null hypothesis is true. $1 - \alpha$ is the pre-assigned confidence level associated with each confidence interval. Recall that the confidence level is the pre-assigned probability of the interval containing the population parameter that the interval is estimating.

Of interest in multiple testing is the overall or **experimentwise significance level**, denoted by $\alpha_e$ and the overall or **experimentwise confidence level** denoted by $CL_e$.

The experimentwise wise significance level $\alpha_e$ is defined to be the probability, assuming that all true means are the same, of at least one Type I error among the $m$ tests. It can be shown that if each of the $m$ hypothesis tests is carried out at the $\alpha$ level, then

$$\alpha_e \leq m\alpha$$

In the context of multiple testing $\alpha$ is now called the **comparison wise Type I error rate** or significance level. Thus if each of $m = 6$ tests is conducted at the $\alpha = 0.05$ comparison wise significance level then the experimentwise error rate $\alpha_e \leq (6)(0.05) = 0.3$. The probability of at least one Type I error among the 6 tests can be as high as 0.3. If there are $t = 5$ treatments and all $m = 10$ tests are conducted then the experimentwise error rate can be as high as $10(0.05) = 0.5$. This is the price one pays for multiple testing. The more tests that one performs the greater the likelihood of concluding at least one significant result if in fact there are no differences among the treatment means.

A similar situation holds for confidence interval estimation. If $m$ confidence intervals are calculated for differences in population means, then the **experimentwise or overall confidence level**, $CL_e$, is defined to be the probability that all $m$ confidence intervals are correct. If the $m$ confidence intervals are conducted, each at confidence level $(1 - \alpha)$ then it can be shown that

$$CL_e \geq 1 - m\alpha$$

In this context $(1 - \alpha)$ is called the **comparison wise confidence level**. So for example if $m = 6$ and the 95% confidence level is used for each of the 6 intervals, then the probability that all 6 intervals are correct is not 95%, but can be as low as $1 - (6)(.05) = 0.7$ or 70%.

There are several methods that have been proposed to reduce the size of the experimentwise error rate, $\alpha_e$ when conducting several tests (or to increase the experimentwise confidence level, $CL_e$, when constructing several intervals). Two of these methods are discussed in Sections 5.4 and 5.5.

## 5.4 Bonferroni method

The Bonferroni approach can be used for general contrasts as well as for pairwise comparisons. The Bonferroni approach recognizes that the experimentwise error rate is

$$\alpha_e \leq m\alpha$$

where $\alpha$ is the comparison wise significance level used for each test. So suppose we want the experimentwise error rate to be at most $\alpha$, rather than $m\alpha$. Then clearly if we carry out each test at comparison level of $\frac{\alpha}{m}$ rather than $\alpha$ we have

$$\alpha_e \leq m\frac{\alpha}{m} = \alpha$$

So if we want the experimentwise error rate to be at most 0.05, then choose the comparison wise level to be $0.05/m$. If $m = 6$ then each $t$ test should be carried out at the comparison wise error rate of $\frac{0.05}{6} = 0.008$. Thus in general to insure that the experimentwise error rate is at most some pre-specified $\alpha$ level for $m$ tests use $\frac{\alpha}{m}$ for the comparison wise error rate.

### 5.4.1 Set of $m$ Contrasts

According to the Bonferroni method, for a set of $m$ pre-planned contrasts, $C_1, C_2, ..., C_m$ the null hypothesis $H_0 : C_i = 0$ would be rejected in favor of the alternative $H_0 : C_i \neq 0$ if $|t| \geq t_{\alpha/2m;N-t}$ where

$$t = \frac{\hat{C}_i}{\sqrt{s^2\{\hat{C}_i\}}} \tag{5.21}$$

Or equivalently reject the null hypothesis for a contrast if P-value $\leq \alpha/2m$. These $m$ tests would have an experimentwise error rate $\alpha_e \leq \alpha$.

The endpoints for the Bonferroni adjusted confidence intervals are:

$$\hat{C}_i \pm t_{\alpha/2m;N-t}\sqrt{s^2\{\hat{C}_i\}} \tag{5.22}$$

These $m$ confidence intervals would have an experimentwise confidence level of at least $1 - \alpha$.

Note that the $t$ percentile is $t_{\frac{\alpha}{2m};N-t}$, the upper $\frac{\alpha}{2m}$ percentile from a $t$ distribution with $N - t$ degrees of freedom. The necessary $t$ probability point for the Bonferroni procedures will not generally be found in the usual $t$ table, Table A.2, since the right tail probability $\frac{\alpha}{2m}$ will usually not correspond to one of the listed right tail probabilities. The appropriate $t$ probability points for the Bonferroni procedures can be obtained from either Table A.3 or Table A.4. Use Table A.3 if the desired experimentise error rate, $\alpha_e$ is to be at most 0.05 (or the desired experimentwise confidence level is to be at least 0.95). Use Table A.4 if the desired experimentise error rate, $\alpha_e$ is to be at most 0.01 (or the desired experimentwise confidence level is to be at least 0.99). Enter either table with the appropriate degrees of freedom $N - t$ for MSE and $m$ equal to the number of comparisons.

The instruction example (Example 5.2) will be reconsidered here. There were $m = 4$ contrasts of interest. Thus to ensure that the experimentwise error rate is at most 0.05 the values of the test statistics need to be compared to $t_{0.05/2(4);30-5} = 2.69$ from Table A.3. Note that the value 2.69 is greater than the critical value used previously of 2.069, and thus the Bonferroni procedure is more conservative. The values of the $t$ statistics for the four contrasts $C_1$, $C_2$, $C_3$, and $C_4$, respectively, were -2.97, 0.04, 1.20, and -1.95. As before only the comparison of satisfaction for the small and large classes, $C_1$, is significant. Since the Bonferroni $t$ probability point is different it is possible for different conclusions to be reached. The Bonferroni confidence intervals for the contrasts $C_1$, $C_2$, $C_3$, and $C_4$ with experimentwise confidence level of at least 95% are

$$\begin{array}{ccc} (-1.33) & \pm & 2.69(\sqrt{0.20}) \\ (0.03) & \pm & 2.69(\sqrt{0.48}) \\ (0.72) & \pm & 2.69(\sqrt{0.36}) \\ (-1.35) & \pm & 2.69(\sqrt{0.48}) \end{array}$$

or

$$\begin{array}{ccc} -1.33 & \pm & 1.20 \\ 0.03 & \pm & 1.86 \\ 0.72 & \pm & 1.61 \\ -1.35 & \pm & 1.86 \end{array}$$

Thus the four Bonferroni intervals with experimentwise confidence level of at least 95% are

$$\begin{array}{ccccc} -2.53 & \leq & C_1 & \leq & -0.13 \\ -1.83 & \leq & C_2 & \leq & 1.89 \\ -0.89 & \leq & C_3 & \leq & 2.33 \\ -3.21 & \leq & C_4 & \leq & 0.51 \end{array}$$

These intervals are wider than unadjusted intervals, again illustrating the conservative nature of the Bonferroni procedure.

### 5.4.2 All Pairwise Comparisons

For all pairwise comparisons of means using the Bonferroni method and experimentwise error rate of at most $\alpha$ one rejects the null hypothesis $H_o : \mu_i = \mu_j$ if

$$|\overline{y}_{i\cdot} - \overline{y}_{j\cdot}| \geq t_{\frac{\alpha}{2m};\nu}\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

If one wants to ensure that $m$ confidence intervals have an experimentwise confidence level of at least $1 - \alpha$ then the comparison wise confidence level for each interval should be $1 - \frac{\alpha}{m}$. The form of the Bonferroni confidence intervals is

$$(\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) \pm (t_{\frac{\alpha}{2m};\nu})\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Determination of the appropriate $t$ probability point is illustrated in Example 5.3.

**Example 5.3** *The Bonferroni confidence intervals will be illustrated with the sales data from Example 5.1. If an experimentwise confidence level of at least* 95% *is desired for the three intervals then from Table A.3 with* $\nu = 15$ *and* $m = 3$, *the appropriate t percentile* $t_{\frac{0.05}{2(3)};15} = 2.69$. *Thus the endpoints of the Bonferroni confidence intervals with experimentwise confidence level of at least* 95% *are:*

$$\begin{array}{rcl} (51.00 - 46.90) & \pm & 2.69\sqrt{2.26}\sqrt{\frac{1}{6}+\frac{1}{6}} \\ (51.00 - 48.42) & \pm & 2.69\sqrt{2.26}\sqrt{\frac{1}{6}+\frac{1}{6}} \\ (46.90 - 48.42) & \pm & 2.69\sqrt{2.26}\sqrt{\frac{1}{6}+\frac{1}{6}} \end{array}$$

*or*

$$\begin{array}{rcl} 4.10 & \pm & 2.33 \\ 2.58 & \pm & 2.33 \\ -1.52 & \pm & 2.33 \end{array}$$

*Thus the three intervals are:*

$$\begin{array}{rcccl} 1.77 & \leq & \mu_1 - \mu_2 & \leq & 6.43 \\ 0.25 & \leq & \mu_1 - \mu_3 & \leq & 4.91 \\ -3.85 & \leq & \mu_2 - \mu_3 & \leq & 0.81 \end{array}$$

Note that the $t$ probability point used for the Bonferroni intervals, 2.69, is larger than the $t$ probability point used for the usual $t$ intervals, 2.131. This results in larger error margins for the differences in the sample means and thus wider confidence intervals for the Bonferroni intervals. In general wider confidence intervals might result in different conclusions because wider intervals are more likely to include 0. However for this example the conclusions are the same. The box cover with the sports hero results in the greatest mean sales. There is no evidence of a difference in mean sales between the box cover with a child and that with a bowl of cereal. We are at least 95% confident in this set of conclusions being correct. A comparison of the unadjusted $t$ and Bonferroni procedures for pairwise comparisons is made in Section 5.6.

## 5.5 Tukey-Kramer Method for Pairwise Comparisons

The Bonferroni method uses a larger $t$ percentile to ensure that the experimentwise error rate is at most some prescribed value. The Tukey-Kramer multiple comparison method also uses a larger percentile, but one from an entirely different distribution, called the Studentized Range distribution.

The null hypothesis $H_o : \mu_i = \mu_j$ for a pairwise comparison is rejected if

$$|\overline{y}_{i\cdot} - \overline{y}_{j\cdot}| \geq \frac{q_{\alpha;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $q_{\alpha;\nu,t}$ is the upper $\alpha$ probability point from the Studentized Range Distribution, tabulated in Tables A.5 ($\alpha = 0.01$) and A.6 ($\alpha = 0.05$). The tables depend upon a degrees of freedom parameter, $\nu$, which for the one-way ANOVA is degrees of freedom associated with $MSE$ and $t$, the number of means being compared.

If the group sizes $n_i$ are all equal, then the experimentwise error rate for the set of all pairwise comparisons is exactly $\alpha$, that is $\alpha_e = \alpha$. If the group sizes are not equal then $\alpha_e \leq \alpha$. Thus one can prescribe the experimentwise error rate or an upper bound for it.

The Tukey-Kramer confidence intervals for the differences $\mu_i - \mu_j$ are

$$(\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

The confidence intervals have an experimentwise confidence level, $CL_e$ of exactly $(1 - \alpha)$ if the group sizes are identical. If the group sizes are not identical then the $CL_e \geq (1 - \alpha)$.

**Example 5.4** *The Tukey-Kramer confidence intervals will be illustrated with the sales data from Example 5.1. If an experimentwise confidence level of* 95% *is desired then Table A.6 has for* $\nu = 15$ *and* $t = 3$, $q_{0.05;15,3} = 3.67$. *Thus the endpoints of the Tukey-Kramer confidence intervals are*

$$\begin{array}{rcl} (51.00 - 46.90) & \pm & \frac{3.67}{\sqrt{2}}\sqrt{2.26}\sqrt{\frac{1}{6} + \frac{1}{6}} \\ (51.00 - 48.42) & \pm & \frac{3.67}{\sqrt{2}}\sqrt{2.26}\sqrt{\frac{1}{6} + \frac{1}{6}} \\ (46.90 - 48.42) & \pm & \frac{3.67}{\sqrt{2}}\sqrt{2.26}\sqrt{\frac{1}{6} + \frac{1}{6}} \end{array}$$

*or*

$$\begin{array}{rcl} 4.10 & \pm & 2.25 \\ 2.58 & \pm & 2.25 \\ -1.52 & \pm & 2.25 \end{array}$$

*Thus the three intervals are:*

$$\begin{array}{rcccl} 1.85 & \leq & \mu_1 - \mu_2 & \leq & 6.35 \\ 0.33 & \leq & \mu_1 - \mu_3 & \leq & 4.83 \\ -3.77 & \leq & \mu_2 - \mu_3 & \leq & 0.73 \end{array}$$

Notice that the multiplier, $\frac{3.67}{\sqrt{2}} = 2.60$, for the Tukey intervals, is larger than the multiplier of 2.131 for the $t$ intervals of Example 5.1, but slightly smaller than the multiplier of 2.69 used in the Bonferroni intervals. Thus the Tukey-Kramer intervals are wider than those for the unadjusted or usual $t$ procedure but not as wide as those for the Bonferroni procedure. The conclusions are the same as for the unadjusted $t$ and Bonferroni procedures. However the conclusions can be different than the two other procedures. A comparison is made between the three procedures in Section 5.6.

## 5.6 Summary and Comparison of the Three Methods

1 Depending upon the research objectives, comparisons among means following a significant F ratio may involve all pairwise comparisons or more general comparisons among the means (contrasts). The type of comparison to be made is specified in the research protocol before the data is collected.

2 A multiple comparison procedure refers to a procedure for making multiple comparisons of means. One possible procedure is to perform a series of $t$ tests for the comparisons. (See sections 5.2.1, 5.2.2). The $t$ tests can be used to make the set of all pairwise comparisons or to make a set of more general comparisons which might include some pairwise comparisons. The usual $t$ tests do not control or adjust for the experimentwise significance level. Subsequently these procedures will be called **unadjusted** $t$ procedures. The Bonferroni procedure controls or adjusts for the experimentwise significance level and can be used if the set of comparisons is all pairwise or some other set of more general contrasts. The group sizes do not have to be of equal size. The Tukey-Kramer procedure is used to control or adjust the experimentwise significance level when the set of comparisons is all pairwise. The group sizes do not have to be of equal size. When the group sizes are the same the Tukey-Kramer procedure is then called the Tukey procedure.

3 Suppose that the set of comparisons of interest is the set of all pairwise comparisons. Then the confidence intervals for all three methods (unadjusted t, Bonferroni, and Tukey-Kramer) can be written as

$$\begin{array}{ll} estimate & \pm(margin\ of\ error) \\ estimate & \pm(multiplier) \times SE(estimate) \end{array}$$

where $estimate = \overline{y}_{i\cdot} - \overline{y}_{j\cdot}$ and $SE(estimate) = \sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$.

Table 5.6 provides the multipliers for the three methods for the set of all pairwise comparisons for $\alpha = 0.05$, when there are $t = 3$ treatments ($m = 3$ comparisons) comparisons and when there are $t = 4$ treatments ($m = 6$ comparisons) for various degrees of freedom $\nu$.

Note that the unadjusted $t$ procedure multiplier is smallest and the Bonferroni procedure multiplier is largest, with Tukey-Kramer procedure multiplier in between. Thus if used with the same set of data (same MSE), the Bonferroni and Tukey-Kramer procedures will have wider intervals than those for the unadjusted t procedure. The Bonferroni intervals will be wider than the Tukey intervals but less so with increasing degrees of freedom. Wider intervals are more likely to include 0 and thus less likely

Table 5.6: Multipliers for Multiple Comparison Procedures

| | | | Multiplier for $\alpha = 0.05$ | | |
|---|---|---|---|---|---|
| t | m | $\nu$ | Unadjusted t | Bonferroni | Tukey-Kramer |
| 3 | 3 | 6 | 2.45 | 3.29 | 3.07 |
| | | 9 | 2.26 | 2.93 | 2.79 |
| | | 12 | 2.18 | 2.78 | 2.67 |
| | | 15 | 2.13 | 2.69 | 2.60 |
| | | 18 | 2.10 | 2.64 | 2.55 |
| | | 21 | 2.08 | 2.60 | 2.52 |
| | | 24 | 2.06 | 2.57 | 2.49 |
| | | 27 | 2.05 | 2.55 | 2.48 |
| 4 | 6 | 8 | 2.31 | 3.48 | 3.20 |
| | | 12 | 2.18 | 3.15 | 2.97 |
| | | 16 | 2.12 | 3.01 | 2.86 |
| | | 20 | 2.09 | 2.93 | 2.80 |
| | | 24 | 2.06 | 2.88 | 2.76 |
| | | 28 | 2.05 | 2.84 | 2.73 |

to conclude significance difference in means. The Bonferroni and Tukey-Kramer procedures are thus more conservative than the unadjusted $t$ procedure with the Bonferroni procedure being more conservative than the Tukey-Kramer procedure. Being more conservative is a good characteristic of a procedure if in fact there are no differences among the means, but not good if there are differences among the means. If there are differences among the means somewhere then the Bonferroni and Tukey-Kramer will have less statistical power to detect those differences than the unadjusted t procedure. Thus a balance has to be struck between Type I error rate and statistical power. The Tukey-Kramer procedure is often used because it offers better protection against the Type I error rate than the unadjusted $t$ test but has better statistical power than the Bonferroni procedure. However a researcher might use the unadjusted $t$ procedure if the purpose of the study is to select among several proposed treatments a few for further study. The Type I error rate may not of major concern in this situation. The Tukey-Kramer or the Bonferroni procedures would be used in future studies of the selected treatments.

## 5.7 P-values for Bonferroni and Tukey Methods

Computer programs, such as SAS and SPSS will report $t$ statistics and two-sided P-values for the Bonferroni and Tukey procedures as well as confidence intervals. The P-values have been adjusted so that conclusions based on these are equivalent to conclusions based on the confidence intervals. The $t$ statistics

Table 5.7: P-values for Pairwise Comparisons of Box Covers

| Cover | Cover | Mean Diff | Std.Error | DF | t Value | P-value | Bonf P | Tukey P |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3.1 | 0.87 | 15 | 4.73 | 0.0003 | 0.0008 | 0.0007 |
| 1 | 3 | 2.58 | 0.87 | 15 | 2.98 | 0.0094 | 0.0281 | 0.0239 |
| 2 | 3 | -1.52 | 0.87 | 15 | -1.75 | 0.1007 | 0.3020 | 0.2201 |

along with P-values are given in Table 5.7 for the cereal data. "P-value" refers to the unadjusted P-value, while "Bon P" and "Tukey P" refer to the Bonferroni and Tukey-Kramer adjusted P-values.

## 5.8 SAS Code for Chapter 5

### 5.8.1 Example 5.1

```
*  Input sales (number of boxes) for
   three types of box covers;

data CEREAL;
  input BoxCover $ NumberBoxes;
datalines;
SportsHero  52.4
SportsHero  47.8
SportsHero  52.4
SportsHero  51.3
SportsHero  50.0
SportsHero  52.1
Child  50.1
Child  45.2
Child  46.0
Child  46.5
Child  47.4
Child  46.2
CerealBowl  49.2
CerealBowl  48.3
CerealBowl  49.0
CerealBowl  47.2
CerealBowl  48.6
CerealBowl  48.2
;

*  Use proc glm to obtain results of F test
   for overall differences in mean sales and
   to obtain pairwise comparisons using
   Multiple t, Bonferroni, and Tukey procedures;
proc glm data = CEREAL;
  class BoxCover;
  model NumberBoxes = BoxCover;
  lsmeans BoxCover / cl pdiff t;
  lsmeans BoxCover / cl pdiff t adjust = Bonferroni;
  lsmeans BoxCover / cl pdiff t adjust = tukey;
run;
```

# Problems for Chapter 5

5.1* Suppose that there are $t = 5$ treatments in a study with 6 replications per treatment. Suppose that the F test for overall differences is significant and interest is in making all pairwise comparisons by constructing confidence intervals for differences in pairs of means? Suppose that MSE is 36.

a. How many possible intervals are there? That is what is the value of $m$?

b. What is the appropriate $t$ percentile for the unadjusted $t$ procedure if the comparison wise error rate is set at 0.01. What is the margin of error associated with each of the differences in sample means? What is the lower bound on the the experimentwise confidence level?

c. Suppose that the Bonferroni procedure is to be used with the experimentwise confidence level required to be at least 0.99. What is the appropriate t-percentile? What is the margin of error associated with each of the differences in sample means?

d. Suppose that the Tukey procedure is to be used with experimentwise confidence level specified to be exactly 0.99. What is the appropriate probability point from the Studentized Range distribution? What is the margin of error associated with each of the differences in sample means?

e. Which procedure would result in the widest confidence intervals? Which would result in the narrowest confidence intervals? Explain.

5.2* Two students, Cheryl Butterworth and Josh Hiller, performed an experiment to study the effect of beverage type on the amount of time for ice cubes to melt. Types of beverage were coca-cola, orange juice, and water. The beverages were left out over night to set them at a constant temperature. Fifteen ice cubes of approximately the same size were randomly assigned to fifteen identical cups. Equal amounts of beverage, five of each kind, were randomly assigned to the cups. The amount of time (minutes) for the ice cubes to melt was recorded and given below.

| | | | | | |
|---|---|---|---|---|---|
| 1. Coca cola | 19 | 17 | 15 | 14 | 18 |
| 2. Orange Juice | 27 | 28 | 30 | 26 | 27 |
| 3. Water | 10 | 11 | 13 | 7 | 9 |

This is the data from Problem 4.4 in the Chapter 4 exercises. The sample mean melting times for the Coca-cola, orange juice, and water treatments, are, respectively, 16.6, 27.6, and 10.0. The F test for overall differences in the beverages on melting time is significant ($F = 102.22, P < 0.0001$). Mean squared error from the ANOVA is 3.87.

a. Construct the Tukey-Kramer confidence intervals for all possible pairwise comparisons of the three population mean melting times. Use an experimentwise confidence level of 99%. Which pairs of means are significantly different?

b. What does the 99% experimentwise confidence level mean?

c. Would your confidence intervals be wider or narrower if the experimentwise confidence level was 95%? Explain.

5.3* Suppose that a study has only $t = 2$ treatments and thus there is only $m = 1$ pairwise comparison of interest. Then since $\alpha/2$ and $\alpha/2m$ are the same when $m = 1$ the t percentiles would be the same for the Multiple t and Bonferroni procedures and thus the two procedures give the same results. Show for the case when $t = 2$, comparison wise confidence of 0.95, and $\nu = 20$ that the unadjusted t procedure and the Tukey-Kramer procedure give the same multiplier on the standard error and thus the same confidence interval.

5.4* In the article "Sex differences in viewing sexual stimuli: An eye-tracking study in men and women" (*Hormones and Behavior* [2007]: Vol. 51, pgs 524-533 ) researchers compared three groups of heterosexuals: males, normal (menstrual) cycling females (NC), and oral contracepting females (OC), on various responses to viewing sexual stimuli. Stimuli were sexually explicit photos of heterosexual couples engaged in oral sex or intercourse. Because researchers thought that any group differences found in the current study could be due to differences in participants' previous experience viewing sexually explicit stimuli, sexual attitudes, sexual motivation, or comfort with visual sexual stimuli, they compared the three groups on these variables using a one factor ANOVA. Only the results of the comparison of the three groups on sexual motivation as measured by the frequency of sexual thoughts and desire to engage in sexual activity in the previous month is given here. The table below gives the means and standard deviations of the sexual motivation variable for the three groups.

| Group | n | Mean | Standard Deviation |
|---|---|---|---|
| Men | 15 | 5.2 | 0.71 |
| NC Women | 15 | 3.8 | 1.18 |
| OC Women | 14 | 4.64 | 0.73 |

The F ratio for comparing the means was statistically significant with F = 8.72 and P = 0.001.

a. Use information provided in the table to calculate MSE (See Chapter 4).

b. Use Tukey-Kramer method to calculate all possible pairwise confidence intervals with experiment-wise confidence level of at least 0.95.

Use Table A.6 to obtain the upper 0.05 percentile from the Studentized Range Distribution. Use the value corresponding to $\nu = 40$ in the table, approximating the true value of $\nu = 41$.

c. Use your intervals from part(a) to determine which means are statistically significantly different.

5.5* Dowdy and Wearden ([9], page 316) describe an experiment comparing five different types of toothpastes on abrasiveness. “The variable of interest is the number of minutes until mechanical brushing of a material similar to tooth enamel exhibits wear. The five toothpastes are all the same except for the absence or presence of certain additives. The material is assigned randomly to the treatments.” The additives for the five different toothpastes are given below.

| Toothpaste | Additive |
|---|---|
| 1 | Whitener |
| 2 | None |
| 3 | Fluoride |
| 4 | Fluoride with freshener |
| 5 | Whitener with freshener |

Group means are given below. Group sizes are all 4.

| Toothpaste | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Mean | 49.4 | 49.8 | 52.8 | 54.0 | 46.6 |

An ANOVA table is given below.

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Toothpaste | 4 | 136.8 | 34.20 | 39.8 | |
| Error | 15 | 13.0 | 0.86 | | |

The investigator planned the following contrasts:

- Mean of groups with additives (1,3,4,5) versus mean without additive (2)
- Mean of groups with whitener (1,5) versus mean of groups with fluoride (3,4)
- Mean of group with Whitener (1) versus mean of group Whitener with freshener (5)
- Mean of group Fluoride (3) versus mean of group Fluoride with freshener (4)

a. Write out the mathematical forms of the four contrasts describe above, that is write out as linear combinations of population means.

b. Determine the estimate of each contrast from part(a) along with the standard error of the estimate.

c. Test each of the following contrasts using a two-sided t test. Be sure to give the null and alternative hypotheses for each test in mathematical form. Give the value of the test statistic. Compare the absolute value of the observed value of the test statistic to an appropriate Bonferroni adjusted t percentile in order to obtain an experiment-wise significance level of at most 0.05. In each case draw a conclusion in context.

   i. Mean of groups with additives (1,3,4,5) versus mean without additive (2).
   ii. Mean of groups with whitener (1,5) versus mean of groups with fluoride (3,4).
   iii. Mean of group with Whitener (1) versus mean of group Whitener with freshener (5).
   iv. Mean of group Fluoride(3) versus mean of group Fluoride with freshener (4).

5.6* In the article "An Ex Vivo Study to Investigate Bond Strengths of Different Tooth Types" (*Journal of Orthodontics* [2001]: Vol. 28, pgs 59 - 65 ) researchers compared the shear bond strengths of orthodontic brackets bonded to six types of teeth using the bonding agent Right-On. The bonding tests were conducted using specimens of extracted human teeth. The six types of teeth used were

A Upper incisors
B Upper canines
C Upper premolars
D Lower incisors
E Lower canines
F Lower premolars

A descriptive summary of the bond strengths (MPa = mega pascals) is given in the table below.

| Tooth Type | n | Mean | SD |
|---|---|---|---|
| A | 20 | 6.95 | 2.85 |
| B | 20 | 12.27 | 2.52 |
| C | 16 | 11.87 | 2.24 |
| D | 18 | 8.95 | 1.63 |
| E | 8 | 12.07 | 2.78 |
| F | 26 | 10.94 | 2.33 |

An ANOVA table is given below.

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| ToothType | 5 | 411.29 | 996.45 | 14.34 | 0.000 |
| Error | 102 | 585.16 | 5.74 | | |
| Total | 107 | 996.45 | | | |

There is evidence at the 0.05 level of an effect of type of tooth on bonding strength. The authors conducted Tukey-Kramer pairwise comparisons of the 6 types of teeth using a 95% experiment-wise confidence level. Endpoints for the confidence intervals are given in the table below.

| Group | Intervals for (column level mean) - (row level mean | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | E |
| B | -7.529 | | | | |
| | -3.126 | | | | |
| C | -7.263 | -1.935 | | | |
| | -2.593 | 2.734 | | | |
| D | -4.265 | 1.063 | 0.533 | | |
| | 0.258 | 5.586 | 5.316 | | |
| E | -8.045 | -2.717 | -3.219 | -6.087 | |
| | -2.221 | 3.107 | 2.809 | -0.171 | |
| F | -6.068 | -0.741 | -1.282 | -4.129 | -1.680 |
| | -1.928 | 3.400 | 3.142 | 0.140 | 3.949 |

a. Do the calculation for the interval comparing types B and E and compare with the tabled result. (Use $\nu = 100$ from the appropriate Studentized Range table.

b. The highest sample mean bonding strength occurred with tooth type B. Use the results of the Tukey confidence intervals in the table to compare bonding strength between tooth type B and bonding strengths for all other possible tooth types. Draw conclusions.

c. In the context of this problem what does the 95% experiment-wise confidence interval mean?

5.7 This example is based on data reported in Oehlert ([24], page 61) on leaflet angle (degrees) from plants in the genus *Albizzia* after exposure to red light. Certain plants from this genus have the ability to fold and unfold their leaves under various light conditions. The researcher selected 15 leaves and subjected them to red light for 3 minutes. The leaves were then divided at random into three groups with 5 leaves per group. The groups were defined by the length of time, 30, 45, or 60 minutes after exposure to the red light when leaflet angle was measured for the leaves. The data are given below.

| Delay (minutes) | Angle (degrees) | | | | |
|---|---|---|---|---|---|
| 30 | 140 | 138 | 140 | 138 | 142 |
| 45 | 140 | 150 | 120 | 128 | 130 |
| 60 | 118 | 130 | 128 | 118 | 118 |

There is evidence of a difference in mean angle among the delay times ($F = 6.56, P = 0.0119$) at the 5% significance level. Mean angle for the three delay times of 30, 45, and 60, are respectively 139.6, 133.6, and 122.4 degrees. Construct Tukey confidence intervals for differences in mean angle among all possible pairs of delay times. Use the 95% overall confidence level. Draw conclusions within the context of this study.

5.8 Suppose that after a significant F ratio at the 0.05 level in a one factor completely randomized design analysis with $t = 4$ treatments Bonferroni confidence intervals are constructed to make all possible pairwise comparisons using an experiment-wise confidence level of (at least) 98%.

a. How many possible intervals are there?

b. What is the comparison-wise confidence level for each interval?

# Chapter 6

# Two Factor Completely Randomized Design - Equal Replications

## 6.1 Introduction and Notation

In this chapter we will consider studies that employ two factors: factor A with $a$ levels denoted by $A_1, A_2, \ldots, A_a$ and factor B with $b$ levels denoted by $B_1, B_2, \ldots, B_b$. The treatments given to the experimental units represent all combinations of the levels of $A$ and $B$ and are said to be in a "factorial" arrangement. For example, A may represent amount of water given to a plant and B amount of fertilizer. Then the word "treatment" refers to a combination of level of water and level of fertilizer.

If there are $a = 2$ levels of $A$ and $b = 3$ levels of $B$ then there are $(2)(3) = 6$ treatments which would be denoted by

$$A_1B_1, A_1B_2, A_1B_3, A_2B_1, A_2B_2, A_2B_3$$

In the completely randomized design studied in this chapter the 6 treatments would be assigned completely at random to the N experimental units. It is assumed in this chapter that the number of replications per treatment is the same and equal to $n$.

## 6.2 Example and the No Interaction Model

Suppose that in an agricultural experiment factor A is type of fertilizer with $a = 2$ levels and factor B is a second factor of interest, watering regimen, with $b = 2$ levels. Thus the four treatments are denoted by

$$A_1B_1, A_1B_2, A_2B_1, A_2B_2$$

Table 6.1: Sample Randomization

| | | | |
|---|---|---|---|
| $A_1B_2$ | $A_2B_1$ | $A_1B_1$ | $A_2B_1$ |
| $A_2B_2$ | $A_1B_1$ | $A_2B_2$ | $A_1B_2$ |
| $A_1B_2$ | $A_2B_1$ | $A_1B_1$ | $A_2B_2$ |
| $A_2B_1$ | $A_1B_2$ | $A_2B_2$ | $A_1B_1$ |

.

Suppose that these four treatments are to be assigned completely at random to 16 plots laid out in a rectangular arrangement with each treatment being applied to 4 plots. A schematic of the resulting randomization is given in Table 6.1 The response variable is tomato production in pounds for a plant.

Let the true mean tomato production (in pounds) for the four treatments be

$$\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}$$

where $\mu_{ij}$ is the mean production for treatment $A_iB_j$.

We can apply the means model from Chapter 4 for each of the treatments resulting in

$$y_{11k} = \mu_{11} + \epsilon_{11k} \tag{6.1}$$

$$y_{12k} = \mu_{12} + \epsilon_{12k} \tag{6.2}$$

$$y_{21k} = \mu_{21} + \epsilon_{21k} \tag{6.3}$$

$$y_{22k} = \mu_{22} + \epsilon_{22k} \tag{6.4}$$

for $k = 1, \ldots, 4$. The errors $\epsilon_{ijk}$ represent as in Chapter 4 the effects of extraneous variables on the tomato production of a plant, such as particular plot soil fertility, genetic composition of the particular plant.

Suppose for the sake of discussion in this chapter that we know the true treatment means to be

$$\mu_{11} = 10, \mu_{12} = 12, \mu_{21} = 6, \mu_{22} = 8$$

These values are given in Table 6.2 along with summaries of these treatment means.

The true "marginal" mean production for fertilizer $A_1$ averaged over the two different watering regimens is $\mu_{1\cdot} = (10 + 12)/2 = 11$. Similarly $\mu_{2\cdot} = 7$ is the true "marginal" mean production for $A_2$ averaged over the two water regimens. The "grand mean" tomato production averaged over all 4 treatments is $\mu_{\cdot\cdot} = (10 + 12 + 6 + 8)/4 = 9$. The true **"main effect"** of fertilizer $A_1$ is defined to be

$$\alpha_1 = \mu_{1\cdot} - \mu_{\cdot\cdot} = 11 - 9 = 2$$

Table 6.2: Table of Treatment Means

| | Watering Regimen | | | |
|---|---|---|---|---|
| | $B_1$ | $B_2$ | | |
| Fertilizer | | | | |
| $A_1$ | $\mu_{11} = 10$ | $\mu_{12} = 12$ | $\mu_{1\cdot} = 11$ | $\alpha_1 = 2$ |
| $A_2$ | $\mu_{21} = 6$ | $\mu_{22} = 8$ | $\mu_{2\cdot} = 7$ | $\alpha_2 = -2$ |
| | $\mu_{\cdot 1} = 8$ | $\mu_{\cdot 2} = 10$ | $\mu_{\cdot\cdot} = 9$ | |
| | $\beta_1 = -1$ | $\beta_2 = 1$ | | |

Similarly $\alpha_2 = -2$ is the true **"main effect"** of fertilizer $A_2$. The true marginal means for the watering regimens $B_1$ and $B_2$, denoted by $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$, and the true main effects of the watering regimens, $\beta_1$ and $\beta_2$, are defined similarly.

Note that each treatment mean can be written as the sum of the grand mean + main effect of fertilizer + main effect of watering regimen:

$$\mu_{11} = 10 = \mu_{\cdot\cdot} + \alpha_1 + \beta_1 = 9 + 2 + (-1) = 10$$

$$\mu_{12} = 12 = \mu_{\cdot\cdot} + \alpha_1 + \beta_2 = 9 + 2 + 1 = 12$$

$$\mu_{21} = 6 = \mu_{\cdot\cdot} + \alpha_2 + \beta_1 = 9 + (-2) + (-1) = 6$$

$$\mu_{22} = 12 = \mu_{\cdot\cdot} + \alpha_2 + \beta_2 = 9 + (-2) + 1 = 8$$

Thus each observed value of the response tomato production can be written

$$y_{11k} = \mu_{11} + \epsilon_{11k} = \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + \epsilon_{11k}$$

$$y_{12k} = \mu_{12} + \epsilon_{12k} = \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + \epsilon_{12k}$$

$$y_{21k} = \mu_{21} + \epsilon_{21k} = \mu_{\cdot\cdot} + \alpha_2 + \beta_1 + \epsilon_{21k}$$

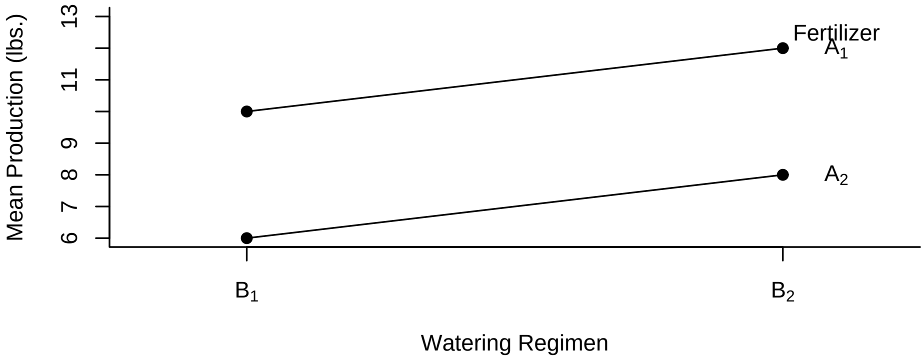
$$y_{22k} = \mu_{22} + \epsilon_{22k} = \mu_{\cdot\cdot} + \alpha_2 + \beta_2 + \epsilon_{22k}$$

for $k = 1, \ldots, 4$. This model is called the **NO INTERACTION MODEL**. Two equivalent characterizations of this model are

- Each true treatment mean can be written as a sum of the grand mean, factor A level main effect, and factor B level main effect.

- The difference between true **treatment** means at two levels of one factor do not depend upon levels of the other factor. This is perhaps the more intuitive condition.

Figure 6.1: Interaction Plot: Tomato Production - No Interaction



In the tomato production example, the difference between the true means for $B_1$ and $B_2$ at level $A_1$, $\mu_{12} - \mu_{11} = 12 - 10 = 2$, is the same as the difference between the true means for $B_1$ and $B_2$ at level $A_2$, $\mu_{22} - \mu_{21} = 8 - 6 = 2$.

Similarly the difference between the true means for $A_1$ and $A_2$ at level $B_1$, $\mu_{11} - \mu_{12} = 10 - 6 = 4$, is the same as the difference between the true means for $A_1$ and $A_2$ at level $B_2$, $\mu_{12} - \mu_{22} = 12 - 8 = 4$.

In other words, the change in mean tomato production when going from one watering regimen to the other does not depend on fertilizer or the change in tomato production when going from one fertilizer to the other does not depend on watering regimen.

The concept of no interaction can be demonstrated with a plot such as that in Figure 6.1. The plot is simply a plot of the treatment means on the vertical axis versus one of the factors on the horizontal axis. Lines are then drawn connecting values having the same values on the 2nd factor. In Figure 6.1 Watering Regimen was put on the horizontal axis and there are two lines corresponding to the two levels of fertilizer. Fertilizer could just as well have been put on the horizontal axis. If the factors do not interact then the lines will be parallel. If the factors interact then the lines will not be parallel. We shall look at an example shortly where interaction exists.

The tomato production example was a hypothetical example where it was assumed that we knew the true means and could plot them. In practice one does not know the true means and only has estimates of these, that is the sample treatment means. Thus in practice one plots the sample means. If the lines are approximately parallel then the no interaction assumption is plausible.

Table 6.3: Table of Treatment Means

| | $B_1$ | $B_2$ | | |
|---|---|---|---|---|
| $A_1$ | $\mu_{11} = 10$ | $\mu_{12} = 16$ | $\mu_{1\cdot} = 13$ | $\alpha_1 = 3$ |
| $A_2$ | $\mu_{21} = 6$ | $\mu_{22} = 8$ | $\mu_{2\cdot} = 7$ | $\alpha_2 = -3$ |
| | $\mu_{\cdot 1} = 8$ | $\mu_{\cdot 2} = 12$ | $\mu_{\cdot\cdot} = 10$ | |
| | $\beta_1 = -2$ | $\beta_2 = 2$ | | |

## 6.3 Interaction Model

Suppose that in the tomato production example the (true) treatment means are instead as in Table 6.3.

The difference between the two treatment means at $B_1$ and $B_2$ for level $A_1$ is $\mu_{12} - \mu_{11} = 16 - 10 = 6$, which is NOT equal to the difference between the true treatment means at $B_1$ and $B_2$ for level $A_2$, $\mu_{22} - \mu_{21} = 8 - 6 = 2$. Similarly the difference in true treatment means at $A_1$ and $A_2$ when watering regimen is at $B_1$, $\mu_{11} - \mu_{21} = 10 - 6 = 4$ is NOT the same as the difference in true treatment means at $A_1$ and $A_2$ when watering regimen is at $B_2$, $\mu_{12} - \mu_{22} = 16 - 8 = 8$.

Thus the change in tomato production when going from one fertilizer to another DOES DEPEND upon the watering regimen and the change in tomato production when going from one watering regimen to another depends on type of fertilizer. A graphical representation is given in Figure 6.2. The lines corresponding to the levels of fertilizer are NOT parallel.

Note also that the true treatment means CANNOT expressed as the sum of the grand mean, fertilizer main effect, and watering regimen main effect:

$$\begin{array}{lcc}
\mu_{11} \neq \mu_{\cdot\cdot} + \alpha_1 + \beta_1 & \text{or} & 10 \neq 10 + 3 + (-2) = 11 \\
\mu_{12} \neq \mu_{\cdot\cdot} + \alpha_1 + \beta_2 & \text{or} & 16 \neq 10 + 3 + 2 = 15 \\
\mu_{21} \neq \mu_{\cdot\cdot} + \alpha_2 + \beta_1 & \text{or} & 6 \neq 10 + (-3) + (-2) = 5 \\
\mu_{22} \neq \mu_{\cdot\cdot} + \alpha_2 + \beta_2 & \text{or} & 8 \neq 10 + (-3) + 2 = 9
\end{array}$$

Thus we need a more complex model to cover possible situations like this. Note that in order for the first equation above to be true we could add $(-1)$ on the right side. Thus

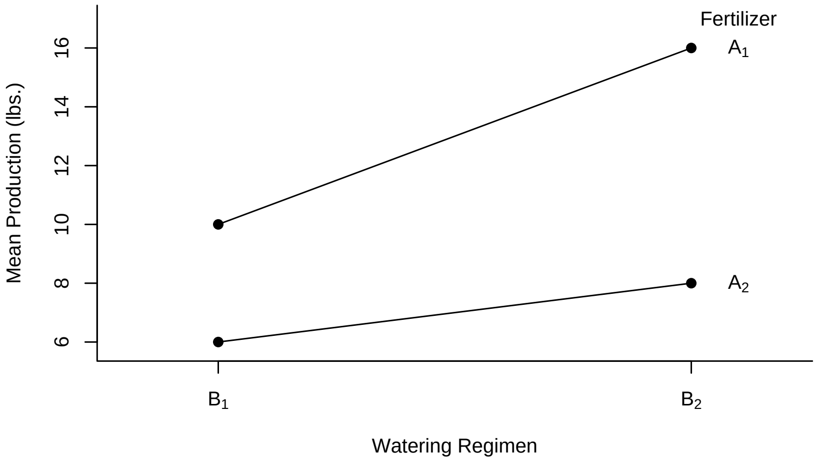
$$10 = 10 + 3 + (-2) + (-1)$$

To get $(-1)$, $\mu_{\cdot\cdot} + \alpha_1 + \beta_1$ was subtracted from $\mu_{11}$. Thus the new equation looks like

$$\mu_{11} = \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + [\mu_{11} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_1)] \quad \text{or} \quad 10 = 10 + 3 + (-2) + [-1]$$

The necessary adjustments are illustrated for all equations below:

Figure 6.2: Tomato Production - Interaction between Watering Regimen and Fertilizer



$$\begin{array}{lcc}
\mu_{11} = \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + [\mu_{11} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_1)] & \text{or} & 10 = 10 + 3 + (-2) + [-1] \\
\mu_{12} = \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + [\mu_{12} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_2)] & \text{or} & 16 = 10 + 3 + 2 + [1] \\
\mu_{21} = \mu_{\cdot\cdot} + \alpha_2 + \beta_1 + [\mu_{21} - (\mu_{\cdot\cdot} + \alpha_2 + \beta_1)] & \text{or} & 6 = 10 + (-3) + (-2) + [1] \\
\mu_{22} = \mu_{\cdot\cdot} + \alpha_2 + \beta_2 + [\mu_{22} - (\mu_{\cdot\cdot} + \alpha_2 + \beta_2)] & \text{or} & 8 = 10 + (-3) + 2 + [-1]
\end{array}$$

The adjustments of -1, 1, 1, and -1 that are made to the above inequalities to make them equalities are called **INTERACTION EFFECTS** and are denoted by $\alpha\beta_{ij}$.

Thus a more general expression for the relationship of the treatment means to effects is given in Equations 6.5. These expressions allow for the possibility of INTERACTION between the two factors A and B. A statistical test involving the $\alpha\beta_{ij}$ that we develop later may conclude that there is no evidence of interaction.

$$\begin{aligned}
\mu_{11} &= \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + [\mu_{11} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_1)] \qquad (6.5) \\
\mu_{11} &= \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + \alpha\beta_{11} \\
\\
\mu_{12} &= \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + [\mu_{12} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_2)] \\
\mu_{12} &= \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + \alpha\beta_{12}
\end{aligned}$$

$$\begin{aligned}
\mu_{21} &= \mu_{..} + \alpha_2 + \beta_1 + [\mu_{21} - (\mu_{..} + \alpha_2 + \beta_1)] \\
\mu_{21} &= \mu_{..} + \alpha_2 + \beta_1 + \alpha\beta_{21}
\end{aligned}$$

$$\begin{aligned}
\mu_{22} &= \mu_{..} + \alpha_2 + \beta_2 + [\mu_{22} - (\mu_{..} + \alpha_2 + \beta_2)] \\
\mu_{22} &= \mu_{..} + \alpha_2 + \beta_2 + \alpha\beta_{22}
\end{aligned}$$

Thus the "full" model, means and effects, for each of the treatments is:

$$\begin{aligned}
y_{11k} &= \mu_{11} + \epsilon_{11k} \\
&= \mu_{..} + \alpha_1 + \beta_1 + \alpha\beta_{11} + \epsilon_{11k} \\[1em]
y_{12k} &= \mu_{12} + \epsilon_{12k} \\
&= \mu_{..} + \alpha_1 + \beta_2 + \alpha\beta_{12} + \epsilon_{12k} \\[1em]
y_{21k} &= \mu_{21} + \epsilon_{21k} \\
&= \mu_{..} + \alpha_2 + \beta_1 + \alpha\beta_{21} + \epsilon_{21k} \\[1em]
y_{22k} &= \mu_{22} + \epsilon_{22k} \\
&= \mu_{..} + \alpha_2 + \beta_2 + \alpha\beta_{22} + \epsilon_{22k}
\end{aligned} \tag{6.6}$$

In its most general form, the **model for the two-factor completely randomized design with interaction** is:

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \tag{6.7}$$

where in general $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n$.

In practice the terms on the right side of the equation are unknown and must be estimated based on the data to draw conclusions about these terms. An assumption of the model is that the $\epsilon_{ijk}{}'s$ are independent normal random variables each with mean 0 and unknown variance $\sigma^2$.

## 6.4 Data Decomposition

When a two factor experiment is conducted the resulting data can be decomposed using the general interaction model (6.7) discussed in the last section. Based on this decomposition an analysis of variable table similar to that in Chapter 4 can be formed.

Suppose the tomato experiment was carried out with the results in Table 6.4 being tomato production in pounds for four replications per treatment combination.

Table 6.4: Tomato Production with Means

| | | $B_1$ | | $B_2$ | | |
|---|---|---|---|---|---|---|
| $A_1$ | $\overline{y}_{11\cdot} = 9.25$ | 8 | $\overline{y}_{12\cdot} = 11.75$ | 11 | $\overline{y}_{1\cdot\cdot} = 10.5$ | $\hat{\alpha}_1 = 1.81$ |
| | | 8 | | 11 | | |
| | | 9 | | 12 | | |
| | | 12 | | 13 | | |
| $A_2$ | $\overline{y}_{21\cdot} = 5.75$ | 5 | $\overline{y}_{22\cdot} = 8.00$ | 7 | $\overline{y}_{2\cdot\cdot} = 6.88$ | $\hat{\alpha}_2 = -1.81$ |
| | | 6 | | 8 | | |
| | | 6 | | 8 | | |
| | | 6 | | 9 | | |
| | | $\overline{y}_{\cdot1\cdot} = 7.50$<br>$\hat{\beta}_1 = -1.19$ | | $\overline{y}_{\cdot2\cdot} = 9.88$<br>$\hat{\beta}_2 = 1.19$ | $\overline{y}_{\cdots} = 8.69$ | |

Note that $\overline{y}_{11\cdot} = 9.25$, the average of the four treatment $A_1B_1$ observations, is an estimate of the true treatment mean $\mu_{11}$. Similarly, $\overline{y}_{12\cdot} = 11.75$, $\overline{y}_{21\cdot} = 5.75$, and $\overline{y}_{22\cdot} = 8.00$ are estimates of the true treatment means $\mu_{12}, \mu_{21}, \mu_{22}$, respectively. The sample marginal means $\overline{y}_{1\cdot\cdot} = 10.5, \overline{y}_{2\cdot\cdot} = 6.88, \overline{y}_{\cdot1\cdot} = 7.50$, and $\overline{y}_{\cdot2\cdot} = 9.88$ are sample estimates of true marginal means $\mu_{1\cdot}, \mu_{2\cdot}, \mu_{\cdot1}, \mu_{\cdot2}$, respectively. The sample main effects $\hat{\alpha}_1 = 1.81$ and $\hat{\alpha}_2 = -1.81$ are estimates of the true main effects $\alpha_1$ and $\alpha_2$. The sample main effects $\hat{\beta}_1 = -1.19$ and $\hat{\beta}_2 = 1.19$ are estimates of the true main effects $\beta_1$ and $\beta_2$. Finally the sample grand mean $\overline{y}_{\cdots} = 8.69$ is an estimate of the true grand mean $\mu_{\cdot\cdot}$.

We can write each observed tomato yield $y$ in terms of the estimated parameters. As an example,

$$\begin{array}{lllll}
y_{111} & = & 8 & = & \overline{y}_{11\cdot} + e_{111} \\
 & & & = & 9.25 + (8 - 9.25) \\
 & & & = & 9.25 + (-1.25) \\
 & & & = & 8.69 + 1.81 + (-1.19) + [9.25 - (8.69 + 1.81 - 1.19)] + (-1.25) \\
 & & & = & 8.69 + 1.81 + (-1.19) + (-0.06) + (-1.25) \\
 & & & = & \overline{y}_{\cdots} + \hat{\alpha}_1 + \hat{\beta}_1 + \widehat{\alpha\beta}_{11} + e_{111}
\end{array}$$

$$\begin{array}{lllll}
y_{123} & = & 12 & = & \overline{y}_{12\cdot} + e_{123} \\
 & & & = & 11.75 + (12 - 11.75) \\
 & & & = & 11.75 + 0.25 \\
 & & & = & 8.69 + 1.81 + (1.19) + [11.75 - (8.69 + 1.81 + 1.19)] + (0.25) \\
 & & & = & 8.69 + 1.81 + 1.19 + 0.06 + 0.25 \\
 & & & = & \overline{y}_{\cdots} + \hat{\alpha}_1 + \hat{\beta}_2 + \widehat{\alpha\beta}_{12} + e_{123}
\end{array}$$

The complete decomposition is given in Table 6.5. The interaction effect of $-0.07$ is in theory the same as the other interaction effects in magnitude but differs because of rounding.

Table 6.5: Decomposition for Two Factor Model

| $y_{ijk}$ | $=$ | $\overline{y}_{...}$ | $+$ | $\hat{\alpha}_i$ | $+$ | $\hat{\beta}_j$ | $+$ | $\widehat{\alpha\beta}_{ij}$ | $+$ | $e_{ijk}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | = | 8.69 | + | 1.81 | + | (-1.19) | + | (-0.06) | + | (-1.25) |
| 8 | = | 8.69 | + | 1.81 | + | (-1.19) | + | (-0.06) | + | (-1.25) |
| 9 | = | 8.69 | + | 1.81 | + | (-1.19) | + | (-0.06) | + | (-0.25) |
| 12 | = | 8.69 | + | 1.81 | + | (-1.19) | + | (-0.06) | + | 2.75 |
| 11 | = | 8.69 | + | 1.81 | + | 1.19 | + | 0.06 | + | (-0.75) |
| 11 | = | 8.69 | + | 1.81 | + | 1.19 | + | 0.06 | + | (-0.75) |
| 12 | = | 8.69 | + | 1.81 | + | 1.19 | + | 0.06 | + | ( 0.25) |
| 13 | = | 8.69 | + | 1.81 | + | 1.19 | + | 0.06 | + | (1.25) |
| 5 | = | 8.69 | + | (-1.81) | + | (-1.19) | + | 0.06 | + | (-0.75) |
| 6 | = | 8.69 | + | (-1.81) | + | (-1.19) | + | 0.06 | + | 0.75 |
| 6 | = | 8.69 | + | (-1.81) | + | (-1.19) | + | 0.06 | + | 0.25 |
| 6 | = | 8.69 | + | (-1.81) | + | (-1.19) | + | 0.06 | + | 0.25 |
| 7 | = | 8.69 | + | (-1.81) | + | (1.19) | + | -0.07 | + | (-1.00) |
| 8 | = | 8.69 | + | (-1.81) | + | (1.19) | + | -0.07 | + | 0.00 |
| 8 | = | 8.69 | + | (-1.81) | + | (1.19) | + | -0.07 | + | 0.00 |
| 9 | = | 8.69 | + | (-1.81) | + | (1.19) | + | -0.07 | + | 1.00 |

Table 6.6: Sums of Squares for Two Factor Example

| | | | | |
|---|---|---|---|---|
| SSTOT | = | $8^2 + 8^2 + \ldots + 9^2$ | = | 1299 |
| SSGM | = | $16(8.69)^2$ | = | 1208.26 |
| SSA | = | $8(1.81)^2 + 8(-1.81)^2$ | = | 52.42 |
| SSB | = | $8(1.19)^2 + 8(-1.19)^2$ | = | 22.66 |
| SSAB | = | $4(-.06)^2 + 4(.06)^2 + 4(0.06)^2 + 4(-0.07)^2$ | = | 0.0471 |
| SSE | = | $(-1.25)^2 + \ldots + (1.00)^2$ | = | 16.25 |

In order to develop hypothesis tests to test for factor A, factor B, and interaction effects, similar to Chapter 4, we will now calculate sums of squared effects across the different observations. These sums of squared effects for the tomato example are given in Table 6.6.

It can be shown that in general

$$SSTOT = SSGM + SSA + SSB + SSAB + SSE$$

In this example because of rounding we have approximate equality:

$$1299 \simeq 1208.26 + 52.42 + 22.66 + 0.0471 + 16.25 = 1299.64$$

The degrees of freedom associated with the different sums of squares are equal to the following:

| SS | Degrees of Freedom | Degrees of Freedom - Tomato Example |
|---|---|---|
| SSTOT | N | 16 |
| SSGM | 1 | 1 |
| SSA | (a-1) | 1 |
| SSB | (b-1) | 1 |
| SSAB | (a-1)(b-1) | 1 |
| SSE | N - ab | 12 |

Note that the degrees of freedom are additive in that degrees of freedom for SSGM, SSA, SSB, SSAB, and SSE add to degrees of freedom for SSTOT.

$$N = 1 + (a-1) + (b-1) + (a-1)(b-1) + (N - ab)$$

or in this example,

$$16 = 1 + 1 + 1 + 1 + 12$$

Typically in computer calculations the grand mean is subtracted from each value of the response $y$ and this difference or deviation from the mean appears on the left side of the decomposition. Then the relevant total sum of squares is the "corrected" total sum of squares, which is the summing of the squares of the deviations. The corrected total sum of squares would then equal

Table 6.7: ANOVA Table for Two Factor Completely Randomized Design

| Source of Variation | df | SS | MS | F | E[MS] |
|---|---|---|---|---|---|
| A | a - 1 | SSA | MSA | MSA/MSE | E[MSA] |
| B | b - 1 | SSB | MSB | MSB/MSE | E[MSB] |
| A*B | (a-1)(b-1) | SSAB | MSAB | MSAB/MSE | E[MSAB] |
| *Error* | ab(n-1) | SSE | MSE | | E[MSE] |

$$SSTOT_C = SST - SSGM$$

Degrees of freedom associated with $SSTOT_C = N - 1$. In this example, $SSTOT_C = 1299 - 1208.26 = 90.74$ and *df* $= N - 1 = 16 - 1 = 15$. When correcting for the grand mean the sums of squares decomposition is

$$SSTOT_C = SSA + SSB + SSAB + SSE$$

Mean squares are defined as in Chapter 4 by dividing sums of squares for effects by their corresponding degrees of freedom. Thus for the tomato example, MSA = 52.42/1 = 52.42, MSB = 22.66/1 = 22.66, MSAB = 0.0471/1 = 0.0471, and MSE = 16.25/12 = 1.35.

## 6.5 F ratios and Hypothesis Testing

In this section we consider three hypothesis tests that can be conducted in a two factor study: a test for interaction between factors A and B, a test for A main effects, and a test for B main effects. The logic proceeds as in Chapter 4. For example, if there are truly no effects of a factor then the size of the mean square for that effect (such as MSA,MSB, or MSAB) based on the data should be roughly the same magnitude as the size of mean squared error (MSE). If there truly are effects of a factor then the mean square of that effect should be larger than mean squared error.

The general form of the ANOVA table for a two factor completely randomized design with expected mean squares, EMS, is given in Table **??**

Formulas for the sums of squares (SS) and mean squares (MS) in Table 6.7 are provided in Section 6.5.1. Expected values of the various mean squares, E[MS] in Table 6.7 can be shown to be

$$\begin{aligned} E[MSE] &= \sigma^2 && (6.8) \\ E[MSA] &= \sigma^2 + \frac{nb\sum_{i=1}^{a}\alpha_i^2}{a-1} && (6.9) \\ E[MSB] &= \sigma^2 + \frac{na\sum_{j=1}^{b}\beta_j^2}{b-1} && (6.10) \end{aligned}$$

$$E[MSAB] \quad = \quad \sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2}{(a-1)(b-1)} \tag{6.11}$$

Thus if there are no main effects for A, that is all $\alpha_i$ are 0, then $E[MSA]$ and $E[MSE]$ are both equal to $\sigma^2$, and we would expect the observed values of $MSA$ and $MSE$ to be about the same. If there are main effects of A, that is not all of the $\alpha_i$ are 0, then $E[MSA] > E[MSE]$ and we would expected the observed value of MSA to be larger than the observed value of MSE. Comparisons like this form the basis for hypothesis testing in the two factor completely randomized design. We first consider the hypothesis test for interaction between A and B since the significance or lack thereof affects the interpretation of the test for A and B main effects.

### 6.5.1 F test for AB interaction

The null and alternative hypotheses for the test of interaction between factors A and B in general form are

$$H_o : \alpha\beta_{ij} = 0 \text{ for each pair } i, j$$

and

$$H_a : \alpha\beta_{ij} \neq 0 \text{ for some pair } i, j$$

The test statistic is

$$F = \frac{MSAB}{MSE} = \frac{SSAB/(a-1)(b-1)}{SSE/(N-ab)}$$

where

$$SSAB = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\widehat{\alpha\beta})_{ij}^2 = n\sum_{i=1}^{a}\sum_{j=1}^{b}[\overline{y}_{ij.} - (\overline{y}_{...} + \hat{\alpha}_i + \hat{\beta}_j)]^2$$

and

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[e_{ijk}]^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[y_{ijk} - \overline{y}_{ij.}]^2$$

The $F$ statistic measures variation in the treatment means from what is expected under the assumption of no interaction, relative to the variation within groups.

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common standard deviation holds, the $F$ ratio above has the “F” distribution with $\nu_1 = (a-1)(b-1)$ numerator degrees of freedom and $\nu_2 = (N-ab)$ denominator degrees of freedom.

At a significance level of $\alpha$ the null hypothesis would be rejected if the observed value of the test statistic, $F_o$, is larger than $F_{\alpha;(a-1)(b-1),N-ab}$, the upper $\alpha$ probability point from the appropriate F distribution.

We will usually use a statistical package to obtain a P-value and use that to make the decision. The null hypothesis is then rejected if the P-value $\leq \alpha$, where P-value $= P[F \geq F_o]$.

### 6.5.2 F test for A main effects

The null and alternative hypotheses for the test of A main effects are

$$H_o : \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$$

or equivalently in terms of A main effect or marginal means,

$$H_0 : \mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{a\cdot}$$

The alternative hypothesis is

$$H_a : \text{ not all } \alpha_i's = 0$$

or equivalently,

$$H_a : \text{ not all } \mu_{i\cdot}'s \text{ are equal}$$

The test statistic is

$$F = \frac{MSA}{MSE} = \frac{SSA/(a-1)}{SSE/(N-ab)}$$

where $SSA = nb\sum_{i=1}^{a}\hat{\alpha}_i^2 = nb\sum_{i=1}^{a}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot\cdot\cdot})^2$ and SSE is as in the test for interaction. Note that $F$ measures variation in the Factor A level marginal means (between group variation) relative to the variation within groups. That is the test statistic is comparing sample marginal means for factor A.

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common standard deviation holds, the $F$ ratio above has the"F" probability distribution with $\nu_1 = (a-1)$ numerator degrees of freedom and $\nu_2 = (N-ab)$ denominator degrees of freedom.

At a significance level of $\alpha$ the null hypothesis would be rejected if the observed value of the test statistic, $F_o$, is larger than $F_{\alpha;(a-1),N-ab}$, the upper $\alpha$ probability point from the appropriate F distribution or equivalently if P-value $\leq \alpha$, where P-value $= P[F \geq F_o]$.

### 6.5.3 F test for B main effects

The null and alternative hypotheses for the test of B main effects are

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_b = 0$$

or equivalently in terms of B main effect or marginal means,

$$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot b}$$

The alternative hypothesis is

$$H_a : \text{ not all } \beta_j's = 0$$

or equivalently,

$$H_a: \text{ not all } \mu'_{.j}s \text{ are equal}$$

The test statistic is

$$F = \frac{MSB}{MSE} = \frac{SSB/(b-1)}{SSE/(N-ab)}$$

where $SSB = na\sum_{j=1}^{b}\hat{\beta}_j^2 = na\sum_{j=1}^{b}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\ldots})^2$ and SSE is as in the test for interaction. Note that $F$ measures variation in the Factor B level marginal means (between group variation) relative to the variation within groups.

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common standard deviation holds, the $F$ ratio above the "F" probability distribution with $\nu_1 = (b-1)$ numerator degrees of freedom and $\nu_2 = (N - ab)$ denominator degrees of freedom.

At a significance level of $\alpha$ the null hypothesis would be rejected if the observed value of the test statistic, $F_o$, is larger than $F_{\alpha;(b-1),N-ab}$, the upper $\alpha$ probability point from the appropriate F distribution or equivalently if P-value $\leq \alpha$, where P-value $= P[F \geq F_o]$.

### 6.5.4 Testing Strategy

Typically the F test for interaction is conducted first. If the F test for interaction is not significant at some prescribed $\alpha$ level then the F test for each of factor A and B main effect (marginal) means $\mu_{i\cdot}$ and $\mu_{\cdot j}$ is conducted at prescribed $\alpha$ levels. If the F test for factor A (or B) main effects is significant then a multiple comparison procedure might be used to determine which of the main effect (marginal) means are different.

If the F test for interaction is significant and the interactions are deemed to be important then the conclusion is that differences in main effect (marginal) means for levels of one factor are not representative of differences in those levels across all levels of the other factor. Comparisons of treatment combination means, $\mu_{ij}$, rather than main effect means are more appropriate. For example treatment combination means involving levels of A are compared at each level of factor B. Or treatment combination means involving levels of B are compared at each level of A. Examples are provided in subsequent sections.

Some practitioners use a liberal significance level for the F test for interaction, such as 0.10 or 0.15, instead of the usual 0.05 level. This increase in the Type I error rate decreases the Type II error rate. The philosophy is that the Type II error rate is more serious. The Type II error would be concluding no interaction when there is interaction. A conclusion of no interaction would then result in comparison of main effect (marginal) means for levels of a factor when these comparisons are not representative of comparisons at the different levels of the other factor. The Type I error would perhaps not be regarded as serious. This would mean concluding interaction and thus comparing treatment combination means when in fact there is no interaction and one could have simplified results by comparing main effect marginal means. The 0.10 level will normally be used for testing interaction in this text unless otherwise stated.

Table 6.8: ANOVA Table for Tomato Example

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Fert | 1 | 52.56 | 52.56 | 38.82 | <.0001 |
| Water | 1 | 22.56 | 22.56 | 16.66 | 0.0015 |
| Fert*Water | 1 | 0.06 | 0.06 | 0.05 | 0.8335 |
| Error | 12 | 16.25 | 1.35 | | |
| Total (Corrected) | 15 | 91.44 | | | |

## 6.6 Examples

### 6.6.1 Tomato Weight Example

Table 6.8 gives an ANOVA table for the tomato example based on computer software. Note that the numbers in this table differ slightly from those of Table 6.6 because of rounding used for that table. The interaction effect is not significant at the $\alpha = 0.10$ level with ($F = 0.05$, P-value $= 0.8335$), providing no evidence of the effects of fertilizer depending upon water (or the effects of water depending upon fertilizer). There is evidence at $\alpha = 0.05$ that both fertilizer ($F = 38.82$, P-value $< 0.0001$) and water ($F = 16.66$, P-value $= 0.0015$) affect tomato production.

### 6.6.2 Paper Towel Example - No Interaction

Kim Cartwright (Spring 2001) conducted an experiment to compare the amounts of three liquids absorbed by three brands of paper towels. The three liquids (Factor A) were

- Water
- Dishwashing Detergent
- Vegetable oil

The three brands of paper towels were

- Coronet
- Kleenex
- Scott

Each liquid was tested with each brand three times for a total of $N = 27$ observations on amount of liquid absorbed. The testing was conducted as follows:

Table 6.9: Paper Towel Example: Amount of Liquid Absorbed (mL)

| | | Liquid | |
|---|---|---|---|
| | Water | Dishwashing Detergent | Vegetable Oil |
| Paper Towel | | | |
| Coronet | 26 | 19 | 22 |
| | 22 | 16 | 25 |
| | 22 | 15 | 29 |
| Kleenex | 43 | 33 | 39 |
| | 41 | 38 | 41 |
| | 41 | 38 | 45 |
| Scott | 27 | 21 | 27 |
| | 26 | 20 | 25 |
| | 25 | 21 | 25 |

Fifty milliliters of each liquid was poured/measured into a graduated cylinder and then poured into a container. The paper towel was then submerged in the container. After 1 minute had passed, the paper towel was removed, letting the excess liquid drip off the towel for 30 seconds. The remaining liquid in the container was then poured back into the graduated cylinder. This remaining amount was then subtracted from 50 to get the amount of liquid absorbed. This was done 27 times, at each time randomly choosing a liquid and brand to use. The amount of liquid absorbed (mL) for the various liquid and brand combinations is given in the Table 6.9

Figure 6.3 is a plot of the amounts of liquid absorbed versus the treatment combination of brand of towel and liquid. A few observations can be made based on the plot. Kleenex appears to have been most absorbent regardless of type of liquid used. The comparison of liquid types is similar across the brands, with the amount of detergent absorbed being less than similar amounts of water and oil. Thus there does not appear to be any evidence of interaction between brand and liquid used.

Mean absorption for the nine treatments is given in Table 6.10 and an interaction plot is given in Figure 6.4.

An interaction plot with Liquid on the horizontal axis and lines for the three brand of paper towel is given in Figure 6.4. Note that the lines are approximately parallel, indicating that the difference in amount absorbed by two paper towel brands is about the same regardless of the liquid.

The model for the data is given by

$$y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \qquad (6.12)$$

with

$i = 1, 2, 3$ representing the $i^{th}$ level (Coronet, Kleenex, Scott) of Paper Towel

Figure 6.3: Amount of Liquid Absorbed versus Towel/Liquid

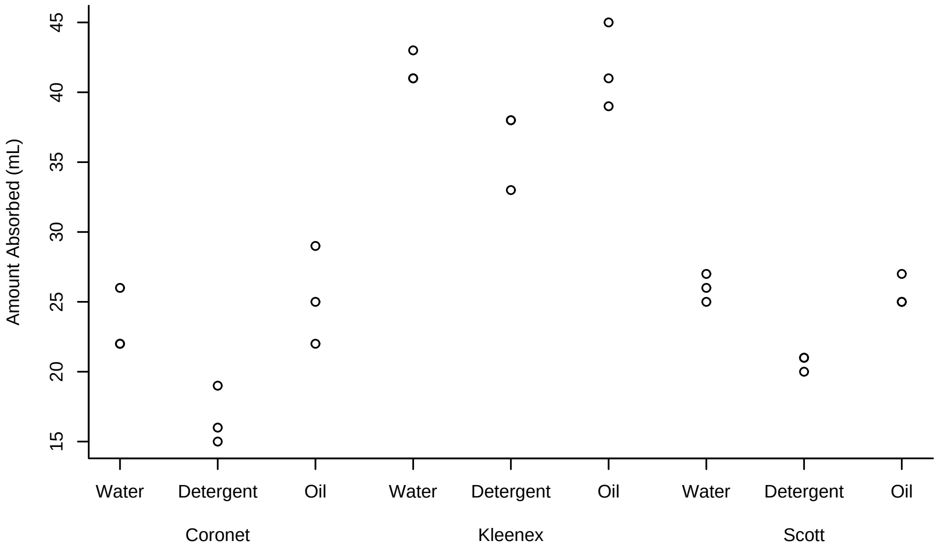
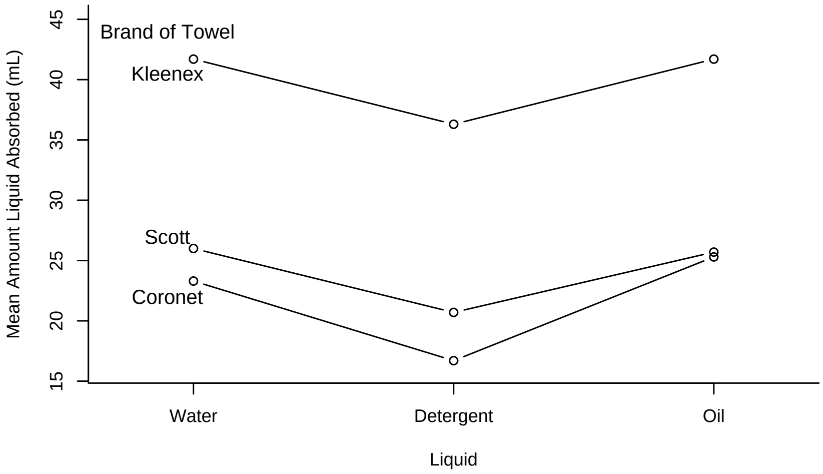


Table 6.10: Means of Amount Absorbed (mL): Paper Towel Example

| | Liquid | | | |
|---|---|---|---|---|
| | Water | Dishwashing Detergent | Vegetable Oil | Marginal Mean |
| Paper Towel | | | | |
| Coronet | 23.3 | 16.7 | 25.3 | 21.8 |
| Kleenex | 41.7 | 36.3 | 41.7 | 39.9 |
| Scott | 26.0 | 20.7 | 25.7 | 24.1 |
| Marginal Mean | 30.3 | 24.6 | 30.9 | |
| | | | | $\overline{y}_{...} = 28.6$ |

Figure 6.4: Mean Amount of Liquid Absorbed versus Liquid by Brand



$j = 1, 2, 3$ representing the $j^{th}$ level (water, detergent, oil) of Liquid

$k = 1, 2, 3$ is an index on a particular amount of liquid absorbed for the $i^{th}$ paper towel and $j^{th}$ liquid

$y_{ijk}$ represents the $k^{th}$ observation on amount absorbed for towel $i$ and liquid $j$

$\mu_{..}$ = the true grand mean of amount absorbed

$\alpha_i$ = the true main effect of the $i^{th}$ level of paper towel on amount absorbed

$\beta_j$ = the true main effect of the $j^{th}$ level of liquid on amount absorbed

$\alpha\beta_{ij}$ = the true interaction effect of the $i^{th}$ level of paper towel and $j^{th}$ level of liquid on amount absorbed

$\epsilon_{ijk}$ = the effects of extraneous variables on the $k^{th}$ amount at the $i^{th}$ paper towel and $j^{th}$ liquid

It is assumed that the 27 errors, $\epsilon_{ijk}$, are values of independent normal random variables, each with mean of 0 and variance $\sigma^2$.

The ANOVA table for the Paper Towel example is given in Table 6.11. There is no evidence of interaction between Towel Brand and Liquid at the 0.10 level of significance with $F = 0.65$, P-value = 0.6350. There is evidence at the 0.05 level of both brand effects ($F = 180.05$, P-value $< .0001$) and Liquid effects ($F = 22.82$, P-value $< .0001$).

Table 6.11: ANOVA Table for Paper Towel Example

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Towel | 2 | 1747.19 | 873.59 | 180.05 | <.0001 |
| Liquid | 2 | 221.41 | 110.70 | 22.82 | <.0001 |
| Towel*Liquid | 4 | 12.59 | 3.14 | 0.65 | 0.6350 |
| Error | 18 | 87.33 | 4.85 | | |
| Total (Corrected) | 26 | 2068.5 | | | |

Since there is no evidence of interaction between brand and liquid, marginal means of amount absorbed will be compared among the three brands using Tukey-Kramer simultaneous confidence intervals. The marginal means of amount absorbed (mL) for the Coronet, Kleenex, and Scott brands, are respectively, $\overline{y}_{1..} = 21.8$, $\overline{y}_{2..} = 39.9$, and $\overline{y}_{3..} = 24.1$

For two levels $i$ and $i'$ of Towel Brand, the general form of the interval for $\mu_{i\cdot} - \mu_{i'\cdot}$ and simultaneous confidence level of 95% is

$$\overline{y}_{i..} - \overline{y}_{i'..} \pm \frac{q_{0.05;\nu,a}}{\sqrt{2}} \sqrt{MSE} \sqrt{\frac{1}{bn} + \frac{1}{bn}}$$

where $\overline{y}_{i..}$ and $\overline{y}_{i'..}$ refer, respectively, to the marginal means of amount absorbed for levels $i$ and $i'$ of brand of towel. The denominator $bn = (3)(3)$ in the denominators refer to the number of observations used to calculate the marginal means. The value of MSE is 4.85 with $\nu = 18$ degrees of freedom. From Table A.6 with $\nu = 18$ and $t = a = 3$ the upper 0.05 probability point $q_{0.05;18,3}$ is 3.61.

Thus the endpoints of the simultaneous 95% Tukey-Kramer confidence intervals are

$$\begin{array}{rcl} 21.8 - 39.9 & \pm & \frac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\frac{1}{9}+\frac{1}{9}} \\ 21.8 - 24.1 & \pm & \frac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\frac{1}{9}+\frac{1}{9}} \\ 39.9 - 24.1 & \pm & \frac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\frac{1}{9}+\frac{1}{9}} \end{array}$$

or

$$\begin{array}{rcl} -18.1 & \pm & 2.7 \\ -2.3 & \pm & 2.7 \\ 15.8 & \pm & 2.7 \end{array}$$

Thus the three intervals are:

$$\begin{array}{rcccl} -20.8 & \leq & \mu_{1\cdot} - \mu_{2\cdot} & \leq & -15.4 \\ -5.0 & \leq & \mu_{1\cdot} - \mu_{3\cdot} & \leq & 0.4 \\ 13.1 & \leq & \mu_{2\cdot} - \mu_{3\cdot} & \leq & 18.5 \end{array}$$

The Kleenex brand results in higher absorption than either of the other two brands. The mean absorption for Kleenex is estimated to be between 15.4 and 20.8 milliliters higher than that for Coronet and between 13.1 and 18.5 milliliters higher than that for Scott. There is no evidence of a difference in mean absorption between the Coronet and Scott brands. These conclusions are based on an experimentwise confidence level of 95%.

Since the F test for overall differences in liquids is significant, a set of simultaneous 95% Tukey-Kramer confidence intervals will be used to compare the three liquids. The marginal means of amount absorbed (mL) for Water, Detergent, and Oil, are, respectively, $\overline{y}_{\cdot 1 \cdot} = 30.3$, $\overline{y}_{\cdot 2 \cdot} = 24.6$, and $\overline{y}_{\cdot 3 \cdot} = 30.9$

The general form of the interval for $\mu_{\cdot j} - \mu_{\cdot j'}$ is

$$\overline{y}_{\cdot j \cdot} - \overline{y}_{\cdot j' \cdot} \pm \frac{q_{0.05;\nu,b}}{\sqrt{2}} \sqrt{MSE} \sqrt{\frac{1}{an} + \frac{1}{an}}$$

where $\overline{y}_{\cdot j \cdot}$ and $\overline{y}_{\cdot j' \cdot}$ refer, respectively, to the marginal means of amount absorbed for levels $j$ and $j'$ of the factor liquid. The value $an = (3)(3)$ in the denominators refer to the number of observations used to calculate the marginal means. The value of MSE is 4.85 with $\nu = 18$ degrees of freedom. From Table A.6 with $\nu = 18$ and $t = b = 3$ the upper 0.05 probability point $q_{0.05;18,3}$ is 3.61.

Thus the endpoints of the simultaneous 95% Tukey-Kramer confidence intervals are

$$\begin{array}{rcl}
30.3 - 24.6 & \pm & \frac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\frac{1}{9} + \frac{1}{9}} \\
30.3 - 30.9 & \pm & \frac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\frac{1}{9} + \frac{1}{9}} \\
24.6 - 30.9 & \pm & \frac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\frac{1}{9} + \frac{1}{9}}
\end{array}$$

or

$$\begin{array}{rcl}
-18.1 & \pm & 2.7 \\
-2.3 & \pm & 2.7 \\
15.8 & \pm & 2.7
\end{array}$$

Thus the three intervals are:

$$\begin{array}{rcccl}
3.0 & \leq & \mu_{\cdot 1} - \mu_{\cdot 2} & \leq & 8.4 \\
-3.3 & \leq & \mu_{\cdot 1} - \mu_{\cdot 3} & \leq & 2.1 \\
-9.0 & \leq & \mu_{\cdot 2} - \mu_{\cdot 3} & \leq & -3.6
\end{array}$$

On average less detergent was absorbed than either water or oil. It is estimated that the mean amount of detergent absorbed is between 3.0 and 8.4 milliliters less than that of water and between 3.6 and 9.0 milliliters less than than of oil. There was no significant difference between the mean amounts of water and oil absorbed. These conclusions are made with an experimentwise confidence level of 95%.

Table 6.12: Yield Data

| | | | Variety | | |
|---|---|---|---|---|---|
| | V1 | V2 | V3 | V4 | V5 |
| Method | | | | | |
| A | 22.1 | 27.1 | 22.3 | 19.8 | 20.0 |
| | 24.1 | 15.1 | 25.8 | 28.3 | 17.0 |
| | 19.1 | 20.6 | 22.8 | 26.8 | 24.0 |
| | 22.1 | 28.6 | 28.3 | 27.3 | 22.5 |
| | 25.1 | 15.1 | 21.3 | 26.8 | 28.0 |
| | 18.1 | 24.6 | 18.3 | 26.8 | 22.5 |
| B | 13.5 | 16.9 | 15.7 | 15.1 | 21.8 |
| | 14.5 | 17.4 | 10.2 | 6.5 | 22.8 |
| | 11.5 | 10.4 | 16.7 | 17.1 | 18.8 |
| | 6.0 | 19.4 | 19.7 | 7.6 | 21.3 |
| | 27.0 | 11.9 | 18.2 | 13.6 | 16.3 |
| | 18.0 | 15.4 | 12.2 | 21.1 | 14.3 |
| C | 19.0 | 20.0 | 16.4 | 24.5 | 11.8 |
| | 22.0 | 22.0 | 14.4 | 16.0 | 14.3 |
| | 20.0 | 25.5 | 21.4 | 11.0 | 21.3 |
| | 14.5 | 16.5 | 19.9 | 7.5 | 6.3 |
| | 19.0 | 18.0 | 10.4 | 14.5 | 7.8 |
| | 16.0 | 17.5 | 21.4 | 15.5 | 13.8 |

### 6.6.3 Example with Interaction

This example is taken from Littel, Stroup, and Freund [17]. An experiment was conducted to compare three seed growth-promoting methods (A,B,C) for five different varieties of turf grass (V1,V2,V3,V4,V5). Seeds from each variety and method combination were planted in 6 pots. The resulting 90 pots were placed in a growth chamber and after four weeks the dry matter was measured for each pot. The resulting yields are given in Table 6.12.

A plot of the yields versus treatment combinations is given in Figure 6.5.

Note that seed growth-promoting method A appears to be the best regardless of the variety. The comparison of methods B and C seems to depend upon the variety. Also variability in yields appear not to depend much on treatment combination.

Mean yield for the nine treatment combinations of method and variety along with marginal means corresponding to levels of each factor are given in Table 6.13. Note that the marginal mean yields and treatment mean yields for method A are consistently higher than the corresponding values for Methods B and C. The marginal and treatment mean yields for method B are all lower than that for method C except for variety V5, indicating possible interaction.

An interaction plot is given in Figure 6.6

Figure 6.5: Plot of Yield versus Method/Variety

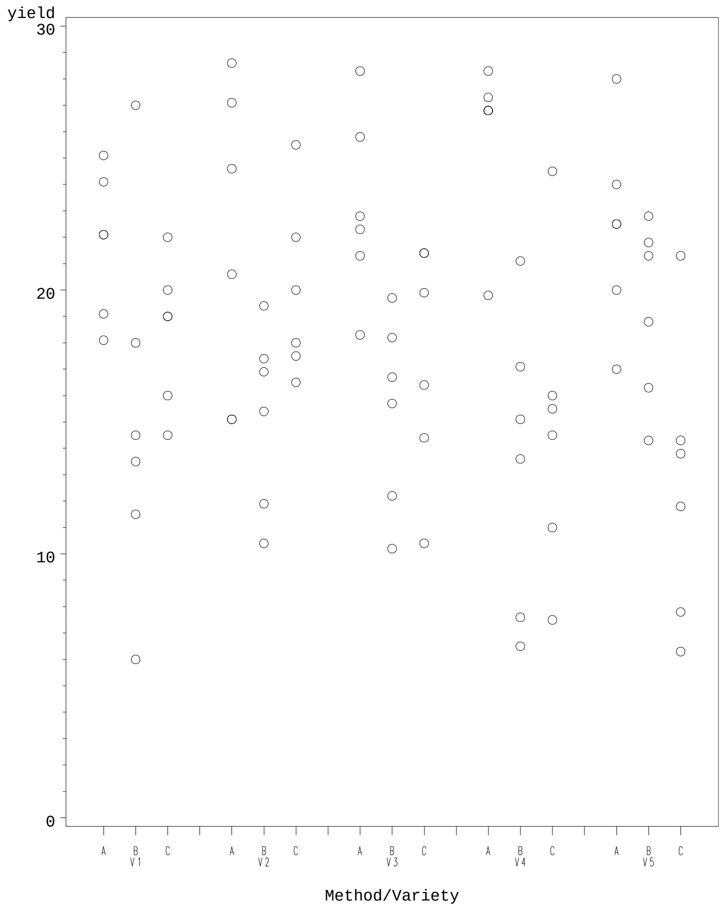
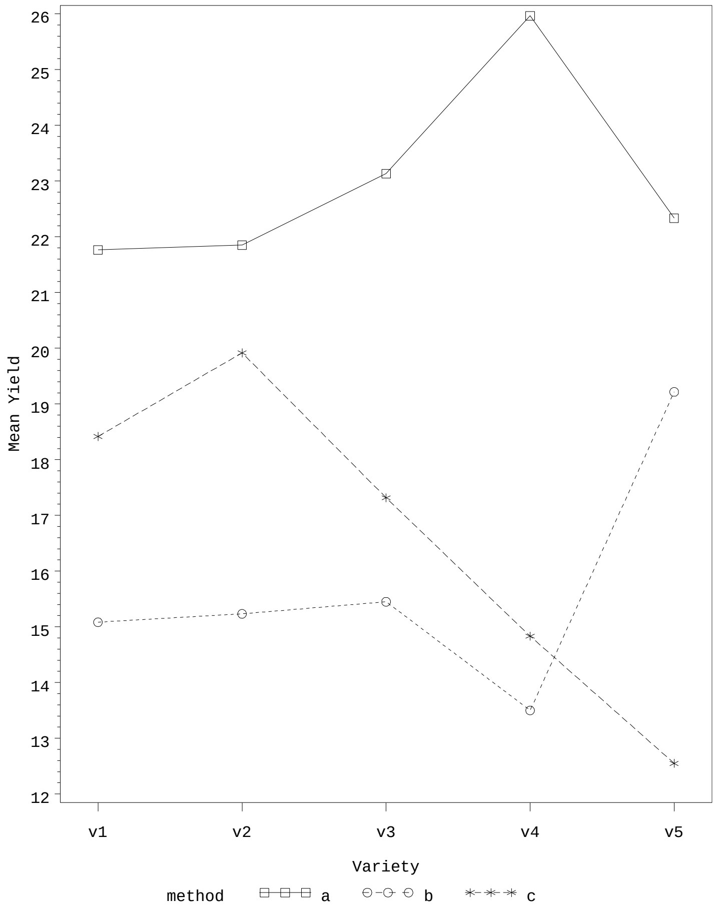


Table 6.13: Yield Means: Grasses Example

| | Variety | | | | | |
|---|---|---|---|---|---|---|
| | V1 | V2 | V3 | V4 | V5 | Marginal Mean |
| Method | | | | | | |
| A | 21.8 | 21.8 | 23.1 | 26.0 | 22.3 | 23.0 |
| B | 15.1 | 15.2 | 15.4 | 13.5 | 19.2 | 15.7 |
| C | 18.4 | 19.9 | 17.3 | 14.8 | 12.6 | 16.6 |
| Marginal Mean | 18.4 | 19.0 | 18.6 | 18.1 | 18.0 | |
| | | | | | | $\overline{y}_{\ldots} = 18.4$ |

Figure 6.6: Interaction Plot Grass Data



The interaction plot more clearly shows evidence of interaction between method and variety as noted in Figure 6.5.

The population (effects) model for the data is given by

$$y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \tag{6.13}$$

where

$i = 1(A), 2(B), 3(C)$ indexes the seed growth promoting method

$j = 1(V1), 2(V2), 3(V3), 4(V4), 5(V5)$ indexes the variety of turf grass

$k = 1, 2, 3, 4, 5, 6$ indexes the yield of dry matter for a particular combination $(i, j)$

$y_{ijk}$ represents the $k^{th}$ observation on yield of dry matter for method $i$ and variety $j$

$\mu_{\cdot\cdot}$ = the true grand mean of yield of dry matter

$\alpha_i$ = the true main effect of the $i^{th}$ method on yield

$\beta_j$ = the true main effect of the $j^{th}$ variety on yield

$\alpha\beta_{ij}$ = the true interaction effect of the $i^{th}$ level of method and $j^{th}$ variety level on yield

$\epsilon_{ijk}$ = the effects of extraneous variables on the $k^{th}$ yield at the $i^{th}$ method and $j^{th}$ variety, such as variations in seeds, pot characteristics, etc.

It is assumed that the 90 errors, $\epsilon_{ijk}$, are values of independent normal random variables, each with mean of 0 and variance $\sigma^2$.

The ANOVA table for the Grasses example is given in Table 6.14. There is evidence of interaction at the 0.10 level of significance ($F$ = 2.38, P-value = 0.0241) consistent with the interaction plot.

When there is interaction comparison of marginal means of levels of a factor may be misleading since this would imply that the comparison of those levels is similar across all levels of the other factor. The appropriate follow-up is a comparison of treatment means rather than marginal means. Treatment means for the 3 methods could be compared for each variety or treatment means for the 5 varieties could be compared for each method. The former comparison will be carried out here using the Tukey-Kramer method. Simultaneous 95% Tukey-Kramer confidence intervals will be calculated for differences in population treatment means $\mu_{ij}$ of the 3 methods at each of the 5 levels of variety.

The Tukey-Kramer confidence intervals for differences in population treatment mean yields for methods A (i = 1), B (i = 2), C (i = 3) when variety is $V1(j = 1)$ , with overall confidence level 0.95 are

Table 6.14: ANOVA Table for Grasses Example

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Method | 2 | 953.16 | 476.58 | 24.25 | <.0001 |
| Variety | 4 | 11.38 | 2.85 | 0.14 | <.0001 |
| Method*Variety | 8 | 374.49 | 46.81 | 2.38 | 0.0241 |
| Error | 75 | 1473.77 | 19.65 | | |
| Total (Corrected) | 89 | 2812.79 | | | |

$$\begin{array}{rcl}
\overline{y}_{11\cdot} - \overline{y}_{21\cdot} & \pm & \frac{q_{0.05;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{1}{n}} \\
\overline{y}_{11\cdot} - \overline{y}_{31\cdot} & \pm & \frac{q_{0.05;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{1}{n}} \\
\overline{y}_{21\cdot} - \overline{y}_{31\cdot} & \pm & \frac{q_{0.05;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{1}{n}}
\end{array}$$

where $q_{0.05;\nu,t}$ is the upper 0.05 probability point from the Studentized range distribution. From the ANOVA table $\nu = 75$ is the degrees of freedom associated with $MSE = 19.65$. The value $n = 6$ is the number of observations contributing to a method mean at a particular variety. Thus the standard error of the difference between two method (sample) means is $\sqrt{\frac{2(19.65)}{6}} = 2.56$. Table A.6 does not have a value for $\nu = 75$ degrees of freedom associated with error; we will use the conservative value of $\nu = 50$. Thus with $t = 3$ levels for the method factor at a particular variety $V1$, Table A.6 gives $q_{0.05;50,3} = 3.42$ for the upper 0.05 probability point from the Studentized Range distribution. Thus the multiplier on the standard error is $\frac{3.42}{\sqrt{2}} = 2.42$. The margin of error for a difference in sample means is thus $(2.42)(2.56) = 6.20$. Thus, using means from Table 6.13, the endpoints of the intervals for the differences $\mu_{11} - \mu_{21}$, $\mu_{11} - \mu_{31}$, and $\mu_{21} - \mu_{31}$ are:

$$(21.8 - 15.1) \pm 6.20 \quad (21.8 - 18.4) \pm 6.20 \quad (15.1 - 18.4) \pm 6.20$$

The simultaneous 95% confidence intervals for the family of comparisons of the three methods for variety $V1$ are:

$$\begin{array}{rcccl}
0.5 & \leq & \mu_{11} - \mu_{21} & \leq & 12.9 \\
-2.8 & \leq & \mu_{11} - \mu_{31} & \leq & 9.6 \\
-9.5 & \leq & \mu_{21} - \mu_{31} & \leq & 2.9
\end{array}$$

Thus for variety $V1(j = 1)$ only method A results in significantly higher yield compared to method B.

The simultaneous 95% confidence intervals for the family of comparisons of the three methods for variety $V2(j = 2)$ are:

$$\begin{array}{rcccl} 0.4 & \leq & \mu_{12} - \mu_{22} & \leq & 12.8 \\ -4.3 & \leq & \mu_{12} - \mu_{32} & \leq & 8.1 \\ -10.9 & \leq & \mu_{22} - \mu_{32} & \leq & 1.5 \end{array}$$

Comparisons of the method means for variety $V2$ are similar to those of variety $V1$.

The simultaneous 95% confidence intervals for the family of comparisons of the three methods for variety $V3$ are:

$$\begin{array}{rcccl} 1.5 & \leq & \mu_{13} - \mu_{23} & \leq & 13.9 \\ -0.4 & \leq & \mu_{13} - \mu_{33} & \leq & 12.0 \\ -8.1 & \leq & \mu_{23} - \mu_{33} & \leq & 4.3 \end{array}$$

Comparisons of the method means for variety $V3$ are similar to those of variety $V1$ and $V2$.

The simultaneous 95% confidence intervals for the family of comparisons of the three methods for variety $V4(j = 4)$ are:

$$\begin{array}{rcccl} 6.3 & \leq & \mu_{14} - \mu_{24} & \leq & 18.7 \\ 5.0 & \leq & \mu_{14} - \mu_{34} & \leq & 17.4 \\ -7.5 & \leq & \mu_{24} - \mu_{34} & \leq & 4.9 \end{array}$$

For variety $V4$, method A results in significantly higher yields when compared to both methods B and C.

The simultaneous 95% confidence intervals for the family of comparisons of the three methods for variety $V5(j = 5)$ are:

$$\begin{array}{rcccl} -3.1 & \leq & \mu_{15} - \mu_{25} & \leq & 9.3 \\ 3.5 & \leq & \mu_{15} - \mu_{35} & \leq & 15.9 \\ 0.4 & \leq & \mu_{25} - \mu_{35} & \leq & 12.8 \end{array}$$

For variety $V5$ the mean yield for method A is not significantly higher than for method B as it was for the other four varieties. Method A results in significantly higher yield when compared to C, similar to variety $V4$. Method B results in significantly higher yield when compared to method C, unlike the insignificant comparisons between these two methods for the other varieties.

## 6.7 SAS Code for Chapter 6

### 6.7.1 Paper Towel Example

```
*  Paper Towel Example;

*  Input data;
data PaperTowel;
  input Towel $ Liquid $ Treatment $ AmountAbsorbed;
datalines;
Coronet  Water      CW  26
Coronet  Water      CW  22
Coronet  Water      CW  22
Coronet  Detergent  CD  19
Coronet  Detergent  CD  16
Coronet  Detergent  CD  15
Coronet  Oil        CO  22
Coronet  Oil        CO  25
Coronet  Oil        CO  29
Kleenex  Water      KW  43
Kleenex  Water      KW  41
Kleenex  Water      KW  41
Kleenex  Detergent  KD  33
Kleenex  Detergent  KD  38
Kleenex  Detergent  KD  38
Kleenex  Oil        KO  39
Kleenex  Oil        KO  41
Kleenex  Oil        KO  45
Scott    Water      SW  27
Scott    Water      SW  26
Scott    Water      SW  25
Scott    Detergent  SD  21
Scott    Detergent  SD  20
Scott    Detergent  SD  21
Scott    Oil        SO  27
Scott    Oil        SO  25
Scott    Oil        SO  25
;
run;
*  Calculate and print means for amount absorbed;
proc means data = PaperTowel;
  class Towel Liquid;
  var AmountAbsorbed;
  output out = Summary  mean = MeanAbsorbed;
run;
proc print data = Summary;
```

```
run;

*  Proc glm for obtaining ANOVA table and Tukey-Kramer pairwise comparisons;
proc glm data = PaperTowel;
  class Towel Liquid;
  model AmountAbsorbed = Towel Liquid Towel*Liquid;
  lsmeans Towel Liquid / pdiff cl t adjust = tukey;
run;
```

# Problems for Chapter 6

6.1* This example is based on an experiment described in Saliva [26], page 382. The effects of two constituents of fertilizers, Potash and Nitrogen, on the yield of tomato plants (pounds per plant) were studied. The yields for three tomato plants are measured at each combination of 2 levels of Potash and 4 levels of Nitrogen. The mean yields at the different combinations are given in the table below.

| | Nitrogen | | | |
|---|---|---|---|---|
| | 5% | 10% | 15% | 20% |
| Potash | | | | |
| 10% | 10.0 | 10.3 | 12.7 | 8.3 |
| 15% | 8.3 | 12.3 | 16.0 | 12.7 |

Suppose that MSE is 3.625.

a. Determine the estimated main effects of Potash

b. Determine the estimated main effects of Nitrogen

c. Determine the estimated interaction effects of Potash and Nitrogen

d. Carry out an F test to determine if there are true interaction effects. Use $\alpha = 0.10$.

e. Carry out an F test to determine if there are true Potash main effects. Use $\alpha = 0.05$.

f. Carry out an F test to determine if there are true Nitrogen main effects. Use $\alpha = 0.05$.

6.2* The author's son used his Nerf gun to shoot at a target on a glass door. The target was a circle having roughly the same diameter of the Nerf bullet. He shot the gun from three ranges:

- Short: 5 feet from the door
- Medium: 10 feet from the door
- Long: 15 feet from the door

He also shot the gun using both his dominant right hand and his left hand. He started each shooting by holding the gun in an upright position. He was then instructed to aim and then after two seconds was instructed to shoot at the target. There were 5 replications of each combination of shooting distance and hand assigned completely at random through time. Thus this is a two-factor completely randomized design.

Accuracy was measured by how far away (to the nearest 1/8 inch) the closest edge of the bullet was from the closest edge of the target. If the bullet touched the target at all, then the accuracy was 0. So smaller values of accuracy here denote closer shots to the target.

| | Left Hand | | Right Hand | |
|---|---|---|---|---|
| | Accuracy | Time Order | Accuracy | Time Order |
| Shooting Distance | | | | |
| Short | 0 | 3 | 3.375 | 1 |
| | 1.500 | 7 | 0.375 | 10 |
| | 0.000 | 13 | 2.125 | 16 |
| | 0.625 | 15 | 0.250 | 24 |
| | 2.000 | 19 | 0.500 | 29 |
| Medium | 3.500 | 5 | 1.000 | 2 |
| | 3.250 | 9 | 4.875 | 18 |
| | 0.125 | 17 | 1.000 | 20 |
| | 3.250 | 21 | 3.250 | 23 |
| | 2.125 | 26 | 4.625 | 28 |
| Long | 13.250 | 4 | 3.125 | 12 |
| | 7.000 | 6 | 1.125 | 14 |
| | 8.125 | 8 | 14.375 | 22 |
| | 7.750 | 11 | 3.375 | 27 |
| | 8.750 | 25 | 9.125 | 30 |

a. Construct an interaction plot putting HAND on the horizontal axis. Describe what you see in the plot. Is there evidence of interaction between hand used and distance.

b. Conduct a test of interaction between hand used and distance using a significance level of 0.10.

   i. If the interaction term is significant, use simultaneous 95% Tukey-Kramer confidence intervals to make pairwise comparisons of the mean accuracies of the three distances when using the left hand. Repeat this procedure for the right hand.

   ii. If the interaction term is not significant, test for differences in distance main effect means. Also test for differences in hand main effect means. Make pairwise comparisons using simultaneous 95% Tukey-Kramer confidence intervals where appropriate.

6.3* Alissa Wunder (2005) did an experiment to study the effect of heat in a microwave on the expansion of a marshmallow. Marshmallows were placed at the bottom of a mug and the mug placed in a microwave at one of two settings, medium and high. Three different brands of marshmallows were also studied (Food Lion, Walmart, and Kraft Jet Puff). The experiment was replicated four times at each combination of microwave setting and brand for a total of 24 marshmallow roastings. Marshmallows were tested one at a time with the particular setting and brand being randomly selected. Thus the experiment is a two factor completely randomized design. The data from the experiment is given in the following table.

| Time Order | Brand | Level | Amount of Time(seconds) |
|---|---|---|---|
| 2 | Food Lion | Medium | 16 |
| 8 | | | 37 |
| 14 | | | 15 |
| 19 | | | 16 |
| 1 | Food Lion | High | 19 |
| 11 | | | 18 |
| 18 | | | 18 |
| 23 | | | 23 |
| 3 | Jet Puff | Medium | 39 |
| 10 | | | 38 |
| 17 | | | 39 |
| 20 | | | 37 |
| 6 | Jet Puff | High | 16 |
| 9 | | | 17 |
| 15 | | | 18 |
| 21 | | | 17 |
| 4 | WalMart | Medium | 15 |
| 12 | | | 44 |
| 16 | | | 44 |
| 22 | | | 43 |
| 5 | WalMart | High | 16 |
| 7 | | | 19 |
| 13 | | | 22 |
| 24 | | | 20 |

a. Construct a plot of amount of time versus combination of brand and microwave level. Draw conclusions based on the plot.

b. Give a model for the data and describe the terms of the model in context.

c. Use a statistical program to obtain an ANOVA table with P-values.

   i. What is the estimate of the variance of the error terms?

   ii. If the interaction term is significant at the 0.10 level, use simultaneous 95% Tukey-Kramer confidence intervals at each level of microwave to make pairwise comparisons of the mean times of the store brands.

   iii. If the interaction term is not significant at the 0.10 level of significance perform the F test for differences in main effect mean amount of time across brands. Also perform the F test for differences in main effect mean amount of time across microwave level. Make pairwise comparisons using simultaneous 95% Tukey-Kramer confidence intervals where appropriate.

6.4* Annie Hambrick and Kristen Haug in 2004 compared the melting times of different brands of butter. The brands used were Land O'Lake, Great Value (Walmart), and Cabot. They were also interested in comparing melting times for different heat sources and thought that perhaps heat source would have an effect on the comparison of the brands. So another factor, heat source, was studied: burner on a stove or toaster oven. The stove burner and toaster oven were turned on at the start of the experiment and remained on during the entire time of the experiment. The heat settings for the two sources were set so that in theory roughly the same temperature was produced. A replication involved the selection of a brand at random and then the selection of a heat source. One tablespoon of the selected brand of butter was then put in a sauce pan if the stove was used or put on a foil covered tray if the toaster oven was selected. The sauce pan was put on the burner for two minutes before placing the butter in it. The saucepan was washed between replications with soap and hot water to prevent the pan from cooling down completely. In the event that the saucepan cooled down, it was left on the burner for two minutes before moving on to the next replication. The tray remained in the toaster oven the entire experiment - only the piece of foil with the butter was removed. The amounts of time to butter meltdown are given in the following table.

| Time Order | Brand | Method | Amount of Time(secs) |
|---|---|---|---|
| 1 | Land-O-Lakes | Stove | 173 |
| 2 | Cabot | Stove | 97 |
| 3 | Great Value | Stove | 150 |
| 4 | Land-O-Lakes | Stove | 125 |
| 5 | Great Value | Stove | 154 |
| 6 | Land-O-Lakes | Toaster Oven | 166 |
| 7 | Land-O-Lakes | Toaster Oven | 179 |
| 8 | Great Value | Stove | 157 |
| 9 | Land-O-Lakes | Stove | 158 |
| 10 | Great Value | Toaster Oven | 206 |
| 11 | Cabot | Stove | 110 |
| 12 | Great Value | Toaster Oven | 195 |
| 13 | Cabot | Toaster Oven | 177 |
| 14 | Cabot | Toaster Oven | 197 |
| 15 | Land-O-Lakes | Toaster Oven | 203 |
| 16 | Cabot | Toaster Oven | 183 |
| 17 | Cabot | Stove | 126 |
| 18 | Great Value | Toaster Oven | 205 |

a. Construct a plot of amount of time versus combination of brand and heat source. Is there evidence of a brand effect? heat source effect?

b. Give a model for the data and describe the terms of the model in context.

c. Conduct a test of interaction between heat source and brand using a significance level of 0.10.

   i. If the interaction is significant, use simultaneous 95% Tukey-Kramer confidence intervals to make pairwise comparisons of the levels of brands only for the oven heat. Repeat this procedure when heat source is the stove. Draw conclusions.

ii. If the interaction term is not significant, then use the F test to test for differences in sources of heat. Use the F test to test for differences in main effect means across brands. Make appropriate pairwise comparisons using Tukey-Kramer simultaneous confidence intervals.

6.5* Consider the following incomplete ANOVA table for a two factor completely randomized design.

| Source of Variation | Df | SS | MS | F |
|---|---|---|---|---|
| A | 3 | 310 | ___ | ___ |
| B | 2 | ___ | ___ | ___ |
| A*B | ___ | 80 | ___ | ___ |
| Error | 24 | 400 | ___ | |
| Total (Corrected) | 35 | 890 | | |

a. How many levels of A are there? How many levels of B?

b. What is the total number of observations on the response variable?

c. At a significance level of $\alpha = 0.05$ is the interaction significant? Use a critical F value from Table A.7.

d. At a significance level of $\alpha = 0.05$ are the A main effects significant? Use a critical F value from Table A.7 to make a decision.

e. At a significance level of $\alpha = 0.05$ are the B main effects significant? use a critical F value from Table A.7 to make a decision.

6.6* In the article "Lipid Pattern in Experimental Canine Atherosclerosis" (*Circular Research* [1964]: Vol. XIV, pgs 61-72) researchers investigated the effects on total serum lipids (mg/100 ml serum) of the addition of cholesterol and thiouracil to the diets of canines. The treatment group sizes, means, and standard deviations are given in the following table.

| Diet | n | Mean | Standard Deviation |
|---|---|---|---|
| Basal | 4 | 556 | 93.8 |
| Basal + Cholesterol | 6 | 879 | 357.6 |
| Basal + Thiouracil | 6 | 1807 | 497.2 |
| Basal + Cholesterol + Thiouracil | 6 | 3393 | 967.5 |

Explain how the four diets can be regarded as a factorial structure involving two factors.

6.7 Rebecca Aaron and Rebecca Redman in 2014 conducted an experiment to study the effects of brand of rubber band and temperature conditioning of the rubberband on how far a rubber band would stretch until breaking. The two brands of rubberbands were Staples and CLI (Douglas Stewart Company). Both brands were approximately 9 cm in length and 0.32 cm in width. Prior to stretching rubberbands were exposed to either cold (freezer for 24 hours), heat (microwave for 1 minute), or room temperature. Rubberbands were tested/stretched one a time. The procedure was as follows. A combination of brand and temperature was randomly selected. A rubberband of the randomly selected brand and temperature conditioning was selected. An apparatus was used to stretch the rubber-band until it broke with a tape measure below the rubberband to measure the stretched length. The stretching was videotaped to aid in the measurement. This process was repeated for a total of 30 rubberbands, 5 for each of the combinations of brand and temperature conditioning. The design is a completely randomized design.

| | Staples | | CLI | |
|---|---|---|---|---|
| | Length | Time Order | Length | Time Order |
| Temperature | | | | |
| Freezer | 71.1 | 1 | 73.5 | 12 |
| | 74.0 | 6 | 66.0 | 13 |
| | 76.2 | 18 | 78.5 | 14 |
| | 73.6 | 22 | 69.1 | 23 |
| | 75.0 | 26 | 72.0 | 27 |
| Room | 77.2 | 5 | 67.5 | 2 |
| | 80.0 | 8 | 76.8 | 9 |
| | 75.0 | 19 | 69.5 | 15 |
| | 69.5 | 24 | 63.5 | 16 |
| | 72.5 | 28 | 70.3 | 20 |
| Heated | 72.6 | 3 | 72.5 | 7 |
| | 76.2 | 4 | 75.5 | 10 |
| | 79 | 25 | 59.2 | 11 |
| | 66.6 | 29 | 70.5 | 17 |
| | 72.5 | 30 | 69.0 | 21 |

a. What are the experiment units?

b. It is most likely that the 15 rubberbands that required cold conditioning were all put in the freezer at the same time. Discuss any potential biases that occur as a result of this.

c. Construct an interaction plot putting Temperature on the horizontal axis. Describe what you see in the plot. Is there evidence of interaction between temperature and brand.

d. Use software to conduct a test of interaction between temperature and brand using a significance level of 0.10.

   i. If the interaction term is significant, use simultaneous 95% Tukey-Kramer confidence intervals to make pairwise comparisons of the mean accuracies of the three temperatures with the Staples rubberbands. Repeat this procedure for the CLI rubberbands.

ii. If the interaction term is not significant, test for differences in brand main effect means. Also test for differences in temperature main effect means. Make pairwise comparisons using simultaneous 95% Tukey-Kramer confidence intervals where appropriate.

6.8 Rand Abdulrazaq and Mabel Adubofour in 2014 conducted an experiment to study the effects of type of liquid and amount of salt on the buoyancy of an egg. The two liquids were water and Sprite. The amounts of salt were 6, 8, and 14 tablespoons. One trial of the experiment was conducted as follows. A combination of liquid type and amount of salt was randomly selected. Four hundred and seventy (470) milliliters of the chosen liquid was poured into a graduated beaker. The selected amount of salt was added to the beaker and the contents were mixed for one minute. Ann egg was put into the solution and the distance from the bottom of the egg to the bottom of the cup was measured (cm). This process was repeated for a total of 30 trials, 5 replicates per each of the 6 combinations of liquid type and Sprite. The same egg and beaker were used for all 30 trials. The design is a completely randomized design. The distances measuring solubility are provided in the table below.

| | Water | | Sprite | |
|---|---|---|---|---|
| | Distance | Time Order | Distance | Time Order |
| Amount of Salt(Tbs) | | | | |
| 6 | 1.5 | 4 | 2.2 | 5 |
| | 1.4 | 9 | 2.2 | 8 |
| | 1.5 | 15 | 2.4 | 16 |
| | 1.5 | 23 | 2.1 | 21 |
| | 1.6 | 25 | 2.1 | 28 |
| 8 | 2.2 | 1 | 2.6 | 11 |
| | 2.3 | 2 | 2.9 | 14 |
| | 2.2 | 3 | 2.8 | 18 |
| | 2.2 | 12 | 2.7 | 20 |
| | 2.0 | 19 | 2.7 | 26 |
| 14 | 2.9 | 7 | 2.8 | 6 |
| | 3.1 | 17 | 2.8 | 10 |
| | 2.5 | 22 | 3.0 | 13 |
| | 2.5 | 24 | 3.1 | 29 |
| | 2.9 | 27 | 2.8 | 30 |

a. What are the two factors in this study? What are the levels of each factor?

b. What are the treatments in this study? How many treatments are there?

c. What are the experimental units?

d. Give two extraneous variables that are being directly controlled in the study.

e. What is the purpose of randomizing the treatments across time? Give an example in your discussion.

f. Construct a dot plot putting combination of liquid type and amount of salt on the horizontal axis and distance on the vertical axis. Comment within the context of this study.

g. Construct an interaction plot putting type of liquid on the horizontal axis. Is there evidence of interaction between liquid type and amount of salt? Yes or no and explain within the context of this study.

h. Give the population effects model for this data. Describe all terms in the model within the context of this study. Be sure to give the assumptions associated with the errors. Give two extraneous variables whose effects are part of the error term.

i. Use software to obtain an ANOVA table.

j. Use software to conduct a test of the interaction between liquid type and amount of salt on buoyancy using a significance level of 0.10.

   i. If the interaction term is significant, use simultaneous 95% Tukey-Kramer confidence intervals to make pairwise comparisons of the mean buoyancies/distances of the three amounts of salt assuming liquid type is Sprite. Repeat this procedure assuming liquid type is water. In all cases provide the treatments means that are being compared.

   ii. If the interaction term is not significant, test for differences in liquid type main effect means. Also test for differences in amount of salt main effect means. Make pairwise comparisons using simultaneous 95% Tukey-Kramer confidence intervals where appropriate. In all cases provide the main effect means that are being compared.

6.9 Refer to the paper towel absorption data given in Table 6.8.

   i. Decompose each of the 27 absorptions according to the two factor model for the completely randomized design and put your decompositions into tabular form similar to Table 6.5, not corrected for the grand mean.

   ii. Based on your decomposition in part (i), calculate the various sums of squares for the absorptions (uncorrected for the grand mean), grand mean, towel effects, liquid effects, towel by liquid interaction effects, and errors. Compare your calculated sums of squares for towel effects, liquid effects, liquid interaction effects, and errors to the same sums of squares given in Table 6.10 obtained by statistical software. Do you agree?

   iii. Find the total sum of squares for the absorptions corrected for the grand mean using your calculated total sum of squares for absorptions not corrected for the grand mean and sum of squares for the grand mean in part (ii). Compare this result to the sum of squares for absorptions corrected for the grand mean given in Table 6.10.

# Chapter 7

# Blocking and the Randomized Complete Block Design

## 7.1 Blocking Designs Compared to Completely Randomized Designs

Recall from Chapter 1 that there are three ways that researchers "control" for the potentially biasing effects of extraneous variables.

a. Randomization

b. Blocking

c. Direct Control

Randomization means assigning the treatments to the experimental units at random so as to balance out the effects of extraneous variables among the treatment groups. It is important to note that the effects of extraneous variables balanced out by randomization are not eliminated altogether. In fact for small group sizes randomization may not work well at all. Any designs which employ only randomization are called **completely randomized designs**.

Direct control means that we only use experimental units which have constant values with regard to some extraneous variable. For example, if gender is an extraneous variable, then we might only use females in the study. In this form of control the effects of the extraneous variable are eliminated altogether but of course, the scope of the conclusions are limited.

Blocking is a form of control whereby experimental units are "blocked" or grouped into homogeneous sets and treatments are then assigned at random within each block. By grouping the units into sets, we can in the analysis remove the effects of the blocking variables from experimental error and thus make for a more precise comparison of treatments. More precisely, the purpose of blocking is to reduce the standard deviation, $\sigma$, of experimental error. In theory blocking will eliminate altogether the effects of an extraneous variable– in practice the effect may not be eliminated altogether, but reduced to a certain extent.

As an example of blocking suppose that a researcher wants to compare four brands of tires for treadwear by having the tires put on cars and driven. Suppose that there are four cars available with thus 16 tire positions. One possible design, the **completely randomized design**, randomly assigns 16 of the tires, 4 of each brand, to the 16 tire positions on the four cars in a completely randomized fashion. The resulting assignment could turn out as follows:

| | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Car 1 | A | C | A | A |
| Car 2 | C | B | A | B |
| Car 3 | D | C | B | D |
| Car 4 | C | B | D | D |

**To emphasize:** this is the result of assigning the brands **completely at random** to the 16 tire positions, that is a **completely randomized design**.

Intuitively, the completely randomized design is not a very good design for this experiment. It is possible, like in this design, that three tires of the same brand get put on the same car (Car 1, Brand A) and if car is a significant extraneous variable, then the resulting comparison between brands would be biased in favor or disfavor of brand A. Numerically, the mean treadwear for Brand A might be smaller/larger than the mean treadwear for the other brands, but it may be due to the effects of Car/Driver 1.

A more intuitively appealing design is a **block design**. The 16 tire positions are naturally grouped by car. Thus use car as a block and assign the four brands within each block or car. Note that **randomization is still possible with blocking**. The brands can be randomly assigned to the 4 tire positions within each car or block. The purpose of the randomization within each car is to balance out the effects of other extraneous variables such as position effects. The random assignment for a block design in this example with cars serving as blocks or groups of tire positions could result as follows:

| | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Car 1 | A | C | B | D |
| Car 2 | C | B | A | D |
| Car 3 | D | B | A | C |
| Car 4 | A | D | C | B |

Notice with this design and random assignment that each brand is exposed to all four cars so comparison of brands will not be unduly influenced by effects of cars. It is possible with the above randomization that all tires of a brand are put on the same wheel position and this could have a confounding effect on the comparison of brands. An alternative blocking design, called the Latin Square

Design, with two blocking criteria, can be used in this situation. A discussion of the Latin Square blocking design will be discussed in Section 7.8.

**Blocking is a restricted form of randomization. This is different than the completely randomized design where there are no restrictions on the randomization**. Determining the kind of randomization can aid in determining the kind of design.

## 7.2 Types of Blocking

Recall in Chapter 3 that we learned how to analyze block designs with only two treatments using the paired samples $t$ test. We also learned about the different types of blocking. Listed again below are the different types reflecting the fact that in this chapter there may be more than two treatments of interest.
**Types of Blocking**

A. Group/sort subjects or objects into blocks, each block containing $t$ subjects, with $t$ being equal to the number of treatments. This would include natural groupings such as twins or litters of animals, tire positions on a car, etc. The $t$ treatments are assigned completely at random to the $t$ subjects within each block. Experimental units in this type of blocking are the subjects or objects that have been grouped.

B. Reuse each subject at different occasions or time slots. There are two types of reusing.

   i. Each subject is reused at $t$ time slots in order to receive different treatments at the time slots, the number of treatments being equal to the number of time slots. The order of the treatments is random.

   ii. Each subject is reused to obtain multiple measurements on the response variable at the different time slots in order to study a time profile. The "factor" in this case would be a time variable associated with the time slots. There would be no randomization of levels of the time variable to the time slots. case.

   Experimental units for the reusing type of blocking (i. or ii.) are the time slots or occasions for each subject. The grouping is of time slots or occasions by subject.

C. Split chunks of material such as a batch of milk, plot of land, etc. into $t$ parts, the number of parts equal to the number of treatments. The $t$ treatments are assigned completely at random to the $t$ parts. Experimental units in this type of blocking are the parts.

In all cases above there is some grouping, whether it is of persons, time slots, or parts of batches of some substance.

It is assumed in this chapter that the number of experimental units within a block is equal to the number of treatments, $t$, and that the treatments are assigned at random to the experimental units with each block independently of other blocks. The design is then called a **randomized complete block**

**design**. The word **complete** refers to the fact that all treatments are used in a block. This is not always the case. Then the design would be an **incomplete block design**, a design not covered in this text.

In some designs subjects are randomly assigned to treatments in a completely randomized design, but then are re-used as in Blocking Type B(ii) to obtain multiple measurements on the response variable. Interest would be in the time profiles of individual treatments and comparing the time profiles for different treatments. Subjects serve as blocks of time slots since they are being reused. For example, .... Designs which use Type B (ii) blocking will be referred to as **repeated measures designs** in this text. The analysis deserves special consideration and will be the subject of Chapter 10. Note that some textbooks refer to a repeated measures design as any design which uses Type B(i) or Type B(ii) blocking.

### 7.2.1 Examples of Type A Blocking

a. The tire brand example of the last section is an example of this type of blocking. This is a natural grouping–the blocks of 4 tire positions come naturally by car. Other examples of this type of blocking would be where litters of animals are used as blocks. Each animal in a litter gets a treatment (assigned at random) and several litters are used. The experimental units are animals within the litters.

b. A study was conducted to investigate the effects of adding the amino acid lysine to the diets of growing kittens. Twenty kittens were weighed, sorted by increasing weight, and then divided into 4 blocks, each block with 5 kittens of similar weight. Within each block the five kittens were assigned at random to one of 5 diets with varying levels of lysine. Kittens were fed individually. Kittens are the experimental units.

c. Cobb ([4], page 247) describes a study of the effect of vitamin B6 on the severity of symptoms of premenstrual syndrome. Subjects were sorted into pairs or blocks based on similar responses to a questionnaire completed before the study regarding the severity of their symptoms. Within each block one subject was randomly assigned vitamin B6 and the other a placebo. Experimental units are the subjects since the treatments (vitamin B6, placebo) are assigned to the subjects.

### 7.2.2 Examples of Type B Blocking

a. To compare three drugs A, B, and C, for their effectiveness in relieving an allergy, each of 10 subjects receives all three drugs in a random order in different time periods. **Experimental units are time slots/periods.**

   The blocking/grouping here is of time slots by person and then **randomization is undertaken within each group of time slots**.

   One version of the completely randomized version of this study would be where the treatments are assigned to the 30 time slots at random. Thus in theory one person could get drug A for all three of his/her time slots. Of course this would not make sense, just as it would not make sense to randomly assign brands to the 16 wheel positions on four cars.

An alternative version of the completely randomized design, but that does not reuse subjects is as follows: thirty subjects are assigned completely at random, ten getting drug A, ten different subjects getting drug B, and ten different subjects getting drug C. **Experimental units are persons here, not time slots**. Thus any differences in persons would be a part of experimental error and thus may not be a very good design, that is we may have imprecise comparisons of the drugs.

b. In the article "Effects of a fibre-enriched milk drink on insulin and glucose levels in healthy subjects" (*Nutrition Journal [2009], 8:45*) researchers compared the effects of three drinks: 1) a lactose-free milk drink, 2) a novel fibre-enriched, fat- and lactose-free milk drink, and 3) normal fat-free milk, on serum glucose and insulin levels and satiety using a randomized complete block design. After fasting overnight, each of the 26 volunteers ingested 200 ml of one of the three drinks on three separate days, the order of the drinks being randomly determined. This is a re-using type of blocking. Each subject is being re-used on three different days. The experimental units are the time slots/days on which the subjects ingested the drinks.

c. **Before and After Studies**. A popular type of block design which reuses subjects is the **before and after** study. A person is measured on the response variable **before** a treatment is given, a treatment is given, and then the person is measured again on the response variable **after** the treatment. This is a special case of the repeated measures design with only two observations on the response variable for each subject. Interest is in the time profile of the response, before the treatment to after the treatment, or the change in the response that might result from the treatment.

An example of this comes from Moore and McCabe ([22], page 560). A bank wonders whether eliminating the annual fee on its credit card customers will increase the amount that the customers charge. A random sample of 100 customers is selected and told that they would not have to pay the fee this year. The amounts that they charged last year (**before elimination of fee**) and the amounts charged this year (**after elimination of the fee**) are compared. This is a block design whereby each customer is used/measured twice. Each subject provides two time slots, two consecutive years. The factor is place in time of the years, "before" and "after," which are **not assigned** to years.

d. It is claimed that an industrial safety program is effective in reducing the loss of working hours due to accidents. The following data are collected concerning the weekly loss of working hours due to accidents in six plants both before and after the safety program is instituted.

| | Plant | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Before | 12 | 29 | 16 | 37 | 28 | 15 |
| After | 10 | 28 | 17 | 35 | 25 | 16 |

This is a block design whereby each plant/set of employees is used/measured twice. Each plant provides two time slots, two different week periods. The factor is point in time of the weeks, "before" and "after," which are not assigned to weeks. This is an example of a repeated measures design.

e. Suppose that 50 high school students agree to take the SAT test twice, once before a special prep course advertised to improve your score, and then again after taking the prep course. The two tests are different versions. Each student serves as a block of two time periods/occassions. The response variable is SAT score. The factor of interest is place in time for the two periods in which the tests are taken, "before" and "after," which are not assigned to the periods.

Note that the above described before and after studies are similar to the drug example in that time slots are experimental units. However in the drug example the treatments/conditions of drug A, B, and C are randomly assigned to time slots. This is to balance out any time effects for the measurements on A, B, and C. Some of the A measurements are taken $1^{st}$, some are taken $2^{nd}$, some $3^{rd}$, etc.

In the before and after study, a comparison is made of **before** time slot measurements and **after** time slot measurements. However, there is one big difference. The conditions **before** and **after** are not assigned (at random) to the time slots, like drugs are. They are inherent characteristics of the time slots in which the measurements are taken. Thus there is no balancing out of time effects among the before measurements and the after measurements. All of the before measurements are taken $1^{st}$ in time, all of the after measurements are taken $2^{nd}$ in time. Thus if there are any extraneous variables which are time related they will be confounded with the before and after measurements.

In the bank study, any effects on amount of money spent related to time, would be confounded with the effects of the no fee option. For example, when the "after no fee option" amounts were recorded, perhaps the economy was more prosperous than when the "before no fee option" amounts were recorded.

In the SAT study, perhaps when students took the test a $2^{nd}$ time, they may have done better, because they had more experience (practice effect) with this kind of test than when they took it the first time.

### 7.2.3 Examples of Type C Blocking

a. In agricultural studies blocks may represent different parts of fields getting different amounts of moisture which is associated with growth of plants. A block would correspond to a part of field or plot which is split into subplots and the subplots would have about the same amount of moisture. For example, there might be 5 plots with each plot being split into 3 subplots. The subplots within each of the 5 plots have about the same moisture. The treatments, which might be 3 types of fertilizer, are applied within each plot to the 3 subplots.

b. (Johnson & Sui,[13]). A food scientist wants to study whether quality differences exist between yogurt made from skim milk with and without the pre-culture of a particular type of bacteria, called Psychotrops(PC). Samples of skim milk are procured from seven dairy farms. One half of the milk sampled from each farm is inoculated with PC, and the other half is not. After yogurt is made with these milk samples, the firmness of the curd is measured, and those measurements are given below.

| | Dairy Farm | | | | | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G |
| With PC | 68 | 75 | 62 | 86 | 52 | 46 | 72 |
| Without PC | 61 | 69 | 64 | 76 | 52 | 38 | 68 |

A block corresponds to a pair of milk samples from a farm. The two samples arose as a result of splitting a larger portion of milk. The blocking eliminates the effects of different farms from the comparisons of the firmness with PC and without PC.

c. In the article “Effect of Rapid Thawing on the Meat Quality Attributes of USDA Select Beef Strip Loin Steaks” (Journal of Food Science, [2011]: Vol. 76, Issue 2, pages S156-S162) researchers compared three methods for thawing frozen beef strip loin steaks on various quality characteristics of the meat. Each of 24 beef strip loins were cut into 3 steaks. The three steaks from each strip loin were randomly assigned to three thawing methods. There was one conventional method (18 to 20 hrs, 4 degrees C), and two rapid thawing methods (20 min, 20 degrees C) or very fast (11 min, 39 degrees C). The rapid thawing methods were conducted in a circulating water bath. The factor of interest is thawing method. The 24 strip loins serve as blocks. The randomization of the thawing methods was done independently from block (loin) to block to the steaks cuts from the strip loins. The experimental units are the steaks. The response variables are the quality characteristics.

## 7.3 Model and Analysis for the Randomized Complete Block Design

### 7.3.1 Block Design Analysis as Analysis for Two Factor Study

Think of the levels of a single blocking variable as levels of one of the two factors, say A, in a two factor study. If there are $a$ levels of the blocking variable A and $b$ levels of the factor of interest B, then let $\mu_{ij}$ be the true mean of the response variable at the $i^{th}$ level of the blocking variable A and $j^{th}$ level of factor B. Then the model from Chapter 6 is

$$\begin{aligned} y_{ij} &= \mu_{ij} + \epsilon_{ij} \\ &= \mu_{..} + \rho_i + \tau_j + (\rho\tau)_{ij} + \epsilon_{ij} \qquad (7.1) \end{aligned}$$

where

- $\mu_{..}$ represents the true grand mean
- $\rho_i$ $(i = 1, ..., a)$ represents the true effect of $i^{th}$ level of the blocking variable
- $\tau_j$ $(j = 1, ..., b)$ represents the $j^{th}$ level of the factor of interest
- $(\rho\tau_{ij})$ represents the interaction between the $i^{th}$ level of the blocking variable and the $j^{th}$ level of the factor of interest, B, and
- $\epsilon_{ij}$ represents as usual the effects of extraneous variables on the observation of $y$ at the $ij^{th}$ combination of the blocking variable and factor of interest.

Note that the subscript $k$ has been dropped because there is only one observation at the $i^{th}$ level of the blocking variable and $j^{th}$ level of the factor of interest. Note also the difference in notation this model uses compared to the two factor model of Chapter 6. The symbol $\rho$ is being used instead of $\alpha$ for the effect of a level of the blocking variable. Also the symbol $\tau$ is being used instead of the symbol $\beta$ for the factor of interest, A. Otherwise the model is the same as that in Chapter 6.

Let us think about estimating parameters in this model. We may proceed as in Chapter 6. Letting

$$\epsilon_{ij} = y_{ij} - \mu_{ij}$$

it is natural to estimate $\epsilon_{ij}$ with $y_{ij} - \overline{y}_{ij.}$ where $\overline{y}_{ij.}$ is the sample mean at the $i^{th}$ level of the blocking variable and $j^{th}$ level of the factor of interest. However there is only one observation at the $i^{th}$ level of the blocking variable and $j^{th}$ level of A. Thus $\overline{y}_{ij.}$ would be the same as $y_{ij}$ and the estimate of the error term would be 0. Obviously this will not give a legitimate estimate of error and hence of MSE. The problem is that there is not enough data "to go around" and estimate all of the parameters in the model.

Note that if we assume that there is no interaction between blocks and treatments, then the model simplifies to

$$y_{ij} = \mu_{\cdot\cdot} + \rho_i + \tau_j + \epsilon_{ij}$$

If we solve this equation for $\epsilon_{ij}$ we get

$$\epsilon_{ij} = y_{ij} - (\mu_{\cdot\cdot} + \rho_i + \tau_j)$$

This suggests estimating the error term $\epsilon_{ij}$ with

$$e_{ij} = y_{ij} - (\overline{y}_{\cdot\cdot} + \hat{\rho}_i + \hat{\tau}_j)$$

The right side is how we estimated in Chapter 6 the interaction effect of the $i^{th}$ level of one factor and the $j^{th}$ level of the other factor (with $y_{ij}$ replaced with $\overline{y}_{ij\cdot}$). So to estimate error in the block design with only one replication per block and treatment combination we use a value that served as interaction effect in Chapter 6. This is legitimate assuming there is no true interaction between blocks and treatments.

### 7.3.2 Model for the One Factor Randomized Complete Block Design

The model for the one factor randomized complete block design with no interaction is summarized below.

$$y_{ij} = \mu_{\cdot\cdot} + \rho_i + \tau_j + \epsilon_{ij} \qquad (7.2)$$

where

- $\mu_{\cdot\cdot}$ represents the true grand mean
- $\rho_i$ $(i = 1, ..., a)$ represents the true effect of $i^{th}$ level of the blocking variable
- $\tau_j$ $(j = 1, ..., b)$ represents the $j^{th}$ level of the factor of interest
- $\epsilon_{ij}$ represents as usual the effects of extraneous variables on the observation of $y$ at the $ij^{th}$ combination of the blocking variable and factor of interest.

It is assumed in the model that the $\epsilon_{ij}$ are independent normal random variables each with mean 0 and standard deviation $\sigma$.

The model assumes no interaction between the blocking variable and the factor of interest. Thus this condition needs to be checked. If this assumption if not reasonable then a transformation of the data may be helpful.

The F test for treatments compares the sample treatment means averaged over levels of the blocks which is appropriate only if there is no interaction.

To illustrate the ideas consider the following example taken from Kutner, Nachtsheim, Neter, and Li [15].

**Example 7.1** *An accounting firm, prior to introducing in the firm widespread training in statistical sampling for auditing, tested three training methods:*

1. *study at home with programmed training materials*
2. *training sessions at local offices conducted by local staff*
3. *training sessions in Chicago conducted by national staff*

*Thirty auditors were grouped into* 10 *blocks of* 3*, according to time elapsed since college graduation, and the auditors in each block were randomly assigned to the 3 training methods. Block* 1 *consists of auditors graduated most recently, ..., block* 10 *consists of those graduated most distantly. At the end of the training, each auditor was asked to analyze a complex case involving statistical application; a proficiency measure based on this analysis was obtained for each auditor. The results are given in Table 7.1*

The model used for the analysis of this example is given below:

$$y_{ij} \quad = \quad \mu_{..} + \rho_i + \tau_j + \epsilon_{ij} \tag{7.3}$$

where

- $\mu_{..}$ represents the true grand mean of proficiency measure
- $\rho_i$ $(i = 1, ..., a = 10)$ represents the true effect of $i^{th}$ level of the blocking variable time elapsed since college graduation on the proficiency measure
- $\tau_j$ $(j = 1, ..., b = 3)$ represents the true effect of the $j^{th}$ level of training method on the proficiency measure
- $\epsilon_{ij}$ represents as usual the effects of extraneous variables on the observation of $y$, proficiency measure at the $ij$ combination of the blocking variable time elapsed since graduation and training method.

The model assumes no interaction between the blocking variable, time elapsed since college graduation, and training method.

Figure 7.1 provides a plot of the proficiency measures versus Block by Method. Note that Method 3 results in the highest measures regardless of the amount of time elapsed since college graduation. Also there is no strong evidence of interaction between block and method.

Let us find the estimate of the error associated with the proficiency measure, $y_{11} = 73$, for block 1, training method 1.

The grand mean proficiency measure is $\overline{y}_{..} = 77.0$. The marginal mean proficiency for block 1 is $\overline{y}_{1.} = 82$ and the marginal mean proficiency for training method 1 is $\overline{y}_{.1} = 70.6$. Thus block 1 effect is $\hat{\rho}_1 = 82 - 77 = 5$ and the training method 1 effect is $\hat{\tau}_1 = 70.6 - 77 = -6.4$. Hence the estimate of error, $e_{11}$, for 77 is

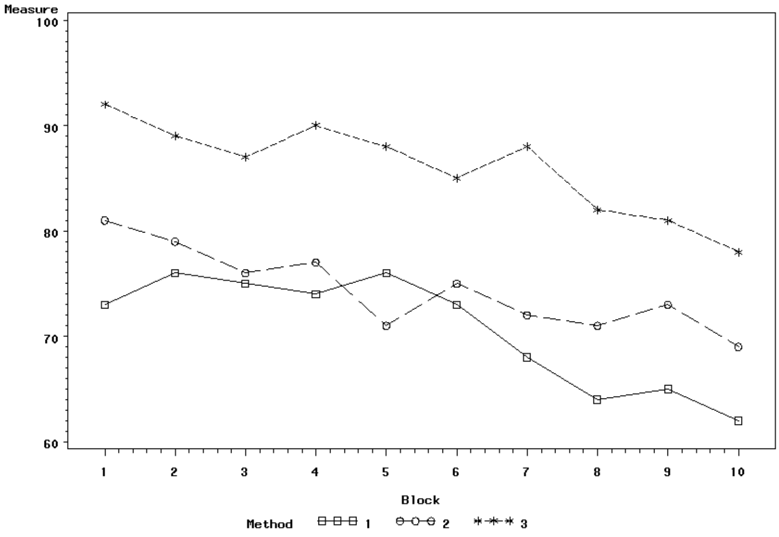
$$e_{11} = 73 - (77 + 5 + (-6.4)) = -2.6.$$

Hence we can decompose $y_{11} = 73$ in the following manner:

Table 7.1: Auditor Proficiency Measures with Marginal and Grand Means

| | Training Method j | | | |
|---|---|---|---|---|
| Block i | 1 | 2 | 3 | $\overline{y}_{i\cdot}$ |
| 1 | 73 | 81 | 92 | 82 |
| 2 | 76 | 79 | 89 | 81.3 |
| 3 | 75 | 76 | 87 | 79.3 |
| 4 | 74 | 77 | 90 | 80.3 |
| 5 | 76 | 71 | 88 | 78.3 |
| 6 | 73 | 75 | 85 | 77.7 |
| 7 | 68 | 72 | 88 | 76 |
| 8 | 64 | 71 | 82 | 72.3 |
| 9 | 65 | 73 | 81 | 73 |
| 10 | 62 | 69 | 78 | 69.7 |
| $\overline{y}_{\cdot j}$ | 70.6 | 74.4 | 86.0 | |
| | | | | $\overline{y}_{\cdot\cdot} = 77.0$ |

Figure 7.1: Plot of Proficiency Measure versus Block by Treatment



$$y_{11} = 73 = 77 + 5 + (-6.4) + (-2.6).$$

Similarly we can do this for the other 29 proficiency measures. Table 7.2 provides the complete decomposition.

If we square the effects in the various columns and add we get the following sums of squares:

- SSTOT with degrees of freedom = $ab$
- SSGM with 1 degree of freedom
- SSBL(Blocks) with degrees of freedom = $a-1$, number of blocks minus 1
- SSTR(Treatments) with degrees of freedom = $b-1$, number of treatments minus 1
- SSE with degrees of freedom = $(a-1)(b-1)$.

Note that degrees of freedom associated with the errors is the same as that used for interaction in Chapter 6. We will use a statistical program to obtain the sums of squares, degrees of freedom, and mean squares for the auditor example. Table 7.3 gives the ANOVA table for the auditor example. Note in the table that total sum of squares corrected for the grand mean is given instead of total sum of squares.

There is evidence of a difference in training method. There is also evidence of a difference in blocks but this is not unexpected since the blocking variable was included to control for differences in experience.

Pairwise comparisons can be made using one of the methods discussed in Chapter 5. The Tukey-Kramer method of multiple comparison is used here. Tukey-Kramer adjusted P-values and simultaneous 95% confidence intervals are in given in Table 7.4. All pairwise comparisons of means are significant at the 0.05 experimentwise level of significance.

The model upon which the inferences is based assumes that there is no interaction between block and method, that is the effect of method does not depend upon the number of years since graduation. One way of checking this assumption is as in Chapter 6, to plot scores versus block and check to see if the graphs representing the different treatments are roughly parallel. Figure 7.1 indicates that the assumption of no interaction between blocks and methods is reasonable for the auditor example.

The method for calculating errors for the block design can be thought of as an adjustment on the errors calculated under the assumption of a completely randomized design of Chapter 4. Consider the proficiency measure of 73 in block 1, training method 1, for Example 7.1. The following equations start with a decomposition of 73 along the lines of Chapter 4 and ends with a decomposition of 73 according to the block design.

$$73 \quad = \quad 77 + (70.6 - 77) + (73 - 70.6)$$

Table 7.2: Decomposition Table - Auditor Data

| i (block) | j(method) | $y_{ij}$ | = | $\overline{y}_{..}$ | + | $\hat{\rho}_i$ | + | $\hat{\tau}_j$ | + | $e_{ij}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 73 | = | 77 | + | 5 | + | (-6.4) | + | (-2.6) |
| 1 | 2 | 81 | = | 77 | + | 5 | + | (-2.6) | + | 1.6 |
| 1 | 3 | 92 | = | 77 | + | 5 | + | 9.0 | + | 1.0 |
| 2 | 1 | 76 | = | 77 | + | 4.3 | + | (-6.4) | + | 1.1 |
| 2 | 2 | 79 | = | 77 | + | 4.3 | + | (-2.6) | + | 0.3 |
| 2 | 3 | 89 | = | 77 | + | 4.3 | + | 9.0 | + | (-1.3) |
| 3 | 1 | 75 | = | 77 | + | 2.3 | + | (-6.4) | + | 2.1 |
| 3 | 2 | 76 | = | 77 | + | 2.3 | + | (-2.6) | + | (-0.7) |
| 3 | 3 | 87 | = | 77 | + | 2.3 | + | 9.0 | + | (-1.3) |
| 4 | 1 | 74 | = | 77 | + | 3.3 | + | (-6.4) | + | 0.1 |
| 4 | 2 | 77 | = | 77 | + | 3.3 | + | (-2.6) | + | (-0.7) |
| 4 | 3 | 90 | = | 77 | + | 3.3 | + | 9.0 | + | 0.7 |
| 5 | 1 | 76 | = | 77 | + | 1.3 | + | (-6.4) | + | 4.1 |
| 5 | 2 | 71 | = | 77 | + | 1.3 | + | (-2.6) | + | (-4.7) |
| 5 | 3 | 88 | = | 77 | + | 1.3 | + | 9.0 | + | 0.7 |
| 6 | 1 | 73 | = | 77 | + | 0.7 | + | (-6.4) | + | 1.7 |
| 6 | 2 | 75 | = | 77 | + | 0.7 | + | (-2.6) | + | (-0.1) |
| 6 | 3 | 85 | = | 77 | + | 0.7 | + | 9.0 | + | (-1.7) |
| 7 | 1 | 68 | = | 77 | + | (-1) | + | (-6.4) | + | (-1.6) |
| 7 | 2 | 72 | = | 77 | + | (-1) | + | (-2.6) | + | (-1.4) |
| 7 | 3 | 88 | = | 77 | + | (-1) | + | 9.0 | + | 3.0 |
| 8 | 1 | 64 | = | 77 | + | (-4.7) | + | (-6.4) | + | (-1.9) |
| 8 | 2 | 71 | = | 77 | + | (-4.7) | + | (-2.6) | + | 1.3 |
| 8 | 3 | 82 | = | 77 | + | (-4.7) | + | (9.0 | + | 0.7 |
| 9 | 1 | 65 | = | 77 | + | (-4) | + | (-6.4) | + | (-1.6) |
| 9 | 2 | 73 | = | 77 | + | (-4) | + | (-2.6) | + | 2.6 |
| 9 | 3 | 81 | = | 77 | + | (-4) | + | (9.0) | + | (-1.0) |
| 10 | 1 | 62 | = | 77 | + | (-7.3) | + | (-6.4) | + | (-1.3) |
| 10 | 2 | 69 | = | 77 | + | (-7.3) | + | (-2.6) | + | 1.9 |
| 10 | 3 | 78 | = | 77 | + | (-7.3) | + | 9.0 | + | (-0.7) |

Table 7.3: ANOVA Table for Auditor Example

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Method | 2 | 1287.2 | 643.6 | 114.2 | <.0001 |
| Block | 9 | 465.3 | 51.7 | 9.17 | <.0001 |
| Error | 18 | 101.5 | 5.6 | | |
| Total (Corrected) | 29 | 1854.0 | | | |

Table 7.4: Tukey Pairwise Comparisons of Methods

| Method | Method | Mean Difference | Std.Error | DF | t Value | P-value | LCL | UCL |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | -3.89 | 1.06 | 18 | -3.58 | 0.0058 | -6.51 | -1.09 |
| 1 | 3 | -15.40 | 1.06 | 18 | -14.50 | <.0001 | -18.11 | -12.69 |
| 2 | 3 | -11.60 | 1.06 | 18 | -10.92 | <.0001 | -14.31 | -8.89 |

$$\begin{aligned} &= 77 + (-6.4) + 2.4 \\ &= 77 + (-6.4) + (5) + (2.4 - 5) \\ &= 77 + (-6.4) + (5) + (-2.6) \end{aligned}$$

The training 1 method effect of $(-6.4)$ is the same for the two models. The error under a completely randomized design is 2.4, which is the difference between 73 and the mean proficiency measure for training method 1. Subtracting 5 (block 1 effect) from 2.4 (and adding 5in the equation) results in an error of $(-2.6)$, which is the error for the block design.

In the model for the completely randomized design there is no term for block effect and thus block effect is contained in the error. In the block design block effects can be estimated and then "removed" (subtracted) from the errors that would normally occur with the completely randomized design, in theory, making the errors for the block design smaller.

### 7.3.3 Summary of F test for Treatment Effects in the Randomized Complete Block Design

The null and alternative hypotheses for the test of treatment effects for $b$ treatments are

$$H_o : \tau_1 = \tau_2 = \ldots = \tau_b = 0$$

or equivalently,

$$H_o : \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot b}$$

The alternative hypothesis is

$$H_a : \text{ not all } {\tau_j}'s = 0$$

or equivalently,

$$H_a : \text{ not all } {\mu_{\cdot j}}'s \text{ are equal}$$

The test statistic is

$$F = \frac{MSTR}{MSE} = \frac{SSTR/(b-1)}{SSE/(a-1)(b-1))}$$

where

$$SSTR = a\sum_{j=1}^{b}\hat{\tau}_j^2 = a\sum_{j=1}^{b}(\overline{y}_{.j} - \overline{y}_{..})^2$$

and

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}[e_{ij}]^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}[y_{ij} - (\overline{y}_{..} + \hat{\rho}_i + \hat{\tau}_j)]^2$$

Note that SSTR is a comparison of the marginal means of the response variable for the treatments which is only appropriate if there is no interaction between the blocking variable and the factor of interest.

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common variance holds, then $MSTR/MSE$ has the "F" probability distribution with $\nu_1 = (b-1)$ numerator degrees of freedom and $\nu_2 = (a-1)(b-1)$ denominator degrees of freedom.

At a significance level of $\alpha$ the null hypothesis would be rejected if the observed value of the test statistic, $F_o$, is larger than $F_{\alpha;(b-1),(a-1)(b-1)}$, the upper $\alpha$ probability point from the appropriate F distribution or equivalently if P-value $\leq \alpha$, where P-value $= P[F \geq F_o]$. Probability points for $\alpha = 0.05$ and $\alpha = 0.01$ are given in Tables A.7 and A.8, respectively. P-values can only be approximated using Table A.7 or A.8. More precise P-values can be obtained using statistical computing software.

A test of block effects is also available although it is not usually of main interest. It is expected that there are block effects since the purpose of blocking is to reduce experimental error associated with the presumed relationship between the blocking variable and the response.

The null and alternative hypotheses for the test of block effects for the $a$ blocks are

$$H_o : \rho_1 = \rho_2 = \ldots = \rho_a = 0$$

or equivalently,

$$H_o : \mu_{1.} = \mu_{2.} = \ldots = \mu_{a.}$$

The alternative hypothesis is

$$H_a : \text{ not all } \rho_i{}'s = 0$$

or equivalently,

$$H_a : \text{ not all } \mu_{i.}{}'s \text{ are equal}$$

The test statistic is

$$F = \frac{MSBL}{MSE} = \frac{SSBL/(a-1)}{SSE/(a-1)(b-1))}$$

where

$$SSBL = b\sum_{i=1}^{a}\hat{\rho}_i^2 = b\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2$$

and

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}[e_{ij}]^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}[y_{ij} - (\overline{y}_{..} + \hat{\rho}_i + \hat{\tau}_j)]^2$$

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common variance holds, then $MSBL/MSE$ has the "F" probability distribution with $\nu_1 = (a-1)$ numerator degrees of freedom and $\nu_2 = (a-1)(b-1)$ denominator degrees of freedom.

At a significance level of $\alpha$ the null hypothesis would be rejected if the observed value of the test statistic, $F_o$, is larger than $F_{\alpha;(a-1),(a-1)(b-1)}$, the upper $\alpha$ probability point from the appropriate F distribution or equivalently if P-value $\leq \alpha$, where P-value $= P[F \geq F_o]$. Probability points for $\alpha = 0.05$ and $\alpha = 0.01$ are given in Tables A.7 and A.8, respectively. P-values can only be approximated using Table A.7 or A.8. More precise P-values can be obtained using statistical computing software.

The expected values of the various mean squares can be shown to be

$$\begin{aligned} E[MSE] &= \sigma^2 && (7.4)\\ E[MSBL] &= \sigma^2 + \frac{b\sum_{i=1}^{a}\rho_i^2}{a-1} && (7.5)\\ E[MSTR] &= \sigma^2 + \frac{a\sum_{j=1}^{b}\tau_j^2}{b-1} && (7.6) \end{aligned}$$

These are the same expected mean squares as those in Chapter 6 under the assumption that there is no interaction between the two factors A and B and that the number of replications is 1 for every treatment combination.

### 7.3.4 Pairwise Comparisons Using the Tukey-Kramer Procedure

The general form of endpoints for the Tukey-Kramer confidence interval for the difference of two treatment means $\mu_{.j} - \mu_{.j'}$, adapted from Chapter 6, is

$$\overline{y}_{.j} - \overline{y}_{.j'} \pm \frac{q_{\alpha;\nu,b}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{a} + \frac{1}{a}}$$

where $\overline{y}_{.j}$ and $\overline{y}_{.j'}$ refer, respectively, to the marginal means of $y$ for levels $j$ and $j'$ of the factor of interest B. The number of levels of the blocking factor $a$ in the denominators refer to the number of observations used to calculate the marginal means. The degrees of freedom $\nu$ refers to degrees of freedom associated with MSE. The set of intervals have an experiment-wise confidence level of $1-\alpha$. The percentile $q_{\alpha;\nu,b}$ can be found in Table A.5 or A.6 with $t$ in the table equalling the number of levels, $b$, of the factor of interest.

## 7.4 More on Block Designs and Analysis

The following are additional items worth noting regarding the RCBD.

a. **Missing Observations**
It was assumed that there was a complete set of observations on the treatments for each block, that is no missing observations on the response. If there are missing observations then the analysis is more complex. For a discussion of this situation see [14], page 287, or [24], page 324.

b. **Time Order and Interference Effects**
In the reusing type of blocking where individuals are given all treatments at different time slots there is the possibility of order effects, that is a tendency for earlier observations to be higher (or lower) than later observations regardless of the treatments. Randomizing the order of the treatments will at least tend to balance out such effects among the treatments and make the errors approximately independent.

   Also for these types of studies the effect of a treatment in one time slot could in theory carry over to the next time slot where the individual is given another treatment. A solution to this potential **"carry-over effect"** is to allow a sufficient amount of time after one treatment application before beginning another treatment.

c. **Loss of Degrees of Freedom** The degrees of freedom or amount of information for estimating error is smaller for the randomized complete block design as compared to the completely randomized design with the same number of observations in the treatment groups. However $MSE$ with the block design may be smaller than that for the completely randomized design ultimately resulting in more precise comparisons of the treatments, which is one reason for using a block design.

d. **Multiple Observations on the Experimental Unit**
In some situations there may be multiple observations on the response variable for each of the experimental units within a block, that is there is subsampling. For example in an animal health study pens of animals may be the experimental units and pens are blocked according to similar weight distribution within the pen. Observations on individual animals within the pen constitutes subsampling. The analysis of the randomized complete block design with subsampling is considered in Chapter 10.

e. **Multiple Experimental Units within Blocks**
In some situations it may be desirable to have more than one replicate of each treatment in a block. The design is then called a generalized randomized complete block design and is considered in Section 7.6.

f. **Random Block Effects**

It has been assumed up to this point that the levels of the blocks and block effects are **"fixed."** This means that the levels of the blocking variable are of particular interest and if the experiment were to be repeated the same levels would be used. Suppose in the auditor example that the actual elapsed times since graduation from college were, respectively, 1, 2, ..., 10 years, corresponding to the 10 blocks. Fixed block effects means that if the experiment were repeated, while the subjects might be different, the elapsed times would be the same.

In many blocking scenarios blocks might represent a random sample of blocks from some population of blocks. The particular blocks used are not of particular interest and interest is in generalization to the population of all blocks. If the experimenter had the opportunity to repeat the experiment he or she would select a different random sample of blocks. In this situation blocks and block effects are **"random"** rather than **"fixed,"**. The analysis would assume that the block effects are values of random variables analagous to the assumption of errors being random variables. Good examples are where blocks are persons who have been randomly selected from some population. Inference would normally not be in those particular subjects but in the population of all subjects from which the random sample was selected.

Unlike for the fixed block effects case, two models may be considered with random block effects, one with random blocks without interaction (additive model) and another with random block and interaction effects. If block effects are considered random then any interaction effects would be also random. Kutner and others ([15], pages 1060 - 1065) gives a detailed discussion of the two models. Standard errors associated with estimates of treatments would not be the same as for the fixed block effects model. However comparisons of treatments use the same F statistic (MSTR and MSE) and degrees of freedom as used for the fixed effects model. The decision to use the additive model or the model with interaction could be based on plots. Regardless of which model is used comparisons of treatments under a model with random block effects are for the population of blocks.

## 7.5 Paired Samples t test Revisited

Recall from Chapter 3 that when there are two treatments and observations are paired (blocks of size 2) the paired samples t test can be used for analysis. Differences between the response values for the two treatments are calculated within each block and then a single sample *t* test is performed on the differences (See Section 3.2). It will be illustrated in this section that a two-sided comparison using the paired samples *t* test is equivalent to analysis with ANOVA for a block design using the F test. An example follows.

Table 7.5: ANOVA Table for Word Recall Example

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| List | 1 | 0.075 | 0.075 | 0.01 | 0.9110 |
| Student | 59 | 1044.425 | 17.702 | 2.97 | <.0001 |
| Error | 59 | 351.425 | 5.956 | | |
| Total (Corrected) | 119 | 1395.925 | | | |

The equivalence will be illustrated with the word recall example from Chapter 3. In Section 3.2 the paired samples t test was used to compare the numbers of words recalled by students after studying two lists of words. One list consisted of 25 concrete words and the other list 25 abstract words. The mean of the differences in numbers of words recalled was $\overline{d} = 0.05$ with $s_d = 3.45$. The observed value of the $t$ test statistic was 0.11 and P-value = 0.9910 based on $df = 59$. Thus there was no evidence that recall of these words depended upon list.

In the context of this chapter student is the blocking variable and list (A or B) is the factor of interest. The response variable is number of words recalled. An ANOVA table for this example is given in Table 7.5.

Note that the P-value for the effect of list, 0.9110 is identical to the P-value obtained from the paired samples t test. What is not obvious is the relation between the values of the t statistic and the F statistic. It can be shown that the square of the t statistic is equal to the F ratio. Note here that $0.11^2 = 0.01$. Also note that the degrees of freedom for the t statistic, 59, is the same as the degrees of freedom for error with the ANOVA. This equivalence between the two methods only holds for the two sided test.

## 7.6 Generalized Randomized Complete Block Design

### 7.6.1 Replication Within Blocks and Reasons for Use

In some studies the blocking factor is such that the number of subjects available or that can be recruited within each block exceeds the number of treatments. Thus the treatments may be assigned to more than one unit within each block. The blocking factor in these situations would typically be natural groupings (Type A) of the subjects such as gender, or type of pain. Kutner, Nachtsheim, Neter, and Li ([15], page 907) provide an example of a study comparing the effects of distraction (low distraction, high distraction) on the time required to complete a task, using eight men and eight women. Gender is the blocking variable regarded as fixed. Within each gender the eight persons are assigned at random to the two types of distraction, with four person assigned to each treatment. Presumably gender is an extraneous variable and blocking was used to reduce the size of error. The effects of gender might also be of interest in and of itself.

One benefit to using multiple units within each block for each treatment is the ability to study possible interaction between the blocking variable and the treatment factor. Recall that with the usual block design it is assumed that there is no interaction which may or may not be the case. This assumption however allows an estimate of experimental error. So in the distraction example above interaction between gender and distraction type may be studied.

Suppose that as before there are $a$ fixed blocks but now there are $sb$ experimental units available for each block. The $sb$ experimental units are assigned at random to the $b$ treatments with $s$ units assigned to each treatment. In the distraction example, there are $a = 2$ gender blocks and 8 subjects within each gender block are assigned at random to the two distraction levels with 4 subjects per distraction level.

The design is referred to as a **Generalized Randomized Complete Block Design or GCBD**. The analysis can be conducted in a manner similar to the analysis of the two factor completely randomized design of Chapter 6. The F tests for block, treatment, and interaction effects compare mean squares associated with these effects to MSE as in Chapter 6. The model and assumptions for the GCBD with fixed block effects are given in the next section.

In other situations researchers may have some control over how blocks are formed, such as sorting based on physical characteristics and thus the number of units within each block could be increased. While this would allow for the testing of interaction the units within a block might be less homogeneous than compared to the standard block design and defeat the purpose of blocking. Thus the generalized randomized block design might not be recommended unless the no interaction assumption is not reasonable and there is a desire to test for it.

Recall in the auditor example (Example 7.1) that there were three auditors in each block equalling the number of treatments. The auditors were blocked according to the number of years since graduation. It is not hard to imagine a situation where the homogeneity could still be retained with multiple units within a block and thus allowing for the formal study of interaction. For example there may be enough auditors available for the study so that blocks could be formed with 6 auditors per block, and for each set of 6 auditors, the elapsed amount of time since graduation from college being the same number of years.

In a study a student wanted to compare the amount of juice that could be obtained from fruit when rolling the fruit as compared to not rolling before squeezing. She recognized that the amount of juice from fruit varies with the type of fruit and considered type as a blocking variable. She also wanted to draw more general conclusions that would apply to many types of fruit. It would perhaps not be difficult to obtain more than two pieces of fruit of the same type and roughly of the same size by weighing them at purchase. She obtained 6 oranges, 6 lemons, and 6 limes from the supermarket. She weighed the fruit so that they were similar in size within each type. The 6 oranges, 6 lemons, and 6 limes were then randomly assigned to the two treatments, rolling and not rolling. The 6 pieces of each type were processed on the same day. Thus there are 3 experimental units within each combination of fruit type and treatment.

When experimental units are physical entities it might be possible to retain the homogeneity within blocks with multiple units as in the fruit example. In other cases it might not be possible to achieve homogeneity and then it is advisable to use single replicates within blocks rather than stacking blocks with multiple units that are not homogeneous.

### 7.6.2 Model and Analysis: Fixed Block Effects

A generalized randomized complete block (GRCBD) design is a block design where the number of experimental units within each block is a multiple of the number of treatments. The model below assumes fixed blocks.

The model is

$$y_{ijk} \quad = \quad \mu_{\cdot\cdot} + \rho_i + \tau_j + (\rho\tau)_{ij} + \epsilon_{ijk} \tag{7.7}$$

where

- $\mu_{\cdot\cdot}$ represents the true grand mean of the response
- $\rho_i$ $(i = 1, ..., a)$ represents the true fixed effect of $i^{th}$ level of the blocking variable
- $\tau_j$ $(j = 1, ..., b)$ represents the true effect of the $j^{th}$ level of the factor of interest
- $(\rho\tau_{ij})$ represents the interaction effect between the $i^{th}$ level of the blocking variable and the $j^{th}$ level of the factor of interest, and
- $\epsilon_{ijk}$ represents as usual the effects of extraneous variables on the $k^{th}$ observation of $y$ at the $ij$ combination of the blocking variable and factor of interest.

The model assumes that the errors are independent normal random variables each with mean 0 and common variance $\sigma^2$.

The formulas for the various sums of square are the same as that for the two factor completely randomized factorial design discussed in Chapter 6.

### 7.6.3 GRCBD Analysis: Random Block Effects

In some experiments it is reasonable to assume that the blocks are random with interest not being in the particular blocks used in the study but in a population of blocks from which the blocks were selected. When the blocks are random and multiple experimental units are possible, that is a generalized block design, then the F test for treatment effects using MSE as the denominator is no longer appropriate. The appropriate F test for treatment effects compares mean square for treatments with mean square for interaction (See Lawson, [16], page 185). Conclusions regarding treatment differences, interaction or no interaction, are general conclusions about the treatments in the population of blocks.

Lawson [16], page 128 gives an example of the comparison of three golf tee heights on distance when driving a golf ball. The blocks are 9 golfers and experimental units are different drives by a golfer. The standard block design would have three drives for each golfer, one drive for each tee height, in a random order. Lawson argues that driving a golf ball does not take a lot of exertion so that homogeneity of drives would still exist if each golfer drove 15 balls, 5 per tee height, instead of 3 balls, 1 per tee height, that is a GRCBD. Golfers are considered random since conclusions are to be drawn about the population of golfers rather than just the particular golfers used in the study.

The model and analysis for the GRCBD with random block effects can be can be found in Lawson ([16], page 185)

### 7.6.4 Subsampling in a RCBD - Not a GRCBD

In some studies there are multiple observations at each block by treatment combination, but the multiple observations do not correspond to different experimental units, but are observations on measurement units of the same experiment unit.

A researcher compared the heights of cupcakes using three different levels of baking powder (low, medium, high). Three batches of cupcake batter were prepared using the same recipe except that the three differed in the level of baking powder used. Six cupcakes were made from each of the batches and put on a tray. The three trays were placed in the oven on randomly located shelves. After baking the cupcakes height was measured for each. This process of cupcake batter preparation and oven run was repeated for 10 different oven runs of the same oven.

Oven run is a blocking factor which might be regarded as random. The treatments are the levels of baking powder. The experimental units within each oven run are the three batches of cupcake mix, that is three experimental units per block. The number of measurements of cupcake height for each experimental unit is 6 per batch for a total of 18 measurements on height for each oven run. Thus there multiple observations per baking powder treatment but these do not correspond to multiple experimental units per treatment. There is only one experimental unit per treatment per block so the GRCBD analysis does not apply. Mean height could be calculated for each set of 6 cupcakes at each oven run by baking powder combination and the resulting means analyzed with the RCBD.

## 7.7 Factorial Treatment Structure in a RCBD

In some studies there is a single blocking variable with two factors in a factorial treatment structure. In the basic design there is only one replicate of each treatment combination in each block.

Brianna Itkin, Jenna Clough, and Derek Wilus (Fall, 2014) studied the effects of boiling status of water (boiled, not boiled) used to make ice cubes and shape of ice cubes (semi-circular, cube, cylindrical) on the amount of time for a tray of ice to melt. Six ice trays, one for each combination of boiling status and shape were filled with the same amount of water and then placed into the freezer. When the water was frozen the six trays were placed upside down on a wire mesh and the cubes allowed to melt. The response variable was the amount of time (in minutes and seconds) that the entire tray of cubes took to melt. This process was repeated on each of 5 days.

The blocking variable is Day with 5 levels. The treatments are combinations of the two factors of interest, boiling status and shape of cube.

The analysis of the RCBD with two factors will not be considered in this text. The interested reader may refer to Kutner, Nachtsheim, Neter, and Li ( [15], page 909 ) or Kuehl ( [14], page 289 ).

## 7.8 Two Blocking Variables – Latin Square Design

Consider the earlier example involving the comparison of the four brands of tires using the 16 tire positions of 4 cars. Suppose car is regarded as the blocking variable and the 4 brands are assigned completely at random to the 4 tire positions within each car. An example of a resulting randomization is the following:

| | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Car 1 | A | C | B | D |
| Car 2 | C | B | A | D |
| Car 3 | A | B | D | C |
| Car 4 | A | D | C | B |

With the above design the effect of wheel position is part of experimental error. When randomly assigning tire brands within each car, it is possible that brand A gets put mostly on the Left Front wheel. This might bias the comparison between brands if location is an extraneous variable.

An alternative design in this experiment would be a block design which consists of two blocking variables: car and wheel position, in general called the row and column blocking variables. In this design all four brands are used for each car and simultaneously all four brands are used at each wheel position as in the following diagram.

| | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Car 1 | A | B | C | D |
| Car 2 | B | C | D | A |
| Car 3 | C | D | A | B |
| Car 4 | D | A | B | C |

Since each wheel position is exposed to all four brands we will be able to estimate in an unbiased fashion the true effects of wheel position and remove this effect from experimental error.

The **Latin Square** design given above is called the **Standard Latin Square Design.** In the standard Latin Square Design the letters A, B, C, ... used to denote the treatments are written in the first row of the row blocking variable and then the remaining rows are obtained by shifting the letters to the left once.

Randomization of treatments for the Latin Square Design is achieved as follows

a. Start with the standard Latin Square Design

b. Randomly permute/arrange the rows: For example after randomly permuting the rows of the standard Latin Square design we may have the following:

| | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Car 1 | D | A | B | C |
| Car 2 | B | C | D | A |
| Car 3 | A | B | C | D |
| Car 4 | C | D | A | B |

Note that the first row of the square was the old fourth row and so on.

c. Randomly permute/arrange the columns of this square, for example

| | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Car 1 | B | C | A | D |
| Car 2 | D | A | C | B |
| Car 3 | C | D | B | A |
| Car 4 | A | B | D | C |

Note that the first column is the old third column and so on.

d. Now randomly assign the treatments to the letters. For example suppose the actual tire brands are Firestone, Goodyear, Goodrich, and UniRoyal. Put these four names on slips of paper and then pull one out at a time. The first one gets assigned to the letter A, the $2^{nd}$ to the letter B, and so on.

The general properties of the Latin Square Design are as follows:

a. There are two blocking variables, in general a row blocking variable and a column blocking variable.

b. The number of levels of the row blocking variable is equal to the number of levels of the column blocking variable, which is equal to the number of treatments.

c. Latin Square Designs can be repeated with additional experimental units to obtain more replications of the treatments.

### 7.8.1 Another Example

Cobb ([4], page 247) describes the following experiment. A study compares recall of words for different learning/recall environments. The subjects in the study are 4 members of a diving club. The treatments are dry/dry, dry/wet, wet/dry, and wet/wet. For example, dry/wet means the word list was studied while the diver was on land and was recalled when the diver was in the water.

This experiment could be carried out using a randomized complete block design with the one blocking variable being subject. Each subject receives all four treatments using different word lists for the treatments. The word lists are randomly assigned to the four treatments to ensure that the treatments are not always assigned to the same word list and to ensure that the effects of word list are random. The treatment/word list is randomly assigned to time slots so that effects of time period is random. In this design the effects of word list is part of experimental error.

The experiment could also be conducted as a Latin Square design. Not only does each person get all four treatments but each list is exposed to all four treatments in the following manner.

| | List 1 | List 2 | List 3 | List 4 |
|---|---|---|---|---|
| Diver 1 | dry/dry | dry/wet | wet/dry | wet/wet |
| Diver 2 | dry/wet | wet/dry | wet/wet | dry/dry |
| Diver 3 | wet/dry | wet/wet | dry/dry | dry/wet |
| Diver 4 | wet/wet | dry/dry | dry/wet | wet/dry |

The above design is a standard Latin Square Design. An alternative design can be obtained by the previously described randomization process. With the Latin Square design, there is a balance in that each list is used with all treatments and thus comparison of list averages would be unbiased estimates of list effect. Thus we can remove list effect from experimental error.

### 7.8.2 Model and Analysis for Latin Square Design

The model for the Latin Square Design is

$$y_{ijk} = \mu_{...} + \rho_i + \kappa_j + \tau_k + \epsilon_{ijk} \tag{7.8}$$

where

- $\mu_{...}$ represents the true grand mean
- $\rho_i$ $(i = 1, ..., r)$ represents the true main effect of the $i^{th}$ level of the row blocking variable
- $\kappa_j$ $(j = 1, ..., r)$ represents the true main effect of the $j^{th}$ level of the column blocking variable
- $\tau_k$ $(k = 1, ..., r)$ represents the true main effect of the $k^{th}$ level of the factor of interest

Table 7.6: ANOVA Table for Latin Square Design

| Source of Variation | df | SS | MS | F | EMS |
|---|---|---|---|---|---|
| Row | r - 1 | SSROW | MSROW | $MSROW/MSE$ | $\sigma^2 + Q_1$ |
| Column | r - 1 | SSCOL | MSCOL | $MSCOL/MSE$ | $\sigma^2 + Q_2$ |
| Treatment | r -1 | SSTR | MSTR | $MSTR/MSE$ | $\sigma^2 + Q_3$ |
| Error | (r-1)(r-2) | SSE | MSE | | $\sigma^2$ |
| Total | $r^2 - 1$ | SSTOTC | | | |

- $\epsilon_{ijk}$ represents as usual the experimental error, the effects of extraneous variables on the observation of $y$ at the $ijk$ combination of the blocking variables and factor of interest.

It is assumed that the errors are values of independent normal random variables each with mean 0 and variance $\sigma^2$. The model also assumes that there is no interaction between the two blocking variables and the factor of interest.

The ANOVA table is derived in a manner similar to how the ANOVA table is derived for other designs. The observed responses can be partitioned into parts representing the grand mean, the effect of the particular level of the row blocking variable, the effect of the particular level of the column blocking variable, the effect of the particular level of the factor of interest, and experimental error. The sum of squares of the deviations of the observed responses from the grand mean can be partitioned into sums describing variability in the effects of the row blocking variable, the column blocking variable, the factor of interest, and the experimental unit effects (error).

$$SSTOT_C = SSROW + SSCOL + SSTR + SSE$$

The general form of the ANOVA table is given in Table 7.6

In Table 7.6 the sums of squares for the various effects are given without formulas. We will rely on computer software to calculate these. Also the values $Q_1$, $Q_2$, and $Q_3$ in the EMS column are, respectively, functions of the true row block main effects, true column block main effects, and true treatment main effects, all of which are zero if the corresponding effects are 0. The column $EMS$ gives the expected or population average mean squares.

Let us consider testing for treatment effects. Under the null hypothesis, the expected mean square for treatments would be identical to the expected mean square for error. Thus under the null hypothesis we would expect the ratio $MSTR/MSE$ to be approximately 1. If the alternative hypothesis is true, then the expected mean square for the treatment factor would be larger than $MSE$. In this case we would expect the ratio $MSTR/MSE$ to be larger than 1. The test statistic for testing for treatment effects is the F ratio $MSTR/MSE$. Assuming the model assumptions hold and the null hypothesis of no treatment effects is true, then the ratio $MSTR/MSE$ has an $F$ probability distribution with numerator degrees of freedom $\nu_1 = r - 1$ and denominator degrees of freedom $\nu_2 = (r - 1)(r - 2)$. We will rely on computer software to calculate the

Table 7.7: Means for Job Satisfaction Scores

| Group | Mean |
|---|---|
| A: 4-day week, day shift | $\overline{y}_A = 14.8$ |
| B: 4-day week, evening shift | $\overline{y}_B = 13.0$ |
| C: 5-day week, day shift | $\overline{y}_C = 9.5$ |
| D: 5-day week, evening shift | $\overline{y}_D = 6.25$ |

$F$ ratio and obtain a P-value for hypothesis testing and for obtaining multiple comparisons of the treatments.

To illustrate the ideas consider the following example taken from Weber and Skillings ([29], page 406).

**Example 7.2** *Four work schedules are compared to see which leads to the best job satisfaction for a group of technicians. The four work schedules studied were:*

1. *A: 4-day week, day shift*
2. *B: 4-day week, evening shift*
3. *C: 5-day week, day shift*
4. *D: 5-day week, evening shift*

*Four technicians served as one blocking variable and week was the other blocking variable. The results are given in the table below.*

| | Week | | | |
|---|---|---|---|---|
| Technician | 1 | 2 | 3 | 4 |
| 1 | 18 (B) | 7 (D) | 13 (A) | 10 (C) |
| 2 | 8 (C) | 15 (A) | 6 (D) | 11 (B) |
| 3 | 18 (A) | 10 (C) | 10 (B) | 5 (D) |
| 4 | 7 (D) | 13 (B) | 10 (C) | 13 (A) |

Table 7.7 provides the marginal means of the job satisfaction scores for the four work schedules. The ANOVA table is given in Table 7.8.

The Tukey-Kramer method of multiple comparisons is used to make pairwise comparisons of the four work schedules. Tukey-Kramer adjusted P-values and simultaneous 95% confidence intervals are given in Table 7.9. The A-C, A-D, and B-D work schedule pairwise comparisons are significant at the 0.05 experimentwise level of significance.

Table 7.8: ANOVA Table for Work Schedule Example

| Source of Variation | df | SS | MS | F | Pvalue |
|---|---|---|---|---|---|
| Technician | 3 | 8.25 | 2.75 | 0.60 | 0.6382 |
| Week | 3 | 24.75 | 8.25 | 1.80 | 0.2473 |
| Work Schedule | 3 | 171.25 | 57.08 | 12.45 | 0.0055 |
| Error | 6 | 27.5 | 4.58 | | |
| Corrected Total | 15 | 231.75 | | | |

Table 7.9: Tukey Pairwise Comparisons of Work Schedules

| Schedule | Schedule | Mean Difference | Std.Error | DF | t Value | P-value | LCL | UCL |
|---|---|---|---|---|---|---|---|---|
| A | B | 1.75 | 1.06 | 6 | 1.16 | 0.6724 | -3.49 | 6.99 |
| A | C | 5.25 | 1.06 | 6 | 3.47 | 0.0496 | 0.01 | 10.49 |
| A | D | 8.50 | 1.06 | 6 | 5.61 | 0.0055 | 3.26 | 13.74 |
| B | C | 3.50 | 1.06 | 6 | 2.31 | 0.1971 | -1.74 | 8.74 |
| B | D | 6.75 | 1.06 | 6 | 4.46 | 0.0168 | 1.51 | 11.99 |
| C | D | 3.25 | 1.06 | 6 | 2.15 | 0.2399 | -1.99 | 8.49 |

# Problems for Chapter 7

7.1* This example is based on a study from Cobb( [4], page 300). Each of 8 premature infants was monitored during sleep during two six-hour periods under one of two conditions: 1) sleeping on a regular bassinet mattress (control) and 2) sleeping on a waterbed. The response variable was the number of interruptions in breathing per hour of sleep during the 6-hour period. The data are given in the table below.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Waterbed | 0.89 | 0.77 | 0.00 | 0.65 | 0.88 | 1.36 | 1.22 | 0.30 |
| Control | 1.36 | 1.66 | 0.11 | 1.44 | 1.63 | 1.52 | 1.53 | 0.48 |

a. The above design is a block design. What are the experimental units? What are the blocks?

b. Explain how the above experiment could have been conducted using a completely randomized design.

7.2* Hicks ([10], page 120) describes a study designed to compare ash content in coal as reported by two laboratories. Each of 10 coal samples were split in half and the halves assigned at random to the two laboratories for analysis. Ash contents reported by the two laboratories are given in the table below.

| Sample | Lab 1 | Lab 2 |
|---|---|---|
| 1 | 5.47 | 5.13 |
| 2 | 5.31 | 5.46 |
| 3 | 5.46 | 5.54 |
| 4 | 5.55 | 5.54 |
| 5 | 5.93 | 6.00 |
| 6 | 5.97 | 5.99 |
| 7 | 6.32 | 6.43 |
| 8 | 6.09 | 6.13 |
| 9 | 5.87 | 5.87 |
| 10 | 5.58 | 5.60 |

a. What type of blocking is this? What are the blocks?

b. What are the experimental units?

c. Explain what the randomization would have been in a completely randomized design.

7.3* This example is taken from Walpole and Myers ([28]). Each of six subjects were given all three diets (see below) in a random order, each diet lasting 3 days.

| | |
|---|---|
| Diet 1: | mixed fat and carbohydrates |
| Diet 2: | high fat |
| Diet 3: | high carbohydrates |

At the end of the 3 day period on a diet the subject was put on a treadmill and time to exhaustion, in seconds, was measured. The times for each subject and diet are given below.

| | Diet | | |
|---|---|---|---|
| Subject | 1 | 2 | 3 |
| 1 | 84 | 91 | 122 |
| 2 | 35 | 48 | 53 |
| 3 | 91 | 71 | 110 |
| 4 | 57 | 45 | 71 |
| 5 | 56 | 61 | 91 |
| 6 | 45 | 61 | 122 |

a. What is the response variable? What are the treatments? What are the experimental units?

b. What is the blocking variable and what extraneous variable is the blocking intended to control?

c. Explain how the above experiment could be carried out using a completely randomized design.

7.4* Ruth Lawlor and Richard Miller (Fall,2002) wanted to determine if color (or the chemicals associated with the colors) were related to the burning rate of candles. Eight-inch candles of four colors: blue, tan, purple, and white were used. The response variable was the amount of time (in minutes) that it took a candle to burn down 3 inches from the top. Four candles, one of each color, were burned on each day and this was repeated over 7 days. On a particular day the one candle of each color was randomly selected from a pool of available candles. The order in which the candles were lit was random and the candles were placed in random positions on a table. The burning times (to the nearest minute) are given in the following table:

| | Color of Candle | | | |
|---|---|---|---|---|
| Replication/Day | Tan | Blue | Purple | White |
| 1 | 201 | 217 | 184 | 167 |
| 2 | 213 | 206 | 158 | 227 |
| 3 | 183 | 116 | 273 | 273 |
| 4 | 300 | 174 | 277 | 271 |
| 5 | 299 | 190 | 228 | 237 |
| 6 | 196 | 159 | 199 | 208 |
| 7 | 259 | 227 | 243 | 262 |

a. This is a block design. What are the blocks.

b. What are the experimental units?

c. What are some extraneous variables that are being controlled by the randomization in each block?

d. Using the same number of observations per color explain how this experiment would be carried out in a completely randomized design.

e. Give the population effects model for the data and describe the terms in the model in context. Give the assumptions associated with the error terms.

f. Construct an ANOVA table. Use it to determine if there are significant differences in burn time among the colors using a 0.05 level of significance. If there are significant differences use the Tukey-Kramer multiple comparison procedure with experiment-wise significance level of 0.05 to rank the colors on burn time.

7.5* Morris ([23], p. 52) describes an experiment where each of four groups of laying hens (each group with 48 birds) were given four different diets over several weeks. The diets varied according to the concentrations of molasses (0, 70, 140, and 210 g/kg). The purpose of the experiment was to determine the effects that the molasses concentrations had on weights of eggs produced by the birds. The table gives the mean egg weight by the group in a period. The The letters A, B, C, and D represent, respectively, the levels 0, 70, 140, and 210 g/kg.

| | Period | | | |
|---|---|---|---|---|
| Group | 1 | 2 | 3 | 4 |
| 1 | 53.5 (D) | 55.4(A) | 55.1 (B) | 53.6 (C) |
| 2 | 56.1 (B) | 54.8(C) | 53.9 (D) | 55.0 (A) |
| 3 | 53.8 (C) | 54.1(D) | 55.2 (A) | 52.9 (B) |
| 4 | 53.1 (A) | 54.4(B) | 53.0 (C) | 51.1 (D) |

a. The design is a Latin Square design. What are the blocking variables?

b. What are the experimental units?

c. Give the population effects model for the data and describe the terms in the model. Give the assumptions associated with the error terms.

d. Use appropriate software to obtain an ANOVA table for the data. Use information from the ANOVA table to determine if there is a significant difference in mean egg weights among the diets. Use a significance level of $\alpha = 0.05$.

e. If your test in part (d) was significant obtain Tukey-Kramer pairwise confidence intervals to determine which diets differ with regard to egg weight. Use an experiment-wise confidence level of 95%.

7.6* Tire wear for the experiment described in Section 7.1 is given below.

| | Position | | | |
|---|---|---|---|---|
| Car | 1 | 2 | 3 | 4 |
| 1 | 30 (B) | 36 (A) | 25 (C) | 22 (D) |
| 2 | 24 (D) | 34 (C) | 18 (A) | 15 (B) |
| 3 | 35 (A) | 30 (D) | 15 (B) | 28 (C) |
| 4 | 32 (C) | 24 (B) | 13 (D) | 14 (A) |

a. Use appropriate software to obtain an analysis of variance table for the data. Use information from the ANOVA table to determine if there is a significant difference in tire wear among the brands. (Use $\alpha = 0.05$.)

b. If your test in part (a) was significant obtain Tukey-Kramer pairwise confidence intervals to determine which brands differ with regard to treadwear. Use an experiment-wise confidence level of 95%.

7.7* In the article "Dose response effects of a caffeine-containing energy drink on muscle performance: a repeated measures design"(*Journal of the International Society of Sports Nutrition* [2012], 9:21) researchers compared the effects of three levels of caffeine (0, 1, 3, mg per kg body weight) in a 100 mL energy drink of the same brand on resting metabolic rate, heart rate and blood arterial pressure for one hour after subjects consumed the drink. Each of twelve subjects consumed all drinks with the three different doses of caffeine on three different days, the order of which was randomly determined.

The design of the experiment is a one factor randomized complete block design.

a. Identify the blocks, treatments, and experimental units.

b. What type of blocking is this? Explain. What is the reason for the randomization of the order of the caffeine doses?

c. The means with standard deviations for the heart rates for the twelve subjects at the three different dose levels of caffeine is given in the table below.

| $MeanHeartRate \pm SD$ | | |
|---|---|---|
| 0 mg/kg | 1 mg/kg | 3 mg/kg |
| $57 \pm 7$ | $59 \pm 8$ | $61 \pm 8$ |

i. Calculate MSTR for an ANOVA.

ii. Is it possible to determine MSE? If yes, then do so. If not possible then explain why, that is what information would still be needed?

7.8 Casella ([2], page 130) describes an experiment conducted by a poultry scientist to investigate the effects of three diets with differing levels of protein (L = low, M = medium, H = High) on the amount of food intake of leghorn chickens. Ninety chickens were randomly assigned to 9 cages, with 10 chickens per cage. Because of space limitations the 9 cages were stacked 3 high and 3 deep with space between the cages. Differing heights of cages was important because of environmental temperature variation which affects food intake. Depth was important because there was only one light source at the front of the cages. A Latin Square design was used to assigned the three treatments to the cages. The food intakes along with the diets assigned to the cages are given in the table below.

| | Height of Cage | | |
|---|---|---|---|
| Depth of Cage | Bottom Row | Middle Row | Top Row |
| Front Stack | (96) (M) | (81)(H) | (106) (L) |
| Middle Stack | (94) (H) | (116)(L) | (114)(M) |
| Back Stack | (100)(L) | (91)(M) | (89) (H) |

a. The design is a Latin Square design. What are the blocking variables?

b. What are the experimental units?

c. Give the population effects model for the data and describe the terms in the model. Give the assumptions associated with the error terms.

d. Use appropriate software to obtain an analysis of variance table for the data. Use information from the ANOVA table to determine if there is a significant difference in pen total food intake among the diets. Use a significance level of $\alpha = 0.05$.

e. If your test in part (d) was significant obtain Tukey-Kramer pairwise confidence intervals to determine which diets differ with regard to food intake. Use an experiment-wise confidence level of 95%.

7.9* Ashlee Schenk and Tanya Blackburn (Fall 2004) conducted an experiment to study the effects of type of container (styrofoam, glass, plastic) and type of liquid (cola, water, juice) on the melting time (seconds) of an ice cube placed into the liquid. Five replications for each combination of container and liquid type were conducted. Because of time constraints the experiment had to be conducted over the course of 5 days with one replication of the 9 combinations on each day. On a given day melting of cubes was done one cube at a time as follows. A combination of type of container and type of liquid was randomly selected. One cup of the liquid was poured into the selected container and then a randomly selected ice cube was added. The amount of time (mins) for the cube to melt was recorded and then this process was repeated until all combinations were done on the given day. This process was repeated on four other days.

a. What are the two factors in the study? What is the response variable?
b. What are the treatments in the study? How many treatment are there?
c. What are the experimental units in the study? How many total experimental units are there?
d. Give two extraneous variables and explain how the effects of these are being controlled.
e. The design is a block design. What are the blocks?

7.10 Consider the infant sleep interruption rate data from Problem 7.1.

a. Use statistical software to give an ANOVA table corresponding to the block model described in Chapter 7. The table should be similar to Table 7.5. Is there evidence of a difference in sleep interruption rates between the two types of mattresses? Use a significance level of 0.05. Give an appropriate null hypothesis, value of test statistic, and P-value. Draw a conclusion.
b. Use statistical software to derive results for a two-sided paired samples t test to compare interruption rates for the two mattresses. Give an appropriate null hypothesis, value of test statistic, and P-value. Draw a conclusion.
c. Compare the results from the two testing procedures in parts (a) and (b) in terms of values of test statistic, P-value, degrees of freedom.

7.11* Devore and Peck ([11], page 635) described a study comparing electricity usage (in KWh) for four different residential air-conditioning systems being proposed for use in tract homes. Twenty homes selected for the experiment were grouped into 5 blocks according to floor space, type of insulation, directional orientation and type of roof and exterior with 4 homes in each block. Within each of the blocks the four homes were randomly assigned to the four systems. The electricity usage during a 1-month period is recorded below.

| | System | | | |
|---|---|---|---|---|
| Block | 1 | 2 | 3 | 4 |
| 1 | 116 | 171 | 138 | 144 |
| 2 | 118 | 131 | 131 | 141 |
| 3 | 97 | 105 | 115 | 115 |
| 4 | 101 | 107 | 93 | 93 |
| 5 | 115 | 129 | 110 | 99 |

a. What is the factor of interest?
b. What are the experimental units?
c. What is the response variable?
d. What is the purpose of blocking the homes on the given variables?

e. Using the same number of observations per treatment explain how this experiment would be carried out using a completely randomized design.

f. Give the population effects model for the data and describe the terms in the model in context. Give the assumptions associated with the errors.

g. Use statistical software to construct an analysis of variable table. Use it to determine if there are significant differences in electricity usage among the four systems. Use a significance level of 0.05. If there are significant differences use the Tukey-Kramer multiple comparison procedure with experiment-wise significance level of 0.05 to rank the systems.

h. Suppose that the data had been incorrectly analyzed as a completely randomized design, disregarding the blocking. What would the incorrect error be for the home getting System 1 with electricity usage of 116? Determine the correct error for this home (if the correct block design analysis had been used) using the incorrect error obtained above and the block 1 effect.

7.12 For each of the following, describe the design that was used: a) one or two factor completely randomized, b) one or two factor block design with one blocking variable, c) Latin Square Design. In each case give the factor or factors of interest, blocking variable(s)with Type (A,B,C), experimental units, and response variable(s). Indicate whether sub-sampling was present and what are the subsamples are.

a. In the article "A prospective study of patients with chronic back pain randomized to group exercise, physiotherapy or osteopathy" (*Physiotherapy 94 (2008) 21-28* ) researchers compared three therapy regimes for chronic back pain. Two hundred and thirty-nine subjects were randomly assigned to group exercise (80), physiotherapy (80), or osteopathy (79) with 32, 59, and 63 completing the therapy. Main outcomes of interest were the Oswestry Disability Index (ODI), EuroQol-5D, shuttle walking test and patients' responses to pain and treatment.

b. Michael Pulomena and Nick Zurlo (Spring 2014) conducted an experiment to compare three brands of dish soap (Gain, Dawn, Ajax) for removing three different types of condiments (ketchup, mustard, barbecue sauce) after the condiments had been applied to a plate and then dried in a microwave for 1 minute and 45 seconds. The response variable was the amount of time (seconds) to completely "clean" the plate. The experiment was conducted as follows. A dish soap and type of condiment was randomly selected. The selected condiment was applied to a clean plate, put in the microwave to dry, and then the plate was cleaned with the selected dish soap. This procedure was repeated for a total of 5 replications per combination of dish soap and type of condiment.

c. A study was conducted to compare two different types of cups, styrofoam and paper, on the ability of the cup to retain heat. Twenty-four ounces of water was heated in a pot to 160° Fahrenheit. The water was then poured into two cups, with half (12 ounces) going into a styrofoam cup and the other half into a paper cup. The water in the two cups was allowed to cool for 10 minutes and the temperature of the water in each cup measured. This process was repeated for 9 other pots of 24 ounces of water and two new cups.

d. A study was conducted to compare a caffeinated energy drink with a placebo drink on jump performance of adolescent basketball players. Each of 30 players was tested on two different days, on one day 60 minutes after ingesting the caffeinated drink and the other day 60 minutes after ingesting a placebo drink. The order of the drinks was random. At each testing session each player performed a series of 15 jumps using a force platform, being instructed to jump as high as possible on each jump. Height (cm) was measured for each jump.

# Chapter 8

# Checking Assumptions of Errors

## 8.1 Assumptions

In the models that we have considered there has always been an error term $\epsilon$ representing the effects of extraneous variables which have not been explicitly accounted for in the design. For example in the model for the one factor completely randomized design,

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij} \end{aligned} \tag{8.1}$$

where $\epsilon_{ij} = y_{ij} - \mu_i$ is the deviation of the $j^{th}$ observation from the $i^{th}$ treatment mean and represents the effects of extraneous variables.

The validity of the P-values associated with the F tests and testing of contrasts depends on the errors satisfying certain statistical assumptions. The assumptions are given below in the order in which they should be assessed.

- The errors are statistically independent.
- The error random variables have the same variance/standard deviation.
- The errors are values of normal random variables.

While $y_{ij}$ is observed the error $\epsilon_{ij}$ is not actually observed since it depends upon the mean of $y_{ij}$, $\mu_i$. So how do we check the assumptions of the errors if we don't actually observe them. We do this by estimating the errors and then using the estimates of the errors to check the assumptions.

The estimates of the errors depend upon the model for the data.

### 8.1.1 Residuals for One Factor Completely Randomized Model

For the one factor completely randomized design of Chapter 4 the error is

$$\epsilon_{ij} = y_{ij} - \mu_i$$

The obvious estimate of $\epsilon_{ij}$ is obtained by substituting for $\mu_i$, the estimate $\overline{y}_{i.}$, and obtaining the estimate of the error, $e_{ij}$, called the **residual**, that is

$$e_{ij} = y_{ij} - \overline{y}_{i.}$$

Note that $e_{ij}$ is not the same as $\epsilon_{ij}$. The estimated error $e_{ij}$ can be calculated – the true error $\epsilon_{ij}$ cannot. Since $\overline{y}_{i.}$ can be thought of as a prediction for the mean of treatment level $i$ we can think of the estimate of the error as

$$\begin{aligned} e_{ij} &= y_{ij} - \overline{y}_{i.} \\ &= \textit{observed} - \textit{predicted} \qquad (8.2) \end{aligned}$$

### 8.1.2 Residuals for Two Factor Completely Randomized Design

The means model for the two factor completely randomized design from Chapter 6 is:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk} \qquad (8.3)$$

and thus the true error is

$$\epsilon_{ijk} = y_{ijk} - \mu_{ij} \qquad (8.4)$$

Estimating $\mu_{ij}$, the population treatment mean, with the sample treatment mean or predicted $Y$, $\overline{y}_{ij.}$ then the estimate $e_{ijk}$ of $\epsilon_{ijk}$ is

$$e_{ijk} = y_{ijk} - \overline{y}_{ij.} \qquad (8.5)$$

A residual for the two factor completely randomized model is the difference between an observed value of the response and the mean of the response in the respective treatment group.

### 8.1.3 Residuals for One Factor Randomized Complete Block Design

The model for the one factor block design (only one blocking variable) with only one replication per combination of block and treatment is from Chapter 7,

$$y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$$

If we solve this equation for $\epsilon_{ij}$ we get

$$\epsilon_{ij} = y_{ij} - (\mu_{..} + \rho_i + \tau_j)$$

This suggests estimating the error term $\epsilon_{ij}$ with

$$\begin{aligned} e_{ij} &= observed - predicted \\ e_{ij} &= y_{ij} - (\overline{y}_{..} + \hat{\rho}_i + \hat{\tau}_j) \end{aligned} \qquad (8.6)$$

where $\hat{\rho}_i = \overline{y}_{i.} - \overline{y}_{..}$, the block effect for the $i^{th}$ block, and $\hat{\tau}_j = \overline{y}_{.j} - \overline{y}_{..}$, the treatment effect for the $j^{th}$ treatment.

## 8.2 Checking for Independence

The $\epsilon_{ij}$'s are independent if the value of one tells you nothing about the value of another error. The most likely cause of lack of independence or dependence are experimental units close in time or space.

If an experiment is conducted through time or arranged in some spatial pattern, then a plot of the estimated errors against time order or spatial arrangement will indicate whether or not the errors are independent or dependent. If the errors are independent then in this plot the estimated errors should be randomly scattered about 0 with no discernible pattern.

**Example 8.1** *This example is taken from Dean and Voss ([6], page 27). The purpose of the study was to compare the life times of four different kinds of batteries:*

- *Battery Type 1: alkaline, name brand*
- *Battery Type 2: alkaline, store brand*
- *Battery Type 3: heavy duty, name brand*
- *Battery Type 4: heavy duty, store brand*

*There were 4 replications per treatment. The experimental units were time slots for battery testings with* 16 *batteries,* 4 *of each kind, being tested in a random order.*

*The data with the time order and the residuals are given in Table 8.1:*

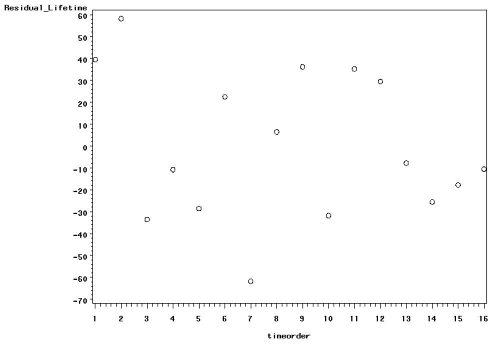
Table 8.1: Lifetime Data

| Battery Type | Lifetime(minutes) | Time Order | Residual |
|---|---|---|---|
| 1 | 602 | 1 | 39.5 |
| 2 | 863 | 2 | 58.25 |
| 1 | 529 | 3 | -33.5 |
| 4 | 235 | 4 | -10.75 |
| 1 | 534 | 5 | -28.5 |
| 1 | 585 | 6 | 22.5 |
| 2 | 743 | 7 | -61.75 |
| 3 | 232 | 8 | 6.5 |
| 4 | 282 | 9 | 36.25 |
| 2 | 773 | 10 | -31.75 |
| 2 | 840 | 11 | 35.25 |
| 3 | 255 | 12 | 29.5 |
| 4 | 238 | 13 | -7.75 |
| 3 | 200 | 14 | -25.5 |
| 4 | 228 | 15 | -17.75 |
| 3 | 215 | 16 | -10.5 |

Table 8.2: Means and Standard deviations for Lifetime Data

| Battery Type | Mean | Standard Deviation |
|---|---|---|
| 1 | 562.50 | 36.52 |
| 2 | 804.75 | 58.25 |
| 3 | 225.50 | 23.61 |
| 4 | 245.75 | 24.53 |

Figure 8.1: Scatterplot of Lifetime Residuals versus Time Order



Note from the Table 8.1 that a battery of type 1 was tested first, a battery of type 2 was tested next, and so on.

The sample means and standard deviations of lifetimes are given in Table 8.2.

The F test for overall differences in lifetime among the battery types was significant at the 0.05 level with $F = 217.53, P < 0.0001$.

The residual corresponding to the first observation would be $602 - 562.50 = 39.5$. The other residuals are in Table 8.1.

A plot of the residuals versus time order is given in Figure 8.1 Note that the residuals appear to be randomly scattered about 0 providing no evidence of dependence.

**Example 8.2** *An experiment (Dean and Voss [6], page 62)) involved an individual blowing up different colored balloons in a random order to compare inflation times. The colors of the balloons used were pink, yellow, orange, and blue. The purpose of the experiment was to see if color affected the amount of time required to blow up the balloons.*

*The inflation times with the time orders are given in Table 8.3*

The sample means and standard deviations for the inflation times are given in Table 8.4.

The F test is significant ($F = 3.85, P = 0.0200$) at the 0.05 level of significance indicating a differences somewhere among mean inflation times among the colors.

A plot of the residuals for inflation times versus time order of the observations is given in Figure 8.2

Note the negative linear relationship between the residuals and time order of the testing with residuals generally being positive for the early trials and negative for the later trials. Thus the inflation times were higher than predicted for the early trials and lower than predicted for the later trials regardless of color, indicating that the experimenter took less time to inflate the balloons as time progressed. One possible solution to this problem is to take account of

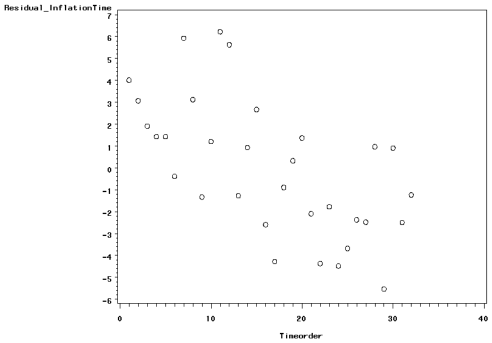
Table 8.3: Inflation Times Data

| Time Order | Color | Inflation Time (secs) |
|---|---|---|
| 1 | pink | 22.4 |
| 2 | orange | 24.6 |
| 3 | pink | 20.3 |
| 4 | blue | 19.8 |
| 5 | blue | 19.8 |
| 6 | yellow | 22.2 |
| 7 | yellow | 28.5 |
| 8 | yellow | 25.7 |
| 9 | orange | 20.2 |
| 10 | pink | 19.6 |
| 11 | yellow | 28.8 |
| 12 | blue | 24.0 |
| 13 | blue | 17.1 |
| 14 | blue | 19.3 |
| 15 | orange | 24.2 |
| 16 | pink | 15.8 |
| 17 | yellow | 18.3 |
| 18 | pink | 17.5 |
| 19 | blue | 18.7 |
| 20 | orange | 22.9 |
| 21 | pink | 16.3 |
| 22 | blue | 14.0 |
| 23 | blue | 16.6 |
| 24 | yellow | 18.1 |
| 25 | yellow | 18.9 |
| 26 | blue | 16.0 |
| 27 | yellow | 20.1 |
| 28 | orange | 22.5 |
| 29 | orange | 16.0 |
| 30 | pink | 19.3 |
| 31 | pink | 15.9 |
| 32 | orange | 20.3 |

Table 8.4: Means and Standard Deviations for Balloon Inflation Time Data

| Balloon Color | Mean | Standard Deviation |
|---|---|---|
| Blue | 18.4 | 2.9 |
| Orange | 21.5 | 3.0 |
| Pink | 18.4 | 2.4 |
| Yellow | 22.6 | 4.5 |

Figure 8.2: Scatterplot of Inflation Time Residuals versus Time Order



order in the statistical model. In theory we could expand our model to include not only color effects but time order as well. The resulting model is called an analysis of covariance model. This approach and other approaches to handling dependent errors will not be considered in this text.

Note how important the randomization was in this example. While there is a problem, the problem could have been even bigger. If the experimenter had not randomized the order of the colors and blown up balloons by color group, such as first all pink, followed by all orange, then yellow, then blue, then any color effects would have been confounded with order effects.

**Example 8.3** *This example uses the data from the two factor study in Problem 6.2 in Chapter 6. The factors are Shooting Distance from a target (Short, Medium, Long) and Hand (Right,Left) used to fire a Nerf bullet from a gun. The response variable is accuracy defined as the absolute distance from a target (to the nearest 1/8 inch).*

*The accuracies along with the time order, residuals, and predicted accuracies are provided in Table 8.5.*

A plot of the accuracies versus the combination of distance and hand is given in Figure 8.3

Treatment means and standard deviations for accuracy are provided in Table 8.6.

Notice that the predicted accuracies in Table 8.5 are just the treatment means as noted earlier. A residual is just the difference between an accuracy and a predicted accuracy or treatment mean. For example the residual of 2.050 associated with the accuracy of 3.375 at time order 1 in the Short,Right group is 3.375 minus the Short,Right mean of 1.325.

Figure 8.4 provides a plot of the residuals versus time order. Note that there is no evidence of a relationship of the residuals with time and thus the assumption of independent errors appears to be satisfied. The plot does indicate one accuracy being slightly outlying.

Table 8.5: Nerf Gun Data

| Distance | Hand | Accuracy | TimeOrder | Predicted | Residual |
|---|---|---|---|---|---|
| Short | Left | 0.000 | 3 | 0.825 | -0.825 |
| Short | Left | 1.500 | 7 | 0.825 | 0.675 |
| Short | Left | 0.000 | 13 | 0.825 | -0.825 |
| Short | Left | 0.625 | 15 | 0.825 | -0.200 |
| Short | Left | 2.000 | 19 | 0.825 | 1.175 |
| Short | Right | 3.375 | 1 | 1.325 | 2.050 |
| Short | Right | 0.375 | 10 | 1.325 | -0.950 |
| Short | Right | 2.125 | 16 | 1.325 | 0.800 |
| Short | Right | 0.250 | 24 | 1.325 | -1.075 |
| Short | Right | 0.500 | 29 | 1.325 | -0.825 |
| Medium | Left | 3.500 | 5 | 2.450 | 1.050 |
| Medium | Left | 3.250 | 9 | 2.450 | 0.800 |
| Medium | Left | 0.125 | 17 | 2.450 | -2.325 |
| Medium | Left | 3.250 | 21 | 2.450 | 0.800 |
| Medium | Left | 2.125 | 26 | 2.450 | -0.325 |
| Medium | Right | 1.000 | 2 | 2.950 | -1.950 |
| Medium | Right | 4.875 | 18 | 2.950 | 1.925 |
| Medium | Right | 1.000 | 20 | 2.950 | -1.950 |
| Medium | Right | 3.250 | 23 | 2.950 | 0.300 |
| Medium | Right | 4.625 | 28 | 2.950 | 1.675 |
| Long | Left | 13.250 | 4 | 8.975 | 4.275 |
| Long | Left | 7.000 | 6 | 8.975 | -1.975 |
| Long | Left | 8.125 | 8 | 8.975 | -0.850 |
| Long | Left | 7.750 | 11 | 8.975 | -1.225 |
| Long | Left | 8.750 | 25 | 8.975 | -0.225 |
| Long | Right | 3.125 | 12 | 6.225 | -3.100 |
| Long | Right | 1.125 | 14 | 6.225 | -5.100 |
| Long | Right | 14.375 | 22 | 6.225 | 8.150 |
| Long | Right | 3.375 | 27 | 6.225 | -2.850 |
| Long | Right | 9.125 | 30 | 6.225 | 2.900 |

Figure 8.3: Plot of Accuracy versus Treatment

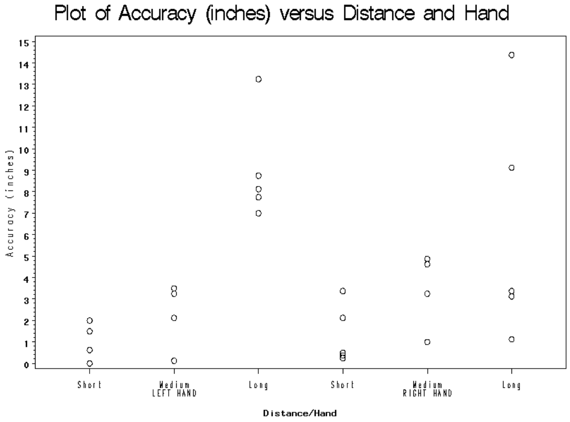
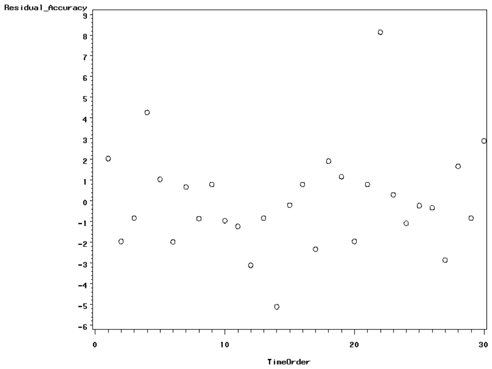


Table 8.6: Nerf Gun Data Means and Standard Deviations

| Distance | Hand | n | Mean | Standard Deviation |
|---|---|---|---|---|
| Short | Left | 5 | 0.825 | 0.900 |
| Short | Right | 5 | 1.325 | 1.377 |
| Medium | Left | 5 | 2.450 | 1.405 |
| Medium | Right | 5 | 2.950 | 1.885 |
| Long | Left | 5 | 8.975 | 2.472 |
| Long | Right | 5 | 6.225 | 5.445 |

Figure 8.4: Plot of Nerf Gun Residuals versus Time Order



## 8.3 Assessing the Assumption of Homogeneous Error Variances

It is assumed in the methods of analysis of variance that the variances of the errors are identical, in particular, equal to some common value, $\sigma^2$. This condition is called **homogeneity** of error variances. Thus we need to check to make sure this assumption holds true, at least approximately. The methods are somewhat robust to deviations from this assumption especially when treatment group sizes are identical. Violation of this assumption is called **heterogeneity** of error variance.

### 8.3.1 Methods for Checking the Assumption of Homogeneity of Error Variance

1. Compare standard deviations of the observations on the response variable for the different treatment groups. A rule of thumb is that the largest standard deviation should be no more than roughly 3 times the smallest standard deviation.
2. Plot the values of the response variable versus the treatments. The vertical spread in the points for the different treatments should be about the same. Recall that in a two-factor factorial treatment structure the treatments are combinations of the levels of the two factors.
3. Plot the residuals or estimated errors from the fitted model against predicted values and treatments. Variation in residuals should be similar across predicted values and treatments.

### 8.3.2 Checking Homogeneity of Variance for the Battery Example

Recall the battery example from Example 8.1 and the means and standard deviations from Table 8.2. A plot of the lifetimes versus battery types is given in Figure 8.5. Note that the vertical spread of the points corresponding to the lifetimes is similar for the four battery types.

The largest standard deviation is 58.25 and the smallest is 23.61. Thus the ratio of the largest to the smallest standard deviation is $58.25/23.61 = 2.47 < 3$.

A plot of the residuals from the model fit against the predicted lifetimes is given in Figure 8.6.

Note the evidence of slightly greater variability of the residuals for the largest predicted lifetime. Predicted lifetimes for this model are just treatment means of the lifetimes. The largest predicted lifetime corresponds to the battery type with the largest mean, which is battery type 2.

A plot of the residuals from the model fit against the treatments (battery types) is given in Figure 8.7. In this example and for the one factor model the plot of the residuals versus treatment is just the residual versus predicted plot with the vertical columns of points in perhaps a different order. Here again

Figure 8.5: Scatterplot of Lifetimes versus Battery Type

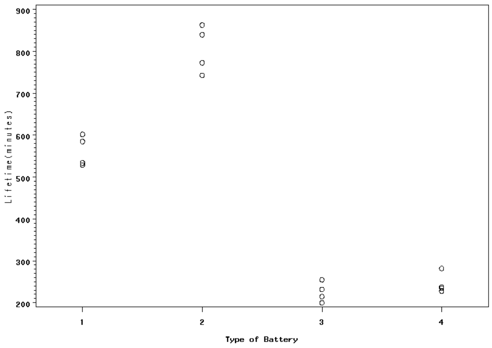


Figure 8.6: Scatterplot of Residuals versus Predicted Lifetimes

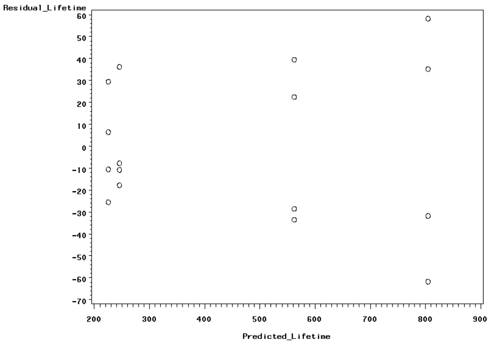
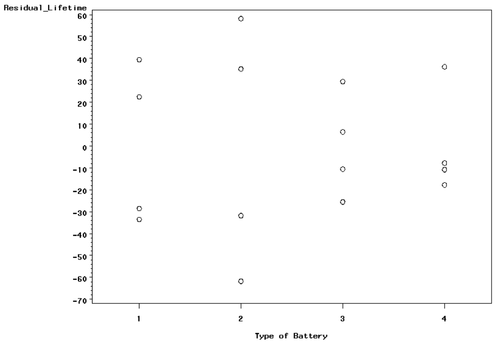


Figure 8.7: Scatterplot of Residuals versus Battery Type



we see that the residuals corresponding to battery type 2 having slightly more spread than the other battery types.

The slightly differing spreads among the lifetimes or residuals is of no practical concern. We will look at another example shortly where the spreads are quite different.

### 8.3.3 Checking Homogeneity of Variance for the Paper Towel Example

Recall the paper towel example from Chapter 6. There were two factors of interest in a completely randomized design. One factor was brand of paper towel with three levels: Coronet, Kleenex, and Scott. The other factor of interest was Liquid with three levels: Water, Dishwashing Detergent, and Vegetable Oil. There were 3 replications of each of the 9 treatment combinations of Brand and Liquid. A plot of the response variable amount of liquid absorbed (mL) versus treatment was given in Figure 6.3. Variation does not appear to differ much among the treatments; however there are only 3 replications per treatment combination.

The means and standard deviations for amount absorbed are given for the treatment combinations in Table 8.7.

The ratio of the largest to the smallest standard deviation is $3.51/0.58 = 6.05$, above our rule of thumb value of 3. Let's also look at a plot of the residuals versus predicted values and check to see if there is any evidence of a trend in spread with increasing predicted amount absorbed.

Figure 8.8 gives a plot of residuals from the fit of the complete two factor model versus predicted amount absorbed based on that model. For this model predicted amount absorbed is just the mean amount absorbed for a treatment combination of brand of paper towel and type of liquid. Note that there is no discernible trend in spread as predicted amount increases. Since the standard deviation ratio was not much larger than 3 (which could have occurred by chance

Table 8.7: Means and Standard Deviations for Amount of Liquid Absorbed

| Towel | Liquid | Mean | Standard Deviation |
|---|---|---|---|
| Coronet | Dishwashing Liquid | 16.67 | 2.08 |
| Coronet | Vegetable Oil | 25.33 | 3.51 |
| Coronet | Water | 23.33 | 2.31 |
| Kleenex | Dishwashing Liquid | 36.33 | 2.89 |
| Kleenex | Vegetable Oil | 41.67 | 3.06 |
| Kleenex | Water | 41.67 | 1.15 |
| Scott | Dishwashing Liquid | 20.67 | 0.58 |
| Scott | Vegetable Oil | 25.67 | 1.15 |
| Scott | Water | 26.00 | 1.00 |

Figure 8.8: Scatterplot of Residuals versus Predicted Amount Liquid Absorbed

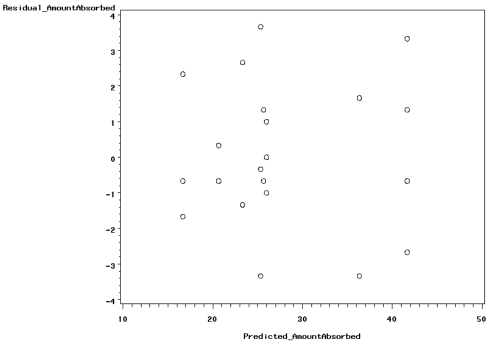
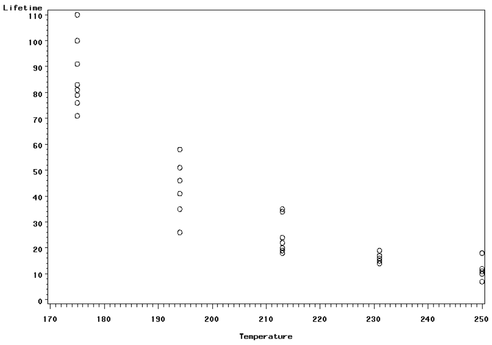


Table 8.8: Lifetimes of a Resin Under Temperature Stress

| Temperature (C) | | | | |
|---|---|---|---|---|
| 175 | 194 | 213 | 231 | 250 |
| 110 | 46 | 34 | 14 | 18 |
| 81 | 51 | 35 | 17 | 7 |
| 100 | 26 | 24 | 15 | 12 |
| 83 | 58 | 20 | 14 | 10 |
| 71 | 46 | 22 | 16 | 12 |
| 91 | 41 | 19 | 19 | 11 |
| 76 | 35 | 18 | 15 | |
| 79 | 46 | 24 | | |

Figure 8.9: Plot of Resin Lifetime versus Temperature



since group sizes was small) and since the residual plot showed no patterns, we will assume that the homogeneity assumption holds approximately.

### 8.3.4 Checking Homogeneity of Variance Data - Resin Example

This example is adapted from Oehlert ([24], page 32). The data given in Table 8.8 represents lifetime in hours of a resin which is used to encapsulate gold-aluminum bonds in integrated circuits when the resin was stressed at different temperatures.

A plot of the resin lifetimes versus temperature is given in Figure 8.9.

Note that not only the average lifetime but also variability in lifetime is affected by temperature violating the homogeneity of variance assumption.

Table 8.9 gives the means and standard deviations of the resin lifetimes at the different temperatures.

Table 8.9: Times to Failure: Means and Standard Deviations

| | Temperature (C) | | | | |
|---|---|---|---|---|---|
| | 175 | 194 | 213 | 231 | 250 |
| Mean | 86.4 | 43.6 | 24.5 | 15.7 | 11.7 |
| StDev | 13.1 | 9.8 | 6.5 | 1.8 | 3.6 |

Table 8.10: ANOVA Table for Resin Lifetime Data

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Temperatures | 4 | 28066.8 | 7016.7 | 99.42 | <.0001 |
| Error | 32 | 2258.5 | 70.6 | | |
| Total (Corrected) | 36 | 30325.3 | | | |

The ratio of the largest to the smallest standard deviation is $13.1/1.8 = 7.3$, somewhat larger than the rule of thumb value of 3.

An ANOVA table for the data is given in Table 8.10. There is strong evidence of a difference in lifetimes among the temperatures.

Figure 8.10 gives the residual plot of residuals versus predicted lifetimes. Note the tendency for the residuals to become more variable as predicted values, here temperature means, increase, creating a funneling effect. Again there is evidence that the error variances are not constant across temperatures.

Figure 8.10: Resin Lifetime Data: Residuals versus Predicted

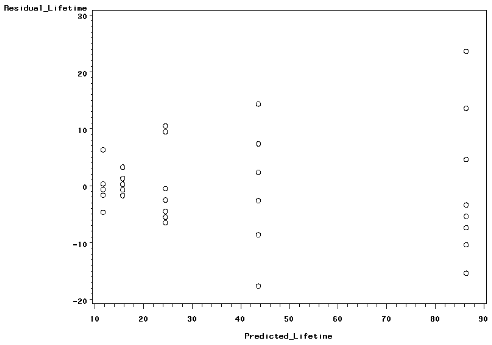
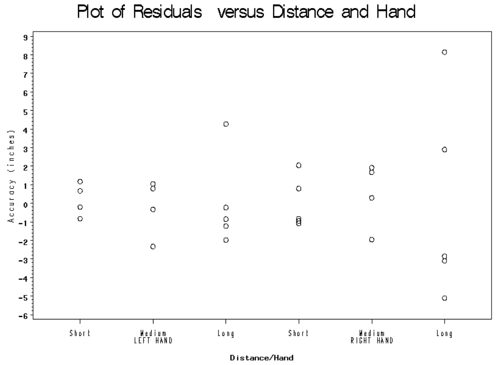


Figure 8.11: Plot of Nerf Gun Residuals versus Treatment



## 8.3.5 Checking Homogeneity of Variance for the Nerf Gun Example

A plot of the accuracies versus treatment was given in Section 8.2, Figure 8.3 Variation in accuracy appears to increase with distance regardless of the hand.

The means and standard deviations for the accuracies at the treatment combinations are given in Table 8.6. The ratio of the largest to the smallest standard deviation is 5.445/0.900 = 6.05, slightly above our rule of thumb value of 3. Let's also look at a plot of the residuals versus the treatments and also residuals versus predicted values and check to see if there is any evidence of a trend in spread with increasing predicted accuracy.

Figure 8.11 is a plot of the residuals versus the combination of distance and hand. This plot reflects what was seen in Figure 8.3, some evidence of greater variation in accuracies for the long distance.

Figure 8.12 is a plot of the residuals versus the predicted accuracies (treatment means). Note that there is a tendency for the variation in the residuals to increase with increasing predicted accuracy. So there is some evidence of heterogeneity in the error variances.

## 8.3.6 Checking Homogeneity of Variance for a Block Design Example

This example refers to the auditor example of Chapter 7 (Example 7.1, Section 7.3). The residuals are given in Table 7.2 in Chapter 7. Since there is only one replication per combination of block and method then a plot of proficiency measure versus treatment (combination of block and method) would not be informative for checking variation in estimated residuals across combinations. Alternative plots are the plotting of residuals against levels of the blocking factor and against levels of the treatment factor, here method of training. Figures 8.13 and 8.14 provide these plots.

Figure 8.15 provides a plot of the residuals versus predicted measures.

Figure 8.12: Plot of Nerf Gun Residuals versus Predicted Accuracies

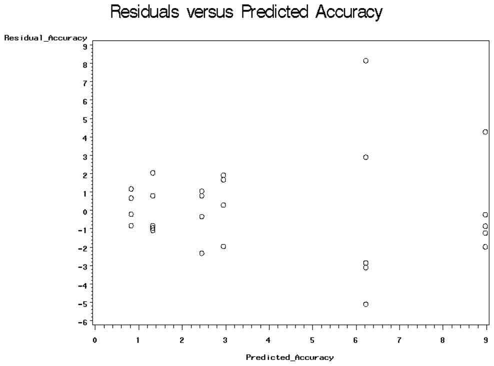


Figure 8.13: Plot of Auditor Example Residuals versus Block

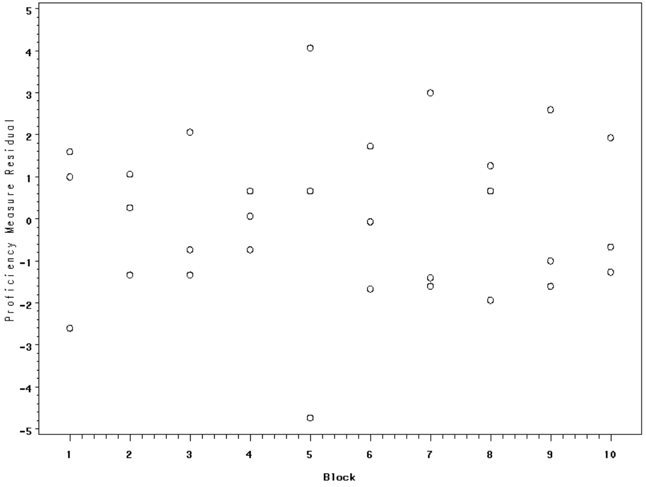


Figure 8.14: Plot of Auditor Example Residuals versus Method

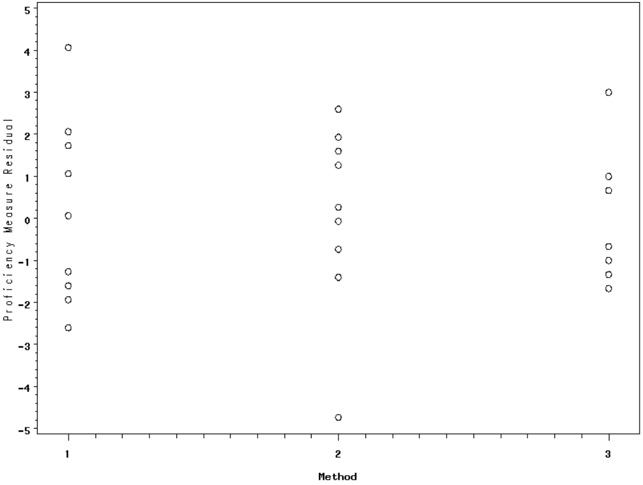
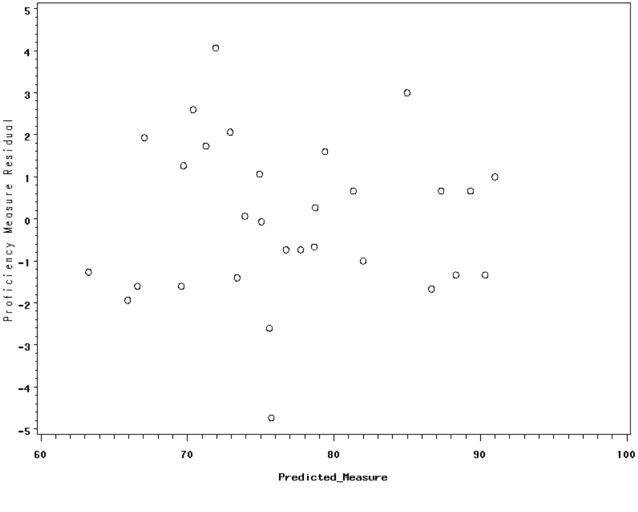


Figure 8.15: Plot of Auditor Example Residuals versus Predicted Measure



There is no evidence of extreme deviations from the assumption of homogeneous error variances. There does appear to be one mildly outlying measure.

## 8.4 Assessing the Assumption of Normality

Recall that if the assumption of constant error variance appears to be satisfied then the analyst should check the assumption of normality of the error terms. There are two graphical procedures that data analysts use to check the assumption of normality of the errors. A histogram or stem and leaf plot of the residuals can be viewed to check for an overall bell shaped distribution. While a histogram is a good start it is not sensitive to departures in the tails of the distributions and needs a relatively large sample size to give a good idea of the true shape of the distribution. Another tool that analysts use is the normal quantile-quantile or Q-Q plot.

Suppose that the residuals are ordered and represented as $r_1, r_2, ..., r_n$ where $n$ is the total number of residuals being investigated and $r_i$ represents the $i^{th}$ smallest residual. Associated with each $r_i$ is $z_i$, the “expected value” of the $i^{th}$ smallest value in a sample of size $n$ from a standard normal or $z$ distribution. For example if $n = 28$ and $i = 14$ then $z_{14}$ would refer to the expected $14th$ smallest z-score in a sample of $n = 28$ $z$ or standard normal scores, which you would expect to be about 0, since the $14^{th}$ smallest z-score in 28 is about halfway through all of the 28 z-scores in the sample. Similarly, if $i = 7$ then $z_7$ would refer to the expected $7^{th}$ smallest z-score in a sample of 28. Or $z_7$ would refer to that z-score for which about $7/28 = 1/4 = 0.25$ of z-scores are smaller. One could go to a standard normal table, such as Table A.1 and use for $z_7$ the 0.25 quantile from that distribution (z score with upper 0.75 area in Table A.1). Statistical programs will calculate the $z_i$ so we do not have to manually do these. Also most programs use a slightly different formula than what we used, $i/n$, to define the appropriate z quantile.

A normal quantile-quantile (Q-Q) plot is a plotting of the pairs $(r_i, z_i)$ in a Cartesian coordinate system. Thus a normal Q-Q plot is just a special kind of scatterplot. If the errors are truly normally distributed with the same variance then the normal probability plot should be roughly linear. If the errors are not normally distributed then the plot should exhibit some type of curvature.

Some examples of typical Q-Q plots are given in the following figures.

Figure 8.16 and Figure 8.17 gives a typical histogram and normal Q-Q plot when the error terms are truly normally distributed. Note the linear relationship between the residuals and the expected standard normal quantiles.

Figure 8.18 and Figure 8.19 give a typical histogram and normal Q-Q plot when the error terms have a “heavy tailed” distribution, that is the tails of the distribution are more spread out than that for a normal distribution.

Figure 8.20 and Figure 8.21 give a typical histogram and and normal Q-Q plot when the error terms have a symmetric “light tailed” distribution, that is the tails of the distribution are less spread out than that for a normal distribution.

Figure 8.16: Residual Histogram: Normal Distribution

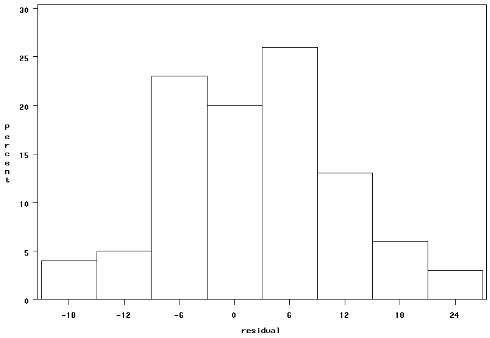


Figure 8.17: Residual QQPlot: Normal Distribution

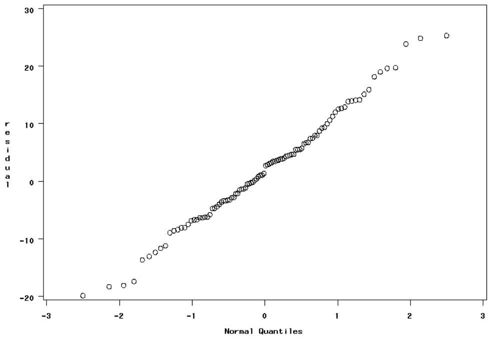


Figure 8.18: Residual Histogram: Heavy Tail Distribution

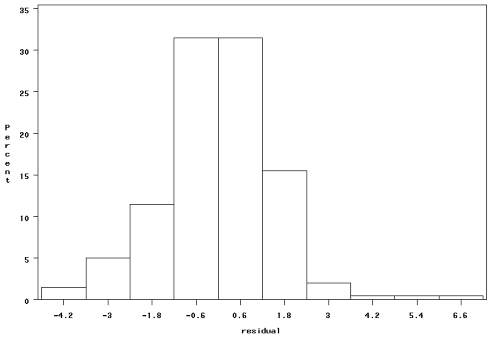


Figure 8.19: Residual QQPlot: Heavy Tail Distribution

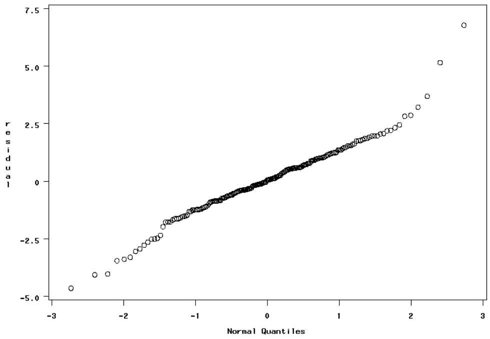


Figure 8.20: Residual Histogram: Light Tail Distribution

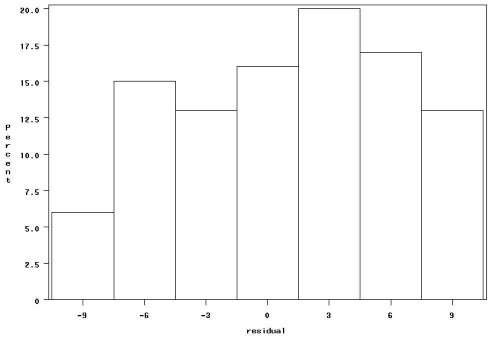


Figure 8.21: Residual Q-Q Plot: Light Tail Distribution

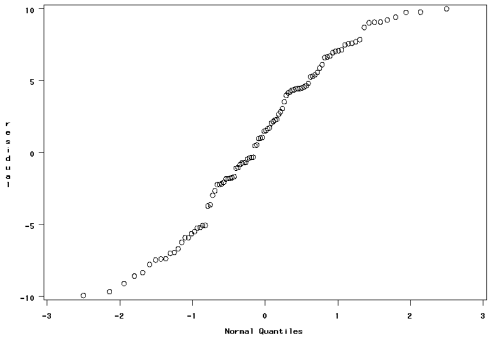


Figure 8.22: Residual Histogram: Right Tail Distribution

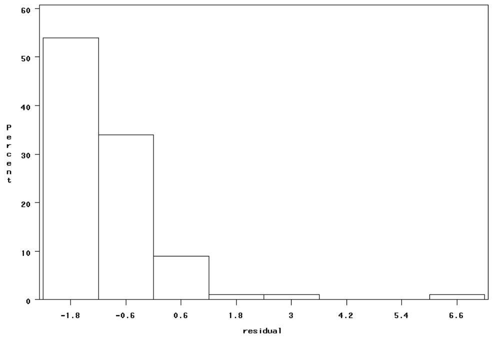


Figure 8.23: Residual Q- Q Plot: Right Tail Distribution

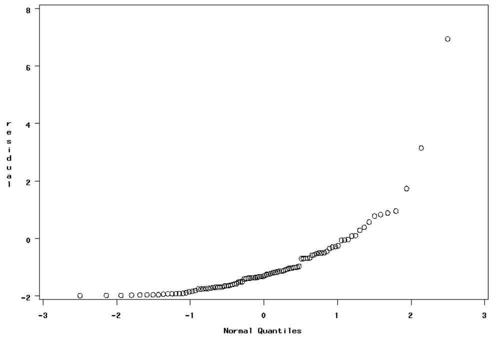


Figure 8.22 and Figure 8.23 give a typical histogram and normal Q-Q Plot when the error terms have an asymmetric "right skewed" distribution, that is the right tail of the distribution is more spread out than the left tail.

Figure 8.24 and Figure 8.25 give a typical histogram and normal Q-Q Plot when the error terms have an asymmetric "left skewed" distribution, that is the left tail of the distribution is more spread out than the right tail.

Note that the QQ plots are nonlinear when the shape of the histogram is not normal or bell-shaped and has various non-linear shapes.

### 8.4.1 Checking Normality of the Errors for the Battery Example

Figures 8.26 and 8.27 provides a histogram and Q-Q plot of the residuals for the battery lifetime data, respectively.

There are no major deviations from normality and so the assumption of normality of the model errors appears to be satisfied approximately. Thus all

Figure 8.24: Residual Histogram: Left Tail Distribution

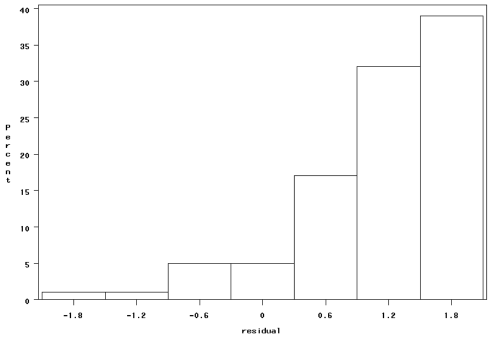


Figure 8.25: Residual Q-Q Plot: Left Tail Distribution

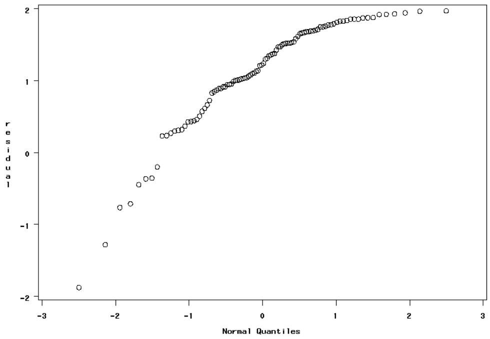


Figure 8.26: Histogram of Residuals for Battery Lifetime Data

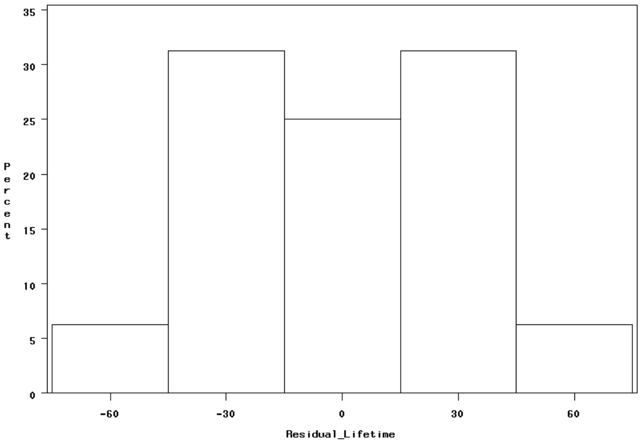
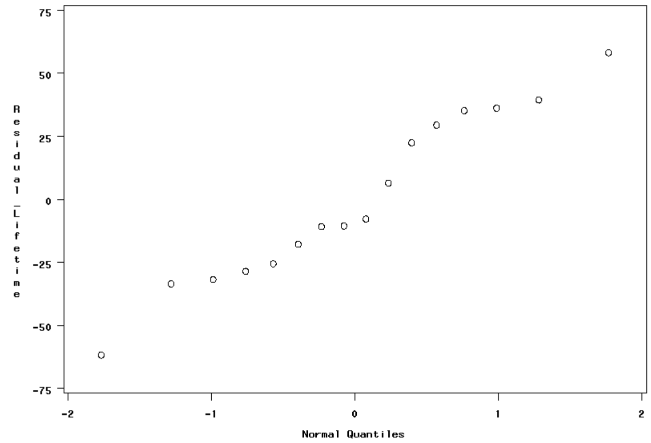


Figure 8.27: QQPlot of Residuals for Battery Lifetime Data



three assumptions appear to hold approximately for this data. See Example 8.1 for the check on independence and Section 8.3.2 for the check on homogeneity of error variances.

### 8.4.2 Checking Normality of the Errors for the Paper Towel Example

Figures 8.28 and 8.29 provide a histogram and QQplot of the residuals for the paper towel absorption data, respectively.

There are no major deviations from normality and so the assumption of normality of the model errors appears to be satisfied approximately. The assumption of homogeneity of error variances was checked earlier (see Section 8.3.3).

### 8.4.3 Checking Normality of the Errors for the Auditor Example

Figures 8.30 and 8.31 provide a histogram and Q-Q plot of the residuals for the auditor proficiency data, respectively. See Section 8.3.6.

There are no major deviations from normality and so the assumption of normality of the model errors appears to be satisfied approximately. Thus the assumptions of homogeneity of error variances and normality appear to be hold approximately for this data.

Figure 8.28: Histogram of Residuals for Paper Towel Absorption Data

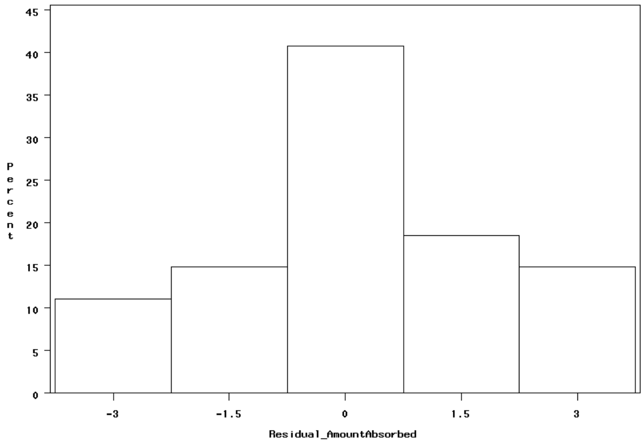


Figure 8.29: QQPlot of Residuals for Paper Towel Absorption Data

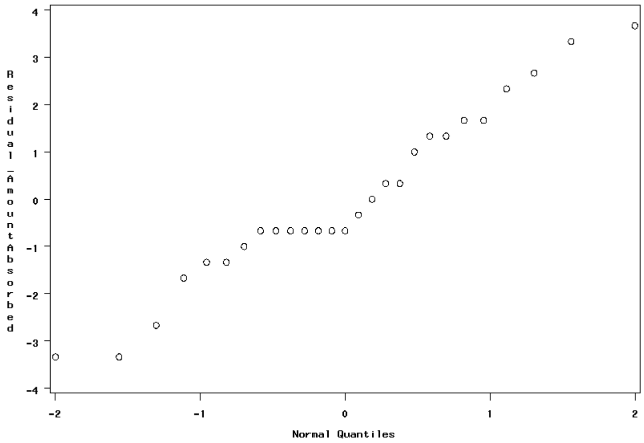


Figure 8.30: Histogram of Residuals for Auditor Proficiency Data

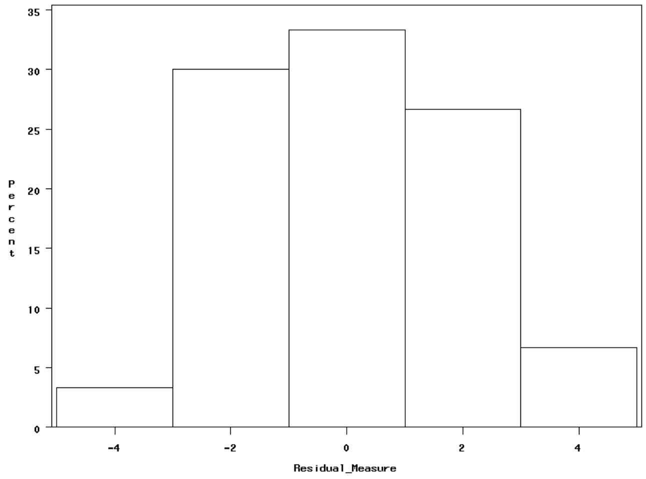
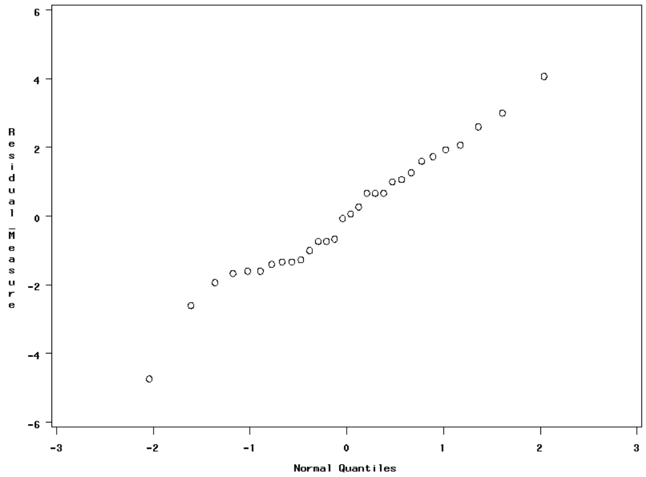
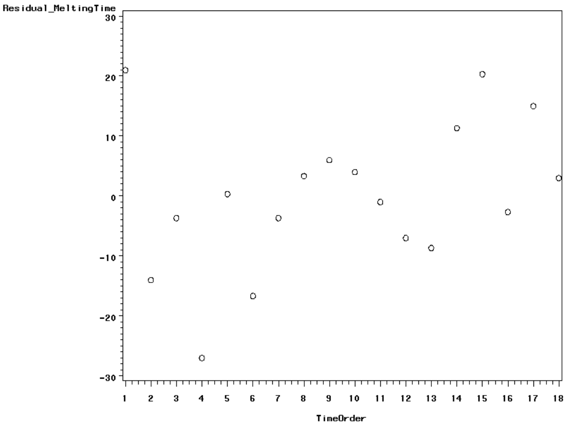


Figure 8.31: QQPlot of Residuals for Auditor Proficiency Data



# Problems for Chapter 8

8.1* Below is a plot of residuals versus time order for the melting butter example of Exercise 6.4. Do you think that the errors are independent? Explain.

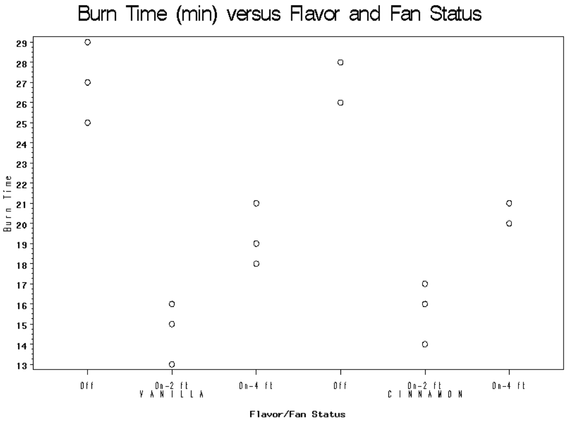


8.2* An incense burning experiment was run by David Gately (Fall 2008) to study the effects of fan status (off, 2 feet from incense, 4 feet from incense) and flavor of incense stick (vanilla, cinnamon) on the amount of time (to the nearest minute) it took the stick of incense to burn out. The experimental was a two-factor completely randomized design with experimental units being time slots. The following table gives the burn times and the time slots at which these burn times were obtained. The residuals and predicted values are left blank.

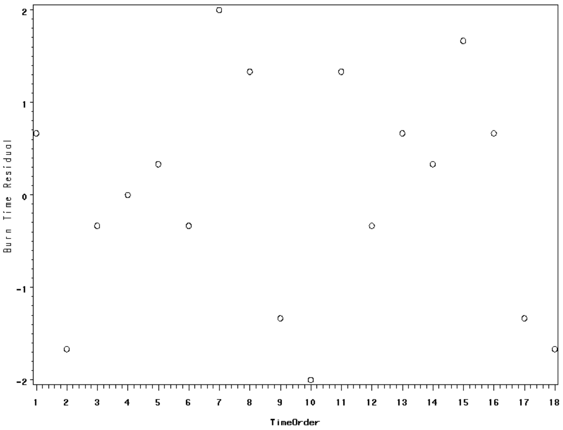
| Fan Status | Flavor | Burning Time | TimeOrder | Predicted | Residual |
|---|---|---|---|---|---|
| On2 | Vanilla | 15 | 5 | ____ | ____ |
| On2 | Vanilla | 16 | 11 | ____ | ____ |
| On2 | Vanilla | 13 | 18 | ____ | ____ |
| On2 | Cinnamon | 14 | 2 | ____ | ____ |
| On2 | Cinnamon | 17 | 8 | ____ | ____ |
| On2 | Cinnamon | 16 | 14 | ____ | ____ |
| On4 | Vanilla | 19 | 6 | ____ | ____ |
| On4 | Vanilla | 21 | 15 | ____ | ____ |
| On4 | Vanilla | 18 | 17 | ____ | ____ |
| On4 | Cinnamon | 21 | 1 | ____ | ____ |
| On4 | Cinnamon | 20 | 3 | ____ | ____ |
| On4 | Cinnamon | 20 | 12 | ____ | ____ |
| Off | Vanilla | 27 | 4 | ____ | ____ |
| Off | Vanilla | 29 | 7 | ____ | ____ |
| Off | Vanilla | 25 | 10 | ____ | ____ |
| Off | Cinnamon | 26 | 9 | ____ | ____ |
| Off | Cinnamon | 28 | 13 | ____ | ____ |
| Off | Cinnamon | 28 | 16 | ____ | ____ |

a. Calculate the residuals and predicted values associated with each of the observations and fill in the blanks. Residual plots based on these residuals are provided in parts (b) - (g).

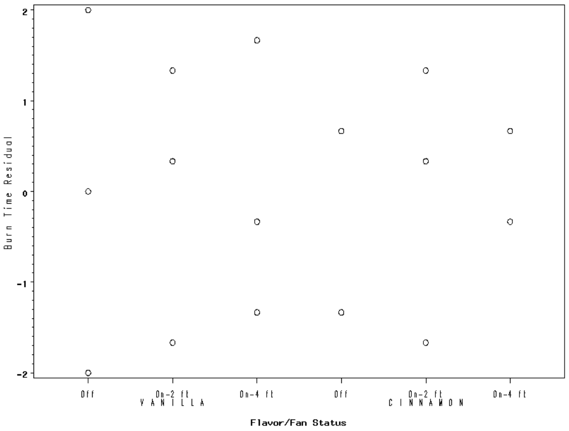
b. Consider the following plot of burn times versus treatment. Can this plot be used to check any of the assumptions about the error terms? Explain in the context of this data.



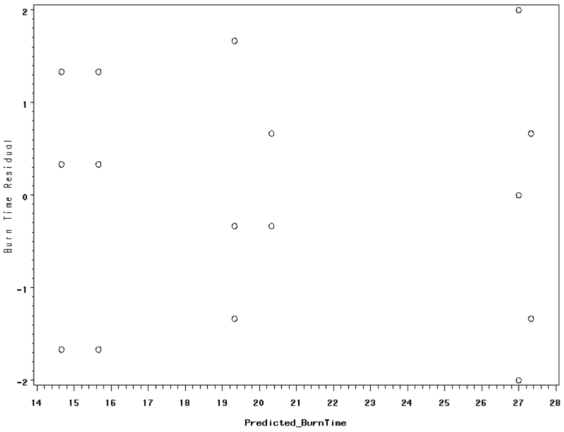
c. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



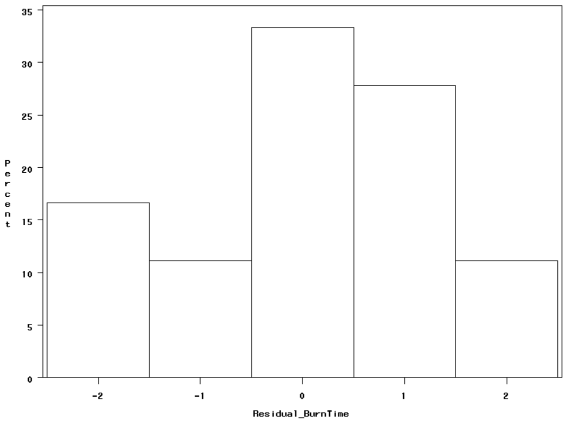
d. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



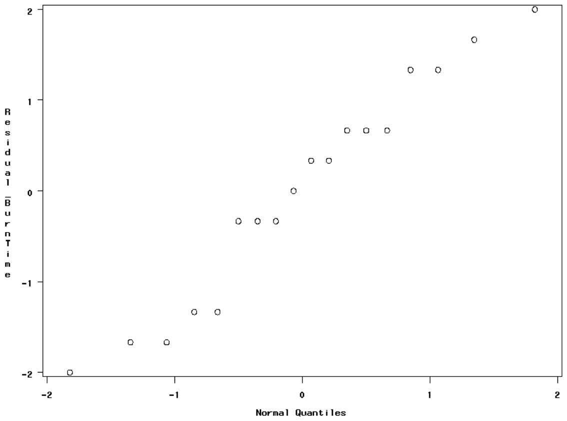
e. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



f. Consider the following histogram. What assumption is this plot used to check? Comment on the assumption for this data.



g. Consider the following Q-Q plot. What assumption is this plot used to check? Comment on the assumption for this data.



8.3* In the article "Lipid Pattern in Experimental Canine Atherosclerosis" (*Circular Research* [1964]: Vol. XIV, pgs 61-72) researchers investigated the effects on total serum lipids (mg/100 ml serum) of the addition of cholesterol and thiouracil to the diets of canines. The treatment group sizes, means, and standard deviations are given in the following table.

| Diet | n | Mean | Standard Deviation |
|---|---|---|---|
| Basal | 4 | 556 | 93.8 |
| Basal + Cholesterol | 6 | 879 | 357.6 |
| Basal + Thiouracil | 6 | 1807 | 497.2 |
| Basal + Cholesterol + Thiouracil | 6 | 3393 | 967.5 |

Is there evidence that any of the assumptions associated with ANOVA is likely violated? Explain.

8.4* Cory Caswell, Selina Cole, and Matt Snow (Spring 2010) ran an experiment to determine the effects of type of cup and type of liquid on the amount of time (minutes) that it took the liquid to cool from 200° to 100° Fahrenheit. The three types of 6 ounce cups were Paper, Plastic, and Styrofoam. The two types of liquid were water and coffee. The 30 trials, 5 replications per treatment, were run randomly through time. The data, including time order, is given in the table below.

| Time Order | Cup | Liquid | Amount of Time(mins) |
|---|---|---|---|
| 1 | Paper | Coffee | 40.3 |
| 2 | Plastic | Coffee | 37.1 |
| 3 | Styrofoam | Water | 47.6 |
| 4 | Plastic | Water | 40.0 |
| 5 | Styrofoam | Coffee | 49.9 |
| 6 | Styrofoam | Coffee | 51.2 |
| 7 | Styrofoam | Coffee | 46.9 |
| 8 | Paper | Coffee | 45.3 |
| 9 | Paper | Coffee | 47.2 |
| 10 | Paper | Water | 43.6 |
| 11 | Plastic | Water | 38.8 |
| 12 | Styrofoam | Water | 51.2 |
| 13 | Styrofoam | Water | 50.4 |
| 14 | Plastic | Coffee | 39.8 |
| 15 | Paper | Coffee | 51.2 |
| 16 | Plastic | Coffee | 43.3 |
| 17 | Plastic | Coffee | 40.7 |
| 18 | Styrofoam | Coffee | 52.9 |
| 19 | Styrofoam | Coffee | 57.2 |
| 20 | Plastic | Water | 42.3 |
| 21 | Paper | Water | 47.0 |
| 22 | Plastic | Water | 42.0 |
| 23 | Paper | Water | 43.6 |
| 24 | Plastic | Coffee | 41.5 |
| 25 | Plastic | Water | 42.6 |
| 26 | Styrofoam | Water | 47.30 |
| 27 | Paper | Water | 42.3 |
| 28 | Paper | Water | 43.3 |
| 29 | Styrofoam | Water | 51.8 |
| 30 | Paper | Coffee | 47.2 |

Use a statistical program to obtain the residuals based on the two factor completely randomized model. Obtain appropriate numerical summaries and residual plots to check on the three assumptions discussed in this chapter. Do the assumptions appear to be approximately satisfied? Comment.

8.5 Refer to the egg buoyancy study in Problem 6.8 of Chapter 6. Obtain appropriate numerical summaries and residual plots to check on the three assumptions discussed in this chapter. Do the assumptions appear to be approximately satisfied? Comment.

8.6 Refer to the leaflet angle data in Problem 4.11 of Chapter 4. Use a statistical program to obtain numerical summaries and residual plots to check on the three assumptions discussed in this chapter. Do the assumptions appear to be approximately satisfied? Comment.

# Chapter 9

# Two Factor Split Plot Designs

The simplest of the split plot designs is a two factor design. The completely randomized two factor design was examined in Chapter 6. In that design there was only one kind of experimental unit. All treatment combinations were assigned completely at random to the same kind of experimental unit. For example in the agricultural example of Chapter 6 all six combinations of type of fertilizer and watering regimen were applied to the same kind of experimental unit, a small plot of land. In some experiments, for reasons to be explored later, the levels of one factor, A, are applied to one type of experimental unit and the levels of factor B are applied to subunits of the units assigned to factor A. For example, suppose in an educational experiment whole classes of students receive one of three types of teaching method (factor A). Suppose within each class the students are divided into two groups with each group receiving a level of another factor B: usage of library or not. The treatment structure is factorial – all combinations of A and B are used. But the levels of A are applied/assigned to one type of experimental unit, the whole class, and the levels of B are applied to another type of experimental unit, subgroups of a class. This is an example of a split plot design. In a completely randomized design for this study there would only be one type of unit, say classes, and the combinations of teaching method and library usage or not would be assigned to classes.

In the general split plot design the levels of factor A are assigned to larger **"whole units"** and the levels of factor B are assigned to smaller **"subunits"** or bf "split units" of the whole units. In the education experiment above the whole units are the classes of students and the two groups of students within each class are the split units. In an agricultural experiment the whole units are often large plots of land and the split units are subdivisions of the large plot, or "split plots." Hence the name split plot design. Factor A is referred to as the whole unit factor and factor B the split unit factor. Since the split units in a split plot design can be arranged/grouped by whole unit the whole unit is also a blocking variable (Chapter 7). Thus in the education experiment the whole unit class is a blocking variable.

Split plot designs are often used when one treatment factor requires larger units than the other treatment factor. The classical example is in agricultural studies where for example different irrigation methods require larger plots of land due to the equipment size but a second factor, such as fertilizer may be applied to smaller plots within each larger plot. In a study to compare the effects of 2 different oven temperatures and 3 different cupcake recipes on the height of cupcakes, temperatures are assigned at random to different oven runs, the larger units, but the different types of cupcakes may be tested within the same oven run using different shelves of the oven, the smaller or split units. This also results in a more efficient use of resources. Using only one combination of temperature and recipe for each oven run would require more oven runs than using an oven run for three combinations, a particular temperature with the 3 different recipes. In medical studies each person may be given each of three different medications for a condition in random order. It is believed that the response to the medications may depend on gender. Two groups of subjects are used, males and females, each receiving the three medications. The design is a split plot design. The larger or whole units are the subjects. Gender is the whole unit classification factor. The smaller units are the time periods when each subject receives the randomly assigned medications.

## 9.1 Arrangement of Whole Units

The levels of factor A in a split plot design can be assigned to the whole units in a completely randomized fashion. Alternatively the whole units could be blocked first and the levels of A then assigned at random to the whole units within each block. Regardless of the treatment design for the whole unit factor A, completely randomized or block, the levels of B are assigned completely at random to the split units within each whole unit.

### 9.1.1 Examples: Split Plot Design with Whole Units in a Completely Randomized Design

The education example described earlier is an example of this situation. Suppose that there are 9 classes available and the three teaching methods (factor A) are assigned completely at random to the 9 classes with 3 classes being assigned to each of the 3 methods. The whole units are the classes. The whole unit factor is teaching method. Within each of the classes 2 groups of students are formed (at random). The two groups within each of the classes form the subunits or split units of the whole unit. The two levels of factor B (library usage or not) are assigned at random to the two subgroups within each class. The split unit factor is library usage or not. Note that each whole unit is actually a block of split units according to Chapter 7. In this example each class (whole unit) is a block of two groups (two split units) of students.

In the article "Changes in Women's Plasma Lipid and Lipoprotein Concentrations Due to Moderate Consumption of Alcohol Are Affected by Dietary Fat Level" (Journal of Nutrition, Vol. 129, pages 1713-1717, 1999) researchers studied the impact of substituting ethanol for dietary carbohydrate, in high-and low-fat diets, on plasma lipids (mmol/L) and lipoprotein concentrations (g/L). Twenty-six women were randomly assigned to consume either a high-fat (n=12) or a low-fact diet (n = 14) for a twelve week period in a controlled feeding study The women were the whole units and the high versus low fat diet was the whole unit factor. During the twelve week period each woman's diet was supplemented either with ethanol or carbohydrate, each during a 6-week period, randomly ordered. The split units are the two 6-week periods for each woman. The split unit factor is the supplement (ethanol or carbohydrate). Each women serves as a block of 2 6-week periods.

Models for the split plot designs are useful in situations where the whole units are (in theory) random samples from populations. The whole unit factor is defined by inherent characteristics of the whole units rather than being assigned to the units. Each sampled whole unit serves as a block of smaller split units which are assigned to levels of a split unit factor. In the article "Enhanced Food Intake Regulatory Responses after a Glucose Drink in Hyperinsulinemic Men" (International Journal of Obesity (2007) 31, 1222-1231) researchers studied the effect of a glucose drink on food intake regulation for two groups of males. One group (n = 33) was classified as hyperinsulinemic (HI), if their fasting plasma insulin was $\geq 41 pmol/l$ or another group (n = 33) normohyperinsulinemic (NI) if their fasting plasma insulin was $< 40 pmol/l$. The whole units are the males. The whole unit "factor" is hyperinsulinemic level (HI or NI). Each male consumed either a noncaloric sweetened drink (placebo) or a glucose-containing drink on two separate occasions in random order. Split units are the time periods for each subject when consuming the drink. Split unit factor is type of drink consumed (placebo or glucose-containing). One of the response variables was food intake based on a pizza meal 1 hour after consuming the drink.

### 9.1.2 Examples: Split Plot Design with Whole Units Arranged in Blocks

Suppose in an agricultural experiment two factors are being studied, irrigation method (factor A) with two levels, and type of fertilizer (factor B) with three levels. Suppose that 10 large plots are grouped into 5 blocks each with 2 plots. The arrangement of the large plots is carried out so that the two large plots within each block are similar with regard to soil composition. The two levels of irrigation are assigned completely at random to the two large plots within each block. The whole units in this experiment are the large plots and are arranged in blocks of two. Each of the two large plots within a block is divided into three subplots. The three fertilizers are assigned at random to the three subplots within each whole plot. The response variable might be yield of a crop planted on all 5 x 2 x 3 = 30 subplots. Note that the 30 subplots can be grouped by whole unit or large plot and thus large plot is a blocking variable. Thus there are two blocking variables in this study. One is soil composition where pairs of large plots have similar soil composition and the other is large plot, each large plot consisting of 3 subplots in close proximity.

In some experiments the blocking of the whole units corresponds to a time or replication variable. Milliken and Johnson ([21], page 468) describe an experiment designed to study the effects of A: amount of pressure in a water line (10, 20, 40, 80 psi) and B: type of nozzle (round hole, oval hole, narrow slit, all of equal area) on the amount of water that flowed through a nozzle. The experiment was conducted as follows. On a particular day a pressure was selected at random from the four and then the pressure in the water line set at that pressure. With the pressure at that level the four nozzle types were tested in a random order. Amount of water flowing (oz) through the nozzle in a fixed amount of time was measured. This process was repeated for the other pressures. The whole process was repeated on two other days resulting in 3 replications per combination of amount of pressure and type of nozzle. The whole units are the larger time slots when the line at a particular pressure is tested. The split units are smaller time slots within the whole units corresponding to the three nozzle tested. The whole units are blocked by Day since all four pressures are used in a Day. The whole unit factor is amount of pressure. Each whole unit is a blocking of the 3 smaller time slots for the different nozzles. The split unit factor is the type of nozzle.

## 9.2 Analysis of Split Plot Design - Whole Units in a Completely Randomized Design

### 9.2.1 The Model

The model for the split plot design where the whole units are assigned to the levels of the whole unit factor A in a completely randomized design is:

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{9.1}$$

where $i = 1, ..., a$, with $a$ being the number of levels of the whole unit factor A, $j = 1, ..., b$, with $b$ being the number of levels of the split unit factor B, $k = 1, ..., n$, with $n$ being the number of whole units assigned to each level of factor A, and

- $y_{ijk}$ is the observation on the response variable for the split unit for the $i^{th}$ level of the factor A, $k^{th}$ whole unit receiving or "nested" within the $i^{th}$ level of A, and receiving the $j^{th}$ level of the factor B.
- $\mu$ is the grand mean of the response variable averaged over a population of split units, all levels of factor A, and all levels of factor B.
- $\alpha_i$ is the true effect of the $i^{th}$ level of the factor A on the response variable
- $\epsilon^w_{k(i)}$ is the error term for the the $k^{th}$ whole unit "nested" within the $i^{th}$ level of the factor A, representing the effect of extraneous variables associated with the whole unit.

- $\beta_j$ is the true effect of the $j^{th}$ level of the factor B on the response variable.
- $\alpha\beta_{ij}$ is the true interaction effect on the response variable of the $i^{th}$ level of A and the $j^{th}$ level of B
- $\epsilon^s_{ijk}$ is the error term for the split unit associated with the $i^{th}$ level of A, $k^{th}$ whole unit nested under the $i^{th}$ level of A, and the $j^{th}$ level of B, representing the effect of extraneous variables with this split unit.

Note that there are two error terms in the model because there are two types of experimental units, the whole unit and the split unit. The whole unit error, denoted by $\epsilon^w_{k(i)}$, represents differences in the whole units getting assigned to the levels of A. The split unit error, denoted by $\epsilon^s_{ijk}$, represents differences in the split units assigned to the levels of B.

The model assumes that the whole unit errors are independent normal random variables each with mean 0 and common variance $\sigma^2_w$ and that the split unit errors are independent normal random variables each with mean 0 and common variance $\sigma^2_s$. It is also assumed that a whole unit error is independent of a split unit error.

The model assumes that the design is balanced. That is there is the same number of observations, $n$, at each treatment combination of the levels of factor A and B.

While the errors are all independent of one another the model hypothesizes that the observations on the response for the split units within each whole unit are correlated since those observations have a common factor, that being a common whole unit, and that the observations are equally correlated.

### 9.2.2 The ANOVA Table

The ANOVA table is derived in a manner similar to how the ANOVA table is derived for other designs. The observed responses can be partitioned into parts representing the grand mean, the effect of the particular level of factor A, the error associated with the whole unit, the effect of the particular level of factor B, the interaction effect, and the error associated with the split unit. The sum of squares of the deviations of the observed responses from the grand mean can be partitioned into sums describing variability in the effects of A, whole unit effects (errors), effects of B, interaction effects, and split unit effects (error).

$$SSTOT_C = SSA + SSE_w + SSB + SSAB + SSE_s$$

The general form of the ANOVA table with expected mean squares is given in Table 9.1

In Table 9.1 the sums of squares for the various effects are given without formulas. We will rely on the computer to calculate these. The column labelled $EMS$ gives the expected or population average mean squares. The values $Q_1$, $Q_2$, and $Q_3$ in the EMS column are, respectively, functions of the true A effects, B effects, and AB effects, which are zero if the corresponding effects are 0, that is the null hypothesis is true.

Table 9.1: ANOVA Table for Two Factor Split Plot Design - Whole Units in Completely Randomized Design

| Source of Variation | df | SS | MS | F | EMS |
|---|---|---|---|---|---|
| A | a - 1 | SSA | MSA | $MSA/MSE_w$ | $\sigma_s^2 + a\sigma_w^2 + Q_1$ |
| $Error_w$ | $a(n-1)$ | $SSE_w$ | $MSE_w$ | | $\sigma_s^2 + a\sigma_w^2$ |
| B | b - 1 | SSB | MSB | $MSB/MSE_s$ | $\sigma_s^2 + Q_2$ |
| A*B | $(a-1)(b-1)$ | SSAB | MSAB | $MSAB/MSE_s$ | $\sigma_s^2 + Q_3$ |
| $Error_s$ | $a(b-1)(n-1)$ | $SSE_s$ | $MSE_s$ | | $\sigma_s^2$ |

Let us consider testing for A effects. Under the null hypothesis, the expected mean square for A would be identical to the expected mean square for the whole plot error. Thus under the null hypothesis we would expect the ratio $MSA/MSE_w$ to be approximately 1. If the alternative hypothesis is true, then the expected mean square for A would be larger than $MSE_w$. In this case we would expect the ratio $MSA/MSE_w$ to be larger than 1. The test statistic for testing for A effects is the F ratio $MSA/MSE_w$. Assuming the model assumptions hold then the ratio $MSA/MSE_w$ has an $F$ distribution with numerator degrees of freedom $\nu_1 = a - 1$ and denominator degrees of freedom $\nu_2 = a(n-1)$. We will rely on the computer to calculate the $F$ ratio and obtain a P value for hypothesis testing.

Similarly test for B main effects and AB interaction can be tested using $F$ ratios. Note however that the denominator mean square error is $MSE_s$, unlike that for the test for A effects, for which the denominator is $MSE_w$. Thus the form of the ratio for the $F$ statistic depends upon the effect being tested.

### 9.2.3 An Example of a Split Plot Study

**Example 9.1** *John Szarka and Zamda Lumbi (Fall 2004) were interested in investigating the effects of type of flour (white, wheat, bread) and length of time in oven (*5*,* 10*,* 15 *minutes) on the change in height of dough after baking. Three rolls of dough were made from each type of flour for a total of nine rolls. Each roll was made using the same ingredients except for the type of flour. Each roll was divided into 3 equal parts and the* 3 *parts put into an oven. One part was baked at* 5 *minutes, another part at* 10 *minutes, and another for* 15 *minutes. Thus one run of the oven involved one roll (*3 *parts). The type of flour used for a particular roll and run of the oven was selected at random. The* 3 *parts of the roll were assigned at random to locations in the oven and time of baking. At the end of the* 5*,* 10*, and* 15 *minute periods, the appropriate parts were taken out of the oven and measured for height change.*

*The data are given in Table 9.2*

This is an example of a split plot design. Type of flour is the whole unit/plot factor, A. The whole plot experimental unit is a roll at a particular baking period or oven run. The design structure for the whole units is completely randomized.

Table 9.2: Change in Height of Dough (mm) for Baking Experiment

| | | Baking Time (Min) | | |
|---|---|---|---|---|
| Type of Flour | | 5 | 10 | 15 |
| | Roll | | | |
| White | 1 | 44 | 46 | 47 |
| | 2 | 42 | 46 | 48 |
| | 3 | 42 | 43 | 43 |
| Wheat | 1 | 40 | 40 | 42 |
| | 2 | 40 | 41 | 41 |
| | 3 | 40 | 41 | 41 |
| Bread | 1 | 43 | 44 | 46 |
| | 2 | 43 | 44 | 45 |
| | 3 | 41 | 43 | 43 |

The types of flour are assigned completely at random to the rolls baked at a particular baking period.

The split unit/plot factor, B, is the amount of time that a part of a roll is baked. The split unit experimental unit is the part of the dough which we will call "biscuit." The biscuits are arranged by roll so roll, the whole unit, also serves as a block.

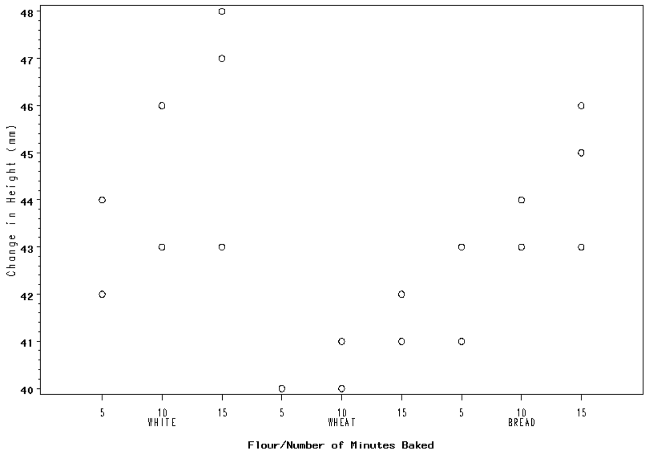
The model for this data is:

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{9.2}$$

with $i = 1(\text{white}), i = 2(\text{wheat}), i = a = 3(\text{bread})$ indexing type of flour, $j = 1(5 \text{ min}), j = 2(10 \text{ min}), j = b = 3(15 \text{ min})$ indexing baking time, and $k = 1, 2, 3(n)$ indexing the roll made with a particular flour at an oven run, and

- $y_{ijk}$ is the observation on height change, in millimeters, at the $i^{th}$ level of flour type, $k^{th}$ roll nested within the $i^{th}$ level of flour type, and $j^{th}$ level of baking time.
- $\mu$ is the grand mean of height change averaged over a population of rolls, all levels of flour type, and all levels of baking time.
- $\alpha_i$ is the true effect of the $i^{th}$ level of the flour type on height change.
- $\epsilon^w_{k(i)}$ is the error term for the the $k^{th}$ roll nested within the $i^{th}$ level of the flour type, representing the effect of extraneous variables associated with the roll, such as differences in amount of kneading, ingredients, etc.
- $\beta_j$ is the true effect of the $j^{th}$ level of baking time on height change of bread.

Figure 9.1: Plot of Height Increase versus Flour/Baking Time



- $\alpha\beta_{ij}$ is the true interaction effect on height change of the $i^{th}$ level of flour type and the $j^{th}$ level of baking time.
- $\epsilon^s_{ijk}$ is the error term for the split unit, here biscuit, associated with the $i^{th}$ level of flour type, $k^{th}$ roll nested under the $i^{th}$ level of flour type, and the $j^{th}$ level of baking time, representing the effect of extraneous variables for this unit, such as slight variations in baking time, variations in temperature of oven, within roll variations such as differences due to uneven mixing of ingredients.

The model assumes that the 9 "roll" errors, $\epsilon^w_{k(i)}$, are independent normal random variables each with mean 0 and common variance $\sigma^2_w$ and that the 27 "biscuit" errors, $\epsilon^s_{ijk}$, are independent normal random variables each with mean 0 and common variance $\sigma^2_s$. It is also assumed that a "roll" error is independent of a "biscuit" error.

While the errors are all independent of one another the model hypothesizes that the observations on height change for the three biscuits of a particular roll are correlated since those observations have a common factor, that being the common roll and a common run of the oven, and that those correlation are all the same.

A plot of change in height versus type of flour and baking time is given in Figure 9.1. Type of flour appears to have an effect with bread and white flour resulting in greater increases in height. As expected increases in baking time are associated with increases in height change.

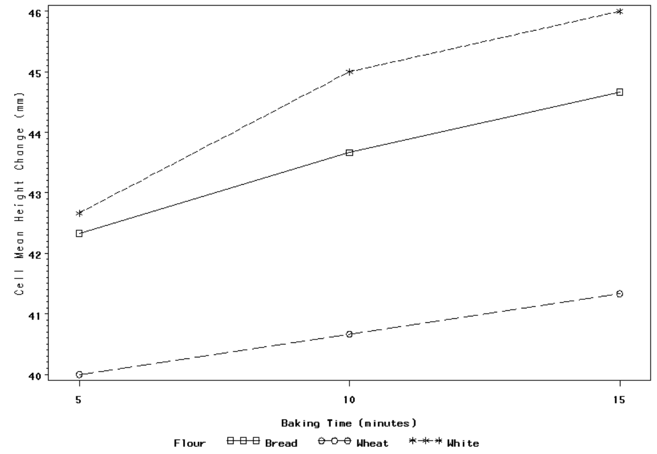
Height change treatment, marginal, and grand means are provided in Table 9.3.

An interaction plot is given in Figure 9.2. There is no strong evidence of

Table 9.3: Height Change Means: Bread Example

| | Baking Time | | | |
|---|---|---|---|---|
| | 5 | 10 | 15 | $\overline{y}_{i..}$ |
| Flour Type | | | | |
| White | 42.7 | 45.0 | 46.0 | 44.6 |
| Wheat | 40.0 | 40.7 | 41.3 | 40.7 |
| Bread | 42.3 | 43.7 | 44.7 | 43.6 |
| $\overline{y}_{.j.}$ | 41.7 | 43.1 | 44.0 | |
| | | | | $\overline{y}_{...} = 42.9$ |

Figure 9.2: InteractionPlot



interaction between type of flour and baking time.

The ANOVA table for the baking experiment is given in Table 9.4.

The null and alternative hypotheses for the test of interaction between type of flour and baking time are $H_o : \alpha\beta_{ij} = 0$ for each pair $(i, j), i = 1, 2, 3; j = 1, 2, 3$ and $H_a : \alpha\beta_{ij} \neq 0$ for some pair $i, j$. There is no evidence of interaction between type of flour and baking time $(F = 1.17, P - value = 0.3701))$ at the 0.10 level.

The null and alternative hypotheses for the test of main effects of type of flour are $H_o : \alpha_1 = \alpha_2 = \alpha_3 = 0$ and $H_a :$ not all $\alpha_i' s = 0$. There is evidence of differences in height change among types of flour $(F = 9.53, P - value = 0.0137$ at the 0.05) level of significance.

The null and alternative hypotheses for the test of main effects of baking time are $H_o : \beta_1 = \beta_2 = \beta_3 = 0$ and $H_a :$ not all $\beta_i' s = 0$, respectively.

Table 9.4: ANOVA Table for Baking Experiment

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Flour | 2 | 73.41 | 36.70 | 9.53 | 0.0137 |
| Error (Roll(Flour)) | 6 | 23.11 | 3.85 | | |
| BakeTime | 2 | 24.96 | 12.48 | 16.85 | 0.0003 |
| Flour*BakeTime | 4 | 3.48 | 0.87 | 1.17 | 0.3701 |
| Error (Biscuit) | 12 | 8.89 | 0.74 | | |

There is evidence of differences in height change among the baking times ($F = 9.53, P - value = 0.0137$ at the 0.05 level of significance.

Tukey-Kramer confidence intervals are used to make pairwise comparisons of marginal means of height change for types of flour and and baking times.

The Tukey-Kramer confidence intervals with overall experiment-wise confidence level $(1 - \alpha)$ corresponding to two levels $i$ and $i'$ of the whole plot factor A, type of flour, are

$$(\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}) \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}} \sqrt{MSE_w} \sqrt{\frac{1}{bn} + \frac{1}{bn}}$$

where $\sqrt{MSE_w}\sqrt{\frac{1}{bn} + \frac{1}{bn}}$ is the standard error of the difference in marginal means $\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}$ and $q_{\alpha;\nu,t}$ is the upper $\alpha$ probability point from the Studentized range distribution. Here $\nu$ refers to degrees of freedom associated with $MSE_w$, whole plot mean squared error and $t = a$, the number of levels of the whole plot factor A.

For the comparisons involving types of flour the appropriate MSE is mean squared error for the roll effect $MSE_w = 3.85$. The value $bn = (3)(3) = 9$ in the denominator is the number of observations contributing to a flour mean. Thus the standard error of the difference between two flour (sample) means is $\sqrt{\frac{2(3.85)}{9}} = 0.92$. Table A.6 with $\nu = 6$ degrees of freedom associated with Roll error and $t = a = 3$ levels for the flour factor gives $q_{0.05;6,3} = 4.34$ for overall 95% confidence. Thus the multiplier on the standard error is $\frac{4.34}{\sqrt{2}} = 3.1$. Thus the endpoints for the intervals for $\mu_{3\cdot} - \mu_{2\cdot}$, $\mu_{1\cdot} - \mu_{2\cdot}$, and $\mu_{1\cdot} - \mu_{3\cdot}$, respectively, are:

$$\begin{array}{c}(43.6 - 40.7) \pm (3.1)(0.92)\\(44.6 - 40.7) \pm (3.1)(0.92)\\(44.6 - 43.6) \pm (3.1)(0.92)\end{array}$$

Thus the Tukey-Kramer simultaneous 95% confidence intervals are:

$$\begin{array}{rcccl}0.1 & \leq & \mu_{3.} - \mu_{2.} & \leq & 5.7\\1.1 & \leq & \mu_{1.} - \mu_{2.} & \leq & 6.7\\-1.8 & \leq & \mu_{1.} - \mu_{3.} & \leq & 3.8\end{array}$$

It is estimated that Bread flour increases mean height change of dough as compared to White flour by between 0.1 and 5.7 mm. It is estimated that White flour increases mean height change of dough compared to Wheat flour by between 1.1 and 3.8 mm. There is not enough evidence of a difference in mean height change between the White and Bread flours. These conclusions are supported by a 95% experiment-wise confidence level.

The Tukey-Kramer confidence intervals with overall confidence level $1-\alpha$ corresponding to the levels $j$ and $j'$ of the split plot factor B, here number of minutes in oven, are

$$(\overline{y}_{.j.} - \overline{y}_{.j'.}) \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}} \sqrt{MSE_s} \sqrt{\frac{1}{an} + \frac{1}{an}}$$

where $\sqrt{MSE_s}\sqrt{\frac{1}{an}+\frac{1}{an}}$ is the standard error of $\overline{y}_{.j.} - \overline{y}_{.j'.}$ and $q_{\alpha;\nu,t}$ is the upper $\alpha$ probability point from the Studentized range distribution. Here $\nu$ refers to the degrees of freedom associated with $MSE_s$, split plot mean squared error and $t = b$ refers to the number of levels of the split plot factor B.

For the comparisons involving the three baking times the appropriate MSE is mean squared error for the split plots $MSE_s = 0.74$. The value $(a)(n) = (3)(3) = 9$ is the number of observations contributing to a baking time marginal mean. Thus the standard error of the difference between two baking time means is $\sqrt{\frac{2(0.74)}{9}} = 0.41$. Table A.6 with $\nu = 12$ degrees of freedom associated with the split plot error and $t = b = 3$ levels for the baking time factor gives $q_{0.05;12,3} = 3.77$. Thus the multiplier on the standard error is $\frac{3.77}{\sqrt{2}} = 2.67$. Thus the endpoints for the intervals for $\mu_{.2} - \mu_{.1}$, $\mu_{.3} - \mu_{.1}$, and $\mu_{.3} - \mu_{.2}$ comparing the baking times 5 and 10, 5 and 15, and 10 and 15 minutes are:

$$\begin{array}{c}(43.1 - 41.7) \pm (2.67)(0.41)\\(44.0 - 41.7) \pm (2.67)(0.41)\\(44.0 - 43.1) \pm (2.67)(0.41)\end{array}$$

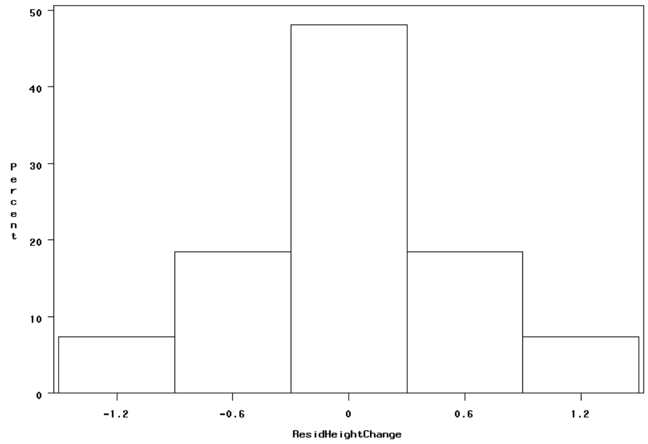
The Tukey-Kramer simultaneous 95% confidence intervals are:

$$\begin{array}{rcccl}0.3 & \leq & \mu_{.2} - \mu_{.1} & \leq & 2.5\\1.2 & \leq & \mu_{.3} - \mu_{.1} & \leq & 3.4\\-0.2 & \leq & \mu_{.3} - \mu_{.2} & \leq & 2.0\end{array}$$

It is estimated that the mean height change of dough for 10 minutes of baking time is between 0.3 and 2.5 mm greater than mean height change for 5 minutes of baking time. It is estimated that mean height change of dough for 15 minutes of baking time is between 1.2 and 3.4 mm greater than mean height change with 5 minutes of baking time. There is not enough evidence of a difference in mean height change of dough at 10 and 15 minutes of baking time. These conclusions are supported by a 95% experiment-wise confidence level.

A check is made of the assumptions of normality and homogeneous error variances associated with the split plot errors. Figure 9.3 gives a histogram

Figure 9.3: Histogram of Split Plot Residuals: Baking Experiment



of the split plot residuals from the model. Normality appears to be satisfied approximately.

Figure 9.4 gives a scatterplot of the split plot residuals versus the predicted height changes with the fitted model. There appears to be no patterns and thus the assumptions of homogeneity of error variance appears to be satisfied approximately.

### 9.2.4 Whole Units in CRD with Interaction

When there is evidence of interaction between the whole unit factor A and the split unit factor B then there might be interest in comparing the levels of B at each level of the A and/or comparing the levels of A at each of the levels of B. Details of the calculations that follow are based on Kutner, Nachtsheim, Neter, and Li ([15], pages 1148-1153).

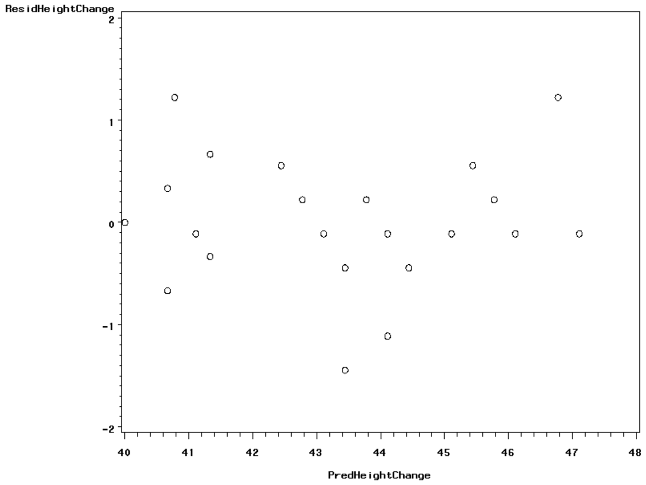
The Tukey-Kramer confidence intervals with experiment-wise confidence level $(1-\alpha)$ for all pairwise comparisons of treatment means with levels $j$ and $j'$ of the split plot factor B at a particular level $i$ of A are:

$$(\overline{y}_{ij\cdot} - \overline{y}_{ij'\cdot}) \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}} \sqrt{MSE_s}\sqrt{\frac{1}{n}+\frac{1}{n}}$$

where $\sqrt{MSE_s}\sqrt{\frac{1}{n}+\frac{1}{n}}$ is the standard error of the difference in two treatment means, $(\overline{y}_{ij\cdot} - \overline{y}_{ij'\cdot})$, and $q_{\alpha;\nu,t}$ is the upper $\alpha$ probability point from the Studentized range distribution. Here $\nu$ refers to the degrees of freedom associated with $MSE_s$, split plot MSE and $t = b$ refers to the number of levels of factor B.

The Tukey-Kramer confidence intervals with approximate experiment-wise confidence level $(1-\alpha)$ for all pairwise comparisons of treatment means with

Figure 9.4: Plot of Split Plot Residuals versus Predicted Change: Baking Experiment



levels $i$ and $i'$ of the whole unit factor A at a particular level $j$ of the split unit factor B are:

$$(\overline{y}_{ij\cdot} - \overline{y}_{i'j\cdot}) \pm \frac{q_{\alpha;\nu_{adj},t}}{\sqrt{2}} \sqrt{MSE_{adj}} \sqrt{\frac{1}{n} + \frac{1}{n}}$$

where $\sqrt{MSE_{adj}}\sqrt{\frac{1}{n} + \frac{1}{n}}$ is the approximate standard error of the difference in two treatment means, $(\overline{y}_{ij\cdot} - \overline{y}_{i'j\cdot})$ and $q_{\alpha;\nu_{adj},t}$ is the upper $\alpha$ probability point from the Studentized range distribution. Here $t = a$, the number of levels of the whole unit factor A, and

$$MSE_{adj} = \frac{a(n-1)MSE_w + a(b-1)(n-1)MSE_s}{ab(n-1)}$$

a pooling of the two mean squares associated with the two types of errors and degrees of freedom associated with $MSE_{adj}$, $\nu_{adj}$, with

$$\nu_{adj} = \frac{[SSE_w + SSE_s]^2}{\frac{[SSE_w]^2}{a(n-1)} + \frac{[SSE_s]^2}{a(b-1)(n-1)}}$$

Computer software will be used to obtain the results of the Tukey pairwise comparisons of treatment means when there is evidence of interaction.

**Example 9.2** *Suppose a study is conducted to investigate the effects of three different types of containers (A: ceramic mug, styrofoam cup, paper cup) and two heated liquids (B: coffee, hot chocolate) on the amount of time for the liquid to cool from 82 °C to 72 °C. The experiment was conducted as follows. A type*

Table 9.5: Cooling Time (seconds)

| | | Cooling Time (sec) | | |
|---|---|---|---|---|
| Liquid Type | | Ceramic Mug | Paper Cup | Styrofoam |
| | Time Slot | | | |
| Coffee | 1 | 78 | 212 | 281 |
| | 2 | 108 | 230 | 259 |
| | 3 | 73 | 234 | 266 |
| | 4 | 92 | 228 | 257 |
| Hot Chocolate | 1 | 96 | 222 | 260 |
| | 2 | 85 | 225 | 244 |
| | 3 | 81 | 237 | 254 |
| | 4 | 84 | 217 | 263 |

Table 9.6: Cooling Time Means (sec): Cooling Example

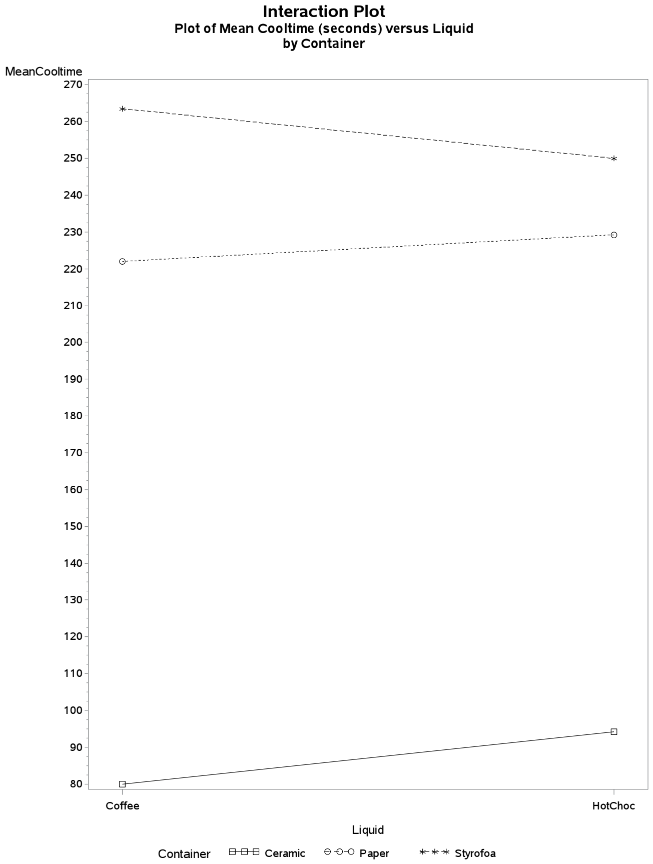
| | Container Type | | | |
|---|---|---|---|---|
| | Ceramic | Paper | Styrofoam | $\overline{y}_{i..}$ |
| Liquid Type | | | | |
| Coffee | 80.0 | 222.0 | 263.5 | 188.5 |
| Hot Chocolate | 94.3 | 222.3 | 250.0 | 191.2 |
| $\overline{y}_{.j.}$ | 87.1 | 225.6 | 256.8 | |
| | | | | $\overline{y}_{...} = 189.8$ |

*of liquid was randomly selected. Eighteen ounces of the liquid was poured into a pot and heated to 82° C. The heated liquid was then poured into containers of the three types, in a random order, with* 6 *ounces per container. The amount of time (seconds) until the liquid in each cup was cooled to 72 ° C was recorded. This process was repeated on 7 other occasions, resulting in* 4 *replications per combination of liquid type and container type. The design is a split plot design with whole units being equal to time slots corresponding to the pouring and heating of eighteen ounces of liquid with whole unit factor being equal to the type of liquid used. Whole units are assigned completely at random to type of liquid. Split units are the* 6 *ounce amounts of liquid in a cup, with split unit factor being type of cup. The data are given in Table 9.5*

Cooling Time treatment, marginal, and grand means are provided in Table 9.6.

An interaction plot is given in Figure 9.5. There is some evidence of inter-

Figure 9.5: InteractionPlot



action between liquid and container type.

The ANOVA table for the cooling experiment is given in Table 9.7. The interaction effect on cooling time between type of liquid and container is significant at the 0.10 level.

Tukey-Kramer comparisons of cooling time for pairs of containers at each of the liquids are given in Table 9.8. For both liquids Coffee and Hot Chocolate the differences between mean cooling time for Ceramic and Paper and Ceramic and Styrofoam are significant at the 0.05 experiment-wise significance level. The difference in cooling time between the Styrofoam and Paper container types is significant for Coffee but not for Hot Chocolate.

Tukey comparisons of cooling time for the two liquids at each of the container types is given in Table 9.9. The difference between cooling time for the two liquids is not significant for any of the container types.

Computer software (SAS Proc GLM and Mixed) was used to obtain the results given for the cooling experiment. The SAS code is provided in Section

Table 9.7: ANOVA Table for Liquid Cooling Experiment

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Beverage | 1 | 42.7 | 42.7 | 0.53 | 0.4924 |
| Whole Unit Error | 6 | 479.3 | 79.9 | | |
| Container | 2 | 130463 | 65232 | 483.5 | $< 0.0001$ |
| Beverage*Container | 2 | 833.1 | 416.5 | 3.09 | 0.0829 |
| Split Unit Error | 12 | 1619.2 | 134.9 | | |

Table 9.8: Tukey Pairwise Comparisons of Container Types for Each Liquid

| Liquid | Container | Container | Mean Difference | Std.Error | df | t Value | P-value | LCL | UCL |
|---|---|---|---|---|---|---|---|---|---|
| Coffee | Ceramic | Paper | -142.0 | 8.21 | 12 | -17.29 | <.0001 | -163.9 | -120.1 |
| Coffee | Ceramic | Styrofoam | -183.5 | 8.21 | 12 | -22.34 | <.0001 | -205.4 | -161.6 |
| Coffee | Paper | Styrofoam | -41.5 | 8.21 | 12 | -5.05 | 0.0008 | -63.4 | -19.6 |
| Hot Chocolate | Ceramic | Paper | -135.0 | 8.21 | 12 | -16.44 | <.0001 | -156.9 | -113.1 |
| Hot Chocolate | Ceramic | Styrofoam | -155.8 | 8.21 | 12 | -14.50 | <.0001 | -177.7 | -133.8 |
| Hot Chocolate | Paper | Styrofoam | -20.8 | 8.21 | 12 | -2.53 | 0.0640 | -42.7 | 1.2 |

Table 9.9: Tukey Pairwise Comparisons of Two Liquids for Each Container

| Container | Liquid | Liquid | Mean Difference | Std.Error | df | t Value | P-value | LCL | UCL |
|---|---|---|---|---|---|---|---|---|---|
| Ceramic | Coffee | Hot Chocolate | -14.3 | 7.63 | 17.15 | -1.87 | 0.0866 | -30.9 | 2.4 |
| Paper | Coffee | Hot Chocolate | -7.3 | 7.63 | 17.15 | -0.95 | 0.3611 | -23.9 | 9.4 |
| Styrofoam | Coffee | Hot Chocolate | 13.5 | 7.63 | 17.15 | 1.77 | 0.1024 | -3.1 | 30.1 |

9.4.2.

## 9.3 Analysis of Split Plot Design - Whole Units Arranged in a Block Design

### 9.3.1 The Model

The model for the split plot design where the whole units are arranged in blocks is:

$$y_{ijk} = \mu + \alpha_i + \rho_k + \epsilon^w_{ik} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{9.3}$$

where $i = 1, ..., a$, with $a$ being the number of levels of factor A, $j = 1, ..., b$, with $b$ being the number of levels of factor B, and $k = 1, ..., n$, with $n$ being the number of blocks of the whole units, and

- $y_{ijk}$ is the observation on the response variable at the $i^{th}$ level of the factor A, $k^{th}$ block of whole units, and $j^{th}$ level of the factor B.
- $\mu$ is the grand mean of the response variable averaged over a population of subjects, all levels of factor A, and all levels of factor B.
- $\alpha_i$ is the true effect of the $i^{th}$ level of the factor A on the response variable.
- $\rho_k$ is the true effect of the $k^{th}$ level of the blocking variable.
- $\epsilon^w_{ik}$ is the error term for the whole unit assigned to factor level $i$ in block $k$ representing the effect of extraneous variables associated with the whole unit.
- $\beta_j$ is the true effect of the $j^{th}$ level of the factor B on the response variable.
- $\alpha\beta_{ij}$ is the true interaction effect on the response variable of the $i^{th}$ level of A and the $j^{th}$ level of B.
- $\epsilon^s_{ijk}$ is the error term for the split unit receiving the $j^{th}$ level of factor B in the whole unit receiving level $i$ of factor A in the $k^{th}$ block, representing the effects of extraneous variables associated with that split unit.

The model assumes that the whole unit errors are independent normal random variables each with mean 0 and common variance $\sigma^2_w$ and that the split unit errors are independent normal random variables each with mean 0 and common variance $\sigma^2_s$. In this chapter it will be assumed that the levels of the blocks are random and that the block effects are independent, normal random variables, each with mean 0 and common variance $\sigma^2_\rho$. It is also assumed that the whole unit errors, split unit errors, and the block effects are statistically independent of one another. While the errors and block effects are all independent of one another the model hypothesizes that the observations on the response for the split units within each whole unit are correlated since those observations have a common factor, that being a common whole unit, and that this correlation is the same for all pairs of responses on the split units.

Table 9.10: ANOVA Table: Split Plot Design - Whole Units Blocked

| Source of Variation | df | SS | MS | F | EMS |
|---|---|---|---|---|---|
| Blocks | n - 1 | SSBlocks | MSBlocks | | $\sigma_s^2 + b\sigma_w^2 + ab\sigma_\rho^2$ |
| A | a - 1 | SSA | MSA | $MSA/MSE_w$ | $\sigma_s^2 + b\sigma_w^2 + Q_1$ |
| $Error_w$ | $(a-1)(n-1)$ | $SSE_w$ | $MSE_w$ | | $\sigma_s^2 + b\sigma_w^2$ |
| B | b - 1 | SSB | MSB | $MSB/MSE_s$ | $\sigma_s^2 + Q_2$ |
| A*B | $(a-1)(b-1)$ | SSAB | MSAB | $MSAB/MSE_s$ | $\sigma_s^2 + Q_3$ |
| $Error_s$ | $a(n-1)(b-1)$ | $SSE_s$ | $MSE_s$ | | $\sigma_s^2$ |

### 9.3.2 The ANOVA Table

The ANOVA table for the split plot design where the whole units are arranged in blocks ([21], Chapter 24)is given in Table 9.10. The sums of squares for the various effects are given without formulas. Computer programs will be used to calculate these. Also the values $Q_1$, $Q_2$, and $Q_3$ are, respectively, functions of the A effects, B effects, and AB interaction effects, which are zero if the effects are 0, that is the null hypothesis is true.

Note again for this design that the mean squared error associated with the whole plots is the appropriate denominator for testing for factor A effects. The appropriate F ratio for testing for B and interaction effects uses mean squared error associated with the split units in the denominator. Again we will obtain F ratios and P-values using computer software.

### 9.3.3 Example

**Example 9.3** *This example is based on an experiment described in Cochran and Cox [5]. The original study was undertaken to investigate the effects of three chocolate cake recipes and 6 baking temperatures on the various quality characteristics of the cakes. The three recipes will simply be referred to as R1, R2, and R3. There were* 6 *temperatures used in the original experiment but we will use only three here, namely* 175, 195, *and* 215 *degrees Fahrenheit. There were three replications of the experiment with replications serving as blocks. So a block here refers to a time frame. At each replication a recipe was selected at random and then enough cake batter was prepared for three cakes. After making a particular batch the batch was split into three equal parts and each part assigned at random to one of the three oven temperatures. There were three ovens available for the experiment. The data is provided in Table 9.11. The response variable is a quality characteristic with higher values indicating greater quality.*

This is an example of a split plot design where the whole units are arranged in blocks. A block corresponds to a replication in which a set of three time slots are available to make three cake batter batches. The whole unit is a batch of cake batter prepared at a particular time slot. The whole plot factor, A, is recipe (R1,R2, and R3) whose levels are assigned at random to the three whole units for a replication. Whole units are blocked according to replication. The split

Table 9.11: Quality Data for Chocolate Cake Experiment

| | Recipe | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Replication/Block | R1 | | | R2 | | | R3 | | |
| Temp(°F) | 175 | 195 | 215 | 175 | 195 | 215 | 175 | 195 | 215 |
| 1 | 28 | 31 | 41 | 31 | 29 | 40 | 21 | 31 | 33 |
| 2 | 24 | 27 | 30 | 21 | 24 | 37 | 26 | 27 | 35 |
| 3 | 26 | 32 | 37 | 21 | 28 | 27 | 21 | 25 | 31 |

units are the three portions of a batch of cake batter prepared at a particular time slot. The three portions are assigned to the three ovens/temperatures. Temperature of oven is the split plot factor, B. The whole units (batches of cake) are blocked by replication. Each whole unit (batch of cake) within a replication serves as a block of three split units (portions of batch).

The model for the split plot design in this example is:

$$y_{ijk} = \mu + \alpha_i + \rho_k + \epsilon^w_{ik} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{9.4}$$

with $i = 1(\text{R1}), i = 2(\text{R2}), i = a = 3(\text{R3})$ indexing recipe, $j = 1(175°\text{F}), j = 2(195°\text{F}), j = b = 3(215°\text{F})$ indexing temperature, and $k = 1, 2, n = 3$ indexing replication.

- $y_{ijk}$ is the observation on the quality characteristic at the $i^{th}$ level of recipe, $k^{th}$ replication, and $j^{th}$ temperature.
- $\mu$ is the grand mean of the quality characteristic averaged over a population of cakes, all levels of recipe, and all levels of temperature.
- $\alpha_i$ is the true effect of the $i^{th}$ level of recipe on quality.
- $\rho_k$ is the true effect of the $k^{th}$ replication.
- $\epsilon^w_{ik}$ is the error term for the batch of cake batter assigned to recipe $i$ in replication $k$ representing the effect of extraneous variables associated with the cake batch.
- $\beta_j$ is the true effect of the $j^{th}$ level of temperature on quality.
- $\alpha\beta_{ij}$ is the true interaction effect on the quality of the $i^{th}$ level of recipe and the $j^{th}$ level of temperature.
- $\epsilon^s_{ijk}$ is the error term for the portion of cake batter batch receiving the $j^{th}$ level of temperature in the $k^{th}$ replication for recipe $i$, representing the effects of extraneous variables associated with the portion. These include within batch variations and variations in ovens.

The model assumes that the batch errors are independent normal random variables each with mean 0 and common variance $\sigma^2_w$ and that the portion errors are independent normal random variables each with mean 0 and common variance $\sigma^2_s$. It is also assumed that a batch error is independent of a portion error. While the errors are all independent of one another the model hypothesizes that the observations on the response quality for the three cakes made

Figure 9.6: Scatterplot of Quality Versus Recipe/Temperature

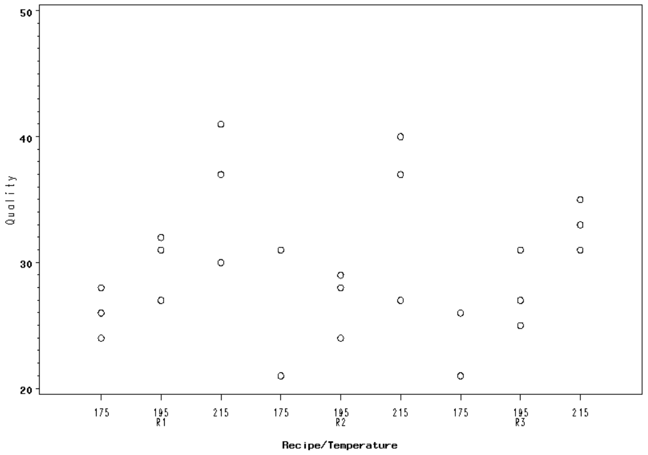


Table 9.12: Quality Means: Chocolate Cake Example

| | Baking Temperature °$F$ (j) | | | |
|---|---|---|---|---|
| | 175(1) | 195(2) | 215(3) | $\overline{y}_{i..}$ |
| Recipe (i) | | | | |
| R1 (1) | 26.0 | 30.0 | 36.0 | 30.7 |
| R2 (2) | 24.3 | 27.0 | 34.7 | 28.7 |
| R3 (3) | 22.7 | 27.7 | 33.0 | 27.8 |
| $\overline{y}_{.j.}$ | 24.3 | 28.2 | 34.6 | |
| | | | | $\overline{y}_{...} = 29.0$ |

from each batch of cake batter are correlated since those observations have a common factor, all cakes made from the same batch of cake batter.

A plot of quality versus recipe and oven temperature is given in Figure 9.6 Recipe does not appear to have an effect on quality. Oven temperature appears to affect quality.

Quality treatment, marginal, and grand means are provided in Table 9.12.

An interaction plot is given in Figure 9.7. There is no strong evidence of interaction between recipe and baking temperature.

The ANOVA table for this example is is given in Table 9.13. Note that there is no evidence of interaction between recipe and temperature ($F = 0.15, \text{P - value} = 0.9571$) at the 0.10 level of significance. The effects of recipe are not significant ($F = 0.82, \text{P-value} = 0.5038$) while the effects of temperature are significant ($F = 26.03, \text{P-value} < 0.0001$) using a 0.05 level of significance in both cases.

Figure 9.7: Interaction Plot

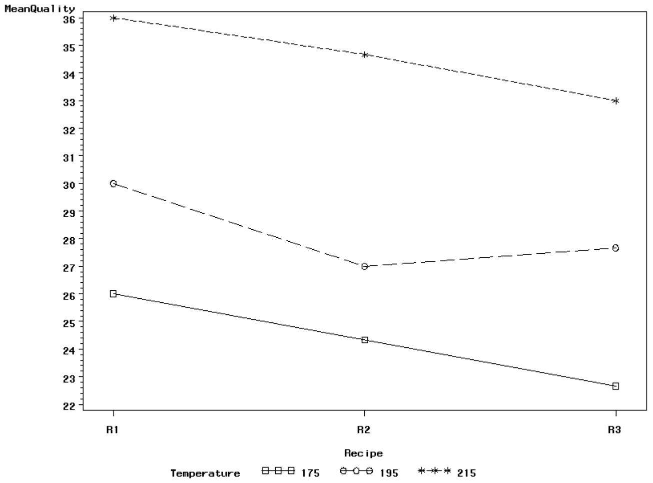


Table 9.13: ANOVA Table: Recipe and Temperature Baking Experiment

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Blocks | 2 | 93.85 | 46.93 | | |
| Recipe | 2 | 39.41 | 19.70 | 0.82 | 0.5038 |
| $Error_w$ | 4 | 96.37 | 24.09 | | |
| Temp | 2 | 479.19 | 239.59 | 26.03 | $< 0.0001$ |
| Recipe*Temp | 4 | 5.70 | 1.43 | 0.15 | 0.9571 |
| $Error_s$ | 12 | 110.44 | 9.20 | | |

Since the recipe effects are not significant pairwise comparisons of the marginal means would normally not be undertaken. However to illustrate the the appropriate mean square error to do Tukey-Kramer comparisons, comparisons of the recipes as well as the temperatures will be calculated.

The Tukey-Kramer confidence intervals with overall confidence level $(1-\alpha)$ for the levels of the whole plot factor A are as before:

$$(\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}) \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}} \sqrt{MSE_w} \sqrt{\frac{1}{bn} + \frac{1}{bn}}$$

where $\sqrt{MSE_w}\sqrt{\frac{1}{bn}+\frac{1}{bn}}$ is the standard error of the difference in two marginal means for factor A, $\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}$, and $q_{\alpha;\nu,t}$ is the upper $\alpha$ probability point from the Studentized range distribution. Here $\nu$ refers to the degrees of freedom associated with whole plot mean squared error, $MSE_w$ and $t = a$, the number of levels of the whole plot factor A.

For the comparisons involving recipes the appropriate MSE is mean squared error for whole unit, here $MSE_w = 24.09$. The value $bn = (3)(3) = 9$ is the number of observations contributing to a recipe mean. Thus the standard error of the difference between two recipe marginal means is $\sqrt{\frac{2(24.09)}{9}} = 2.31$. Table A.6 with $\nu = 4$ degrees of freedom associated with whole unit error and $t = a = 3$ levels for the recipe factor gives $q_{0.05;4,3} = 5.04$. Thus the multiplier on the standard error is $\frac{5.04}{\sqrt{2}} = 3.56$. Thus the endpoints for the intervals for the differences $\mu_{1.} - \mu_{2.}$, $\mu_{1.} - \mu_{3.}$, and $\mu_{2.} - \mu_{3.}$ for the comparisons of recipes R1 and R2, R1 and R3, and R2 and R3 are:

$$2.0 \pm (3.56)(2.31), \quad 2.9 \pm (3.56)(2.31), \quad 0.9 \pm (3.56)(2.31)$$

Thus the simultaneous 95% Tukey-Kramer confidence intervals are:

$$\begin{array}{rcccl} -6.2 & \leq & \mu_{1.} - \mu_{2.} & \leq & 10.2 \\ -5.3 & \leq & \mu_{1.} - \mu_{3.} & \leq & 11.1 \\ -7.3 & \leq & \mu_{2.} - \mu_{3.} & \leq & 9.1 \end{array}$$

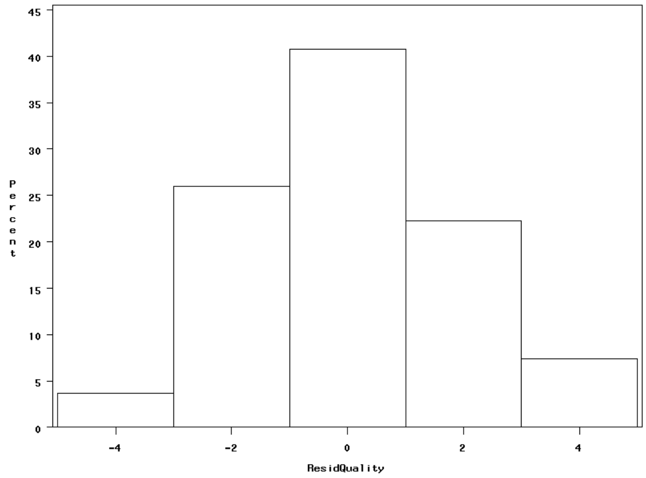
All three intervals contain zero and thus the comparisons are consistent with the results from the F test.

The Tukey-Kramer confidence intervals with overall confidence level $(1-\alpha)$ for the levels of the split plot factor B, here temperature, are

$$(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdot j'\cdot}) \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}} \sqrt{MSE_s} \sqrt{\frac{1}{an} + \frac{1}{an}}$$

where $\sqrt{MSE_s}\sqrt{\frac{1}{an}+\frac{1}{an}}$ is the standard error of the difference in two marginal means for factor B, $\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdot j'\cdot}$, and $q_{\alpha;\nu,t}$ is the upper $\alpha$ probability point from the Studentized range distribution. Here $\nu$ refers to the degrees of freedom associated with split plot mean squared error, $MSE_s$ and $t = b$, the number of levels of the split plot factor B.

Figure 9.8: Histogram of Residuals from Cake Experiment



For the comparisons involving the three oven temperatures the appropriate MSE is mean squared error for the split units (cake batter batch portion), $MSE_s = 9.20$. The value $an = (3)(3) = 9$ is the number of observations contributing to a temperature mean. Thus the standard error of the difference between two temperature marginal means is $\sqrt{\frac{2(9.20)}{9}} = 1.43$. Table A.6 with $\nu = 12$ degrees of freedom associated with the split plot error and $t = b = 3$ levels for the temperature factor gives $q_{0.05;12,3} = 3.77$. Thus the multiplier on the standard error is $\frac{3.77}{\sqrt{2}} = 2.67$. Thus the endpoints for the intervals for differences in temperatures $\mu_{\cdot 2} - \mu_{\cdot 1}$, $\mu_{\cdot 3} - \mu_{\cdot 1}$, and $\mu_{\cdot 3} - \mu_{\cdot 2}$ comparing the temperatures 175° and 195°, 175° and 215°, and 195° and 215° are:

$$\begin{array}{c}(28.2 - 24.3) \pm (2.67)(1.43)\\(34.6 - 24.3) \pm (2.67)(1.43)\\(34.6 - 28.2) \pm (2.67)(1.43)\end{array}$$

The Tukey-Kramer simultaneous 95% confidence intervals are:

$$\begin{array}{rcccl}0.08 & \leq & \mu_{\cdot 2} - \mu_{\cdot 1} & \leq & 7.7\\6.5 & \leq & \mu_{\cdot 3} - \mu_{\cdot 1} & \leq & 14.1\\2.6 & \leq & \mu_{\cdot 3} - \mu_{\cdot 2} & \leq & 10.2\end{array}$$

All pairwise comparisons of recipe mean quality are significant.

A check is made of the assumptions of normality and homogeneous error variances associated with the split plot errors. Figure 9.8 gives a histogram of the split plot residuals from the model. Normality appears to be satisfied approximately.

Figure 9.9: Scatterplot of Residuals versus Predicted for Cake Baking Experiment

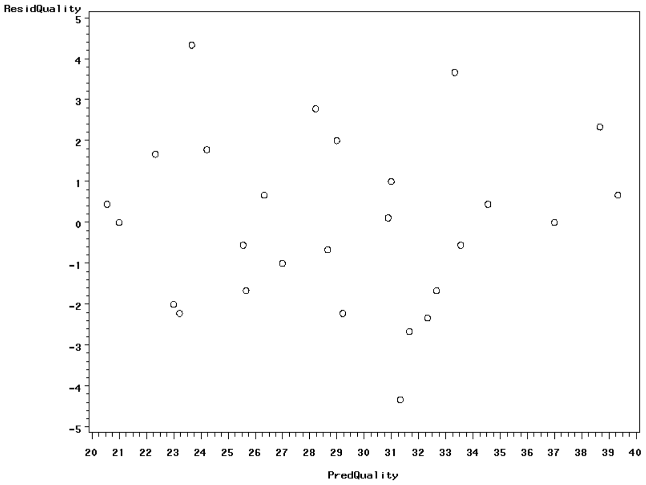


Figure 9.9 gives a plot of the split plot residuals versus the predicted quality for the fitted model. There appears to be no patterns and thus the assumption of homogeneity of error variance appears to be satisfied approximately.

## 9.4 SAS Code

### 9.4.1 Example 9.1

```
*  SAS Code for Example 9.1;


*  Input data;
data bake;
   input Flour $ Roll $  BakeTime   Treatment $ HeightChange;
datalines;
White   1   5    White5   44
White   1   10   White10  46
White   1   15   White15  47
White   2   5    White5   42
White   2  10    White10  46
White   2  15    White15  48
White   3   5    White5   42
White   3   10   White10  43
White   3   15   White15  43
Wheat   1   5    Wheat5   40
Wheat   1   10   Wheat10  40
Wheat   1   15   Wheat15  42
Wheat   2   5    Wheat5   40
Wheat   2  10    Wheat10  41
Wheat   2  15    Wheat15  41
Wheat   3   5    Wheat5   40
Wheat   3   10   Wheat10  41
Wheat   3   15   Wheat15  41
Bread   1   5    Bread5   43
Bread   1   10   Bread10  44
Bread   1   15   Bread15  46
Bread   2   5    Bread5   43
Bread   2  10    Bread10  44
Bread   2  15    Bread15  45
Bread   3   5    Bread5   41
Bread   3   10   Bread10  43
Bread   3   15   Bread15  43
;
run;

* Calculate and print means of height change;
proc means data = Bake;
  class Flour BakeTime;
  var HeightChange;
  output out = Summary  mean = MeanHeightChange;
```

```
run;
proc print data = Summary;
run;

* Compute ANOVA table and construct Tukey-Kramer pairwise comparisons;
proc glm data = bake;
  title1;
  class Flour Roll BakeTime;
  model HeightChange = Flour Roll(Flour) BakeTime Flour*Baketime;
  random Roll(Flour) / test;
  lsmeans Flour / pdiff tdiff adjust = tukey e = Roll(Flour);
  lsmeans BakeTime / pdiff tdiff adjust = tukey;
run;
```

### 9.4.2 Example 9.2

```
*  SAS Code for Example 9.2;


*  Input Cooling;
data bake;
   input Liquid $ TimeSlot $  Container $ Treatment $ CoolingTime;
datalines;
Coffee 1 Ceramic    CC 78
Coffee 1 Paper      CP 212
Coffee 1 Styrofoam  CS 281
Coffee 2 Ceramic    CC 85
Coffee 2 Paper      CP 225
Coffee 2 Styrofoam  CS 244
Coffee 3 Ceramic    CC 73
Coffee 3 Paper      CP 234
Coffee 3 Styrofoam  CS 266
Coffee 4 Ceramic    CC 84
Coffee 4 Paper      CP 217
Coffee 4 Styrofoam  CS 263
HotChoc 1 Ceramic   HC 96
HotChoc 1 Paper     HP 222
HotChoc 1 Styrofoam HS 230
HotChoc 2 Ceramic   HC 108
HotChoc 2 Paper     HP 230
HotChoc 2 Styrofoam HS 259
HotChoc 3 Ceramic   HC 81
HotChoc 3 Paper     HP 237
HotChoc 3 Styrofoam HS 254
HotChoc 4 Ceramic   HC 92
HotChoc 4 Paper     HP 228
HotChoc 4 Styrofoam HS 257
;
run;

* Calculate and print means of cooling times;

proc means data = Cooling;
  class Liquid Container;
  var CoolTime;
  output out = Summary  mean = MeanCoolTime;
run;
proc print data = Summary;
run;

* Use proc glm to compute ANOVA table;

proc glm data = Cooling;
  class Liquid TimeSlot Container;
  model CoolTime = Liquid TimeSlot(Liquid) Container Liquid*Container;
  random TimeSlot(Liquid) / test;
run;


* Use proc glimmix to construct tukey comparisons;

proc glimmix data = Cooling nobound;
  class TimeSlot Liquid Container;
  model CoolTime = Liquid Container Liquid*Container / ddfm = kr;
  random TimeSlot(Liquid);
  lsmeans Liquid*Container / slicediff = (Liquid Container) cl adjust = Tukey;
```

### 9.4.3 Example 9.3

```
*  SAS Code for Example 9.3;

*  Input data;
data Cake;
   input Block Recipe $ Temperature Treatment $ Quality;
datalines;
1 R1 175 R1_175 28
1 R1 195 R1_195 31
1 R1 215 R1_215 41
1 R2 175 R2_175 31
1 R2 195 R2_195 29
1 R2 215 R2_215 40
1 R3 175 R3_175 21
```

```
1 R3 195 R3_195 31
1 R3 215 R3_215 33
2 R1 175 R1_175 24
2 R1 195 R1_195 27
2 R1 215 R1_215 30
2 R2 175 R2_175 21
2 R2 195 R2_195 24
2 R2 215 R2_215 37
2 R3 175 R3_175 26
2 R3 195 R3_195 27
2 R3 215 R3_215 35
3 R1 175 R1_175 26
3 R1 195 R1_195 32
3 R1 215 R1_215 37
3 R2 175 R2_175 21
3 R2 195 R2_195 28
3 R2 215 R2_215 27
3 R3 175 R3_175 21
3 R3 195 R3_195 25
3 R3 215 R3_215 31
;
run;

* Calculate and print quality means;
proc means data = Cake;
  class Recipe Temperature;
  var Quality;
  output out = Summary  mean = MeanQuality;
run;
proc print data = Summary;
run;

* Calculate ANOVA table and results of Tukey-Kramer pairwise comparisons;
proc glm data = Cake;
  class Block Recipe Temperature;
  model Quality = Block Recipe Block*Recipe Temperature Recipe*Temperature;
  random Block Block*Recipe / test;
  lsmeans Recipe / pdiff tdiff adjust = tukey e = Block*Recipe;
  lsmeans Temperature / pdiff tdiff adjust = tukey;
run;
```

# Problems for Chapter 9

9.1* A researcher was interested in comparing the growths of three strains of petunias (A,B,C) grown at different temperatures. The plants were to be grown in growth chambers where temperature could be controlled. Nine growth chambers altogether were used, three chambers randomly assigned to each of 70, 75, and 80 degree temperatures. Within each growth chamber three saplings, one of each strain, were assigned at random to three pots and locations within the growth chamber. The saplings were grown in the chambers for one month. At the end of the month the growth in inches was recorded. This is an example of a split plot experiment.

a. What is the whole plot factor? What is the whole plot experimental unit? Give some extraneous variables that contribute to whole plot experimental error.

b. Are the whole plot experimental units arranged in a completely randomized design or a block design? Explain.

c. What is the split plot factor? What is the split plot experimental unit? Give some extraneous variables that contribute to split plot experimental error.

d. Describe the blocking variable(s) used in this study.

e. Give the population effects model for this experiment and describe the terms in the model including the error terms. Give the assumptions associated with the errors.

9.2* An experiment was conducted to investigate the effects of background music and font color in memorizing a list of words. Three kinds of music were investigated: classical, reggae, and jazz. Three font colors were used in the list: red, blue, and black. There was a total of nine testing sessions with three subjects tested at a particular session. The type of music to be played at a session was selected at random with three sessions used for each of the types of music. At a particular session three subjects were assigned to study the same list of 50 words except that subjects had a different font color. After studying the list for 1 minute the subjects were then asked to recall and write down the words that he/she could remember. The score on this memorization test was the fraction of the 50 words that were correctly remembered.

This is an example of a split plot experiment.

a. What is the whole plot factor? What is the whole plot experimental unit? Give some extraneous variables that contribute to whole plot experimental error.

b. Are the whole plot experimental units arranged in a completely randomized design or a block design? Explain.

c. What is the split plot factor? What is the split plot experimental unit? Give some extraneous variables that contribute to split plot experimental error.

d. Describe the blocking variable(s) used in the study.

e. Give the population effects model for this experiment and describe the terms in the model including the error terms. Give the assumptions associated with the errors.

9.3* Casey Gundersen (Fall 2004) investigated the effects of oven temperature and type of ice cube on the amount of time for the ice cube to melt. The experiment was carried out using nine oven sessions. The temperature used for a particular oven session was randomly selected from one of 250, 300, 350 degrees Fahrenheit, with three sessions per temperature. At each session three ice cubes about of equal size were put into the oven, one per Pyrex bowl. One ice cube was made from bottle water, one from tap water, and one from bottle water with salt. The response variable was the amount of time in seconds that it took for a cube to melt. This is an example of a split plot experiment. The data are given in the following table.

| | | Ice Type | | |
|---|---|---|---|---|
| Oven Temp | | Tap | Bottle | Salt |
| | Run | | | |
| 250 | 1 | 753 | 707 | 525 |
| | 2 | 786 | 728 | 648 |
| | 3 | 650 | 658 | 596 |
| 300 | 1 | 546 | 528 | 567 |
| | 2 | 629 | 598 | 485 |
| | 3 | 665 | 612 | 628 |
| 350 | 1 | 563 | 602 | 484 |
| | 2 | 642 | 521 | 443 |
| | 3 | 608 | 498 | 438 |

a. What is the whole unit factor? What is the whole unit? Give some extraneous variables that contribute to whole unit experimental error.

b. Are the whole units arranged in a completely randomized design or a block design? Explain.

c. What is the split unit factor? What is the split unit? Give some extraneous variables that contribute to split unit experimental error.

d. Give the population effects model for this experiment and describe the terms in the model including the error terms. Give the assumptions associated with the model.

e. Use a statistical computing program to obtain an ANOVA table for the data.

f. Is there evidence of interaction between oven temperature and ice type? Use a 0.10 level of significance.

g. From part (f) there is no evidence of interaction between oven temperature and ice type. Thus test for main effects of oven temperature and ice type. Use a significance level of 0.05 for each type. Make appropriate pairwise comparisons using the Tukey-Kramer confidence intervals with an overall confidence level of 0.95.

h. Check the assumptions of normality and homogeneity of split unit error variance with appropriate plots. Comment.

9.4* Milliken and Johnson [20], page 297 describe an experiment in which a field is divided into two blocks, each with four plots. Each of four fertilizers $(F1, F2, F3, F4)$ is randomly assigned to one of the plots within each block. Each plot is split into two smaller plots. Each smaller plot within the plot is randomly assigned to one of two wheat varieties $(W1, W2)$. The response variable is yield (lbs) of the variety of wheat on the smaller plot. This is an example of a split plot experiment. The yields are given in the following table.

| Block | F1 | | F2 | | F3 | | F4 | |
|---|---|---|---|---|---|---|---|---|
| | $W1$ | $W2$ | $W1$ | $W2$ | $W1$ | $W2$ | $W1$ | $W2$ |
| 1 | 35.4 | 37.9 | 36.7 | 38.2 | 34.8 | 36.4 | 39.5 | 40.0 |
| 2 | 41.6 | 40.3 | 42.7 | 41.6 | 43.6 | 42.8 | 44.5 | 47.6 |

a. What is the whole plot factor? What is the whole plot experimental unit?

b. Are the whole plot experimental units arranged in a completely randomized design or a block design? Explain.

c. What is the split plot factor? What is the split plot experimental unit?

d. Give the population effects model for this experiment and describe the terms in the model including the error terms. Also give the assumptions associated with the model.

e. Use a statistical computing program to obtain an ANOVA table for the yields.

f. Is there evidence of interaction between fertilizer and wheat variety? Use a 0.10 level of significance.

g. From part(f) there was no statistical evidence of interaction between fertilizer and wheat variety. Thus test for fertilizer and wheat variety main effects. Use a significance level of 0.05 for each test. Make appropriate pairwise comparisons using Tukey-Kramer confidence intervals with an overall confidence level of 0.95.

9.5 Megan Ragghiani and Lauren Saunders (Spring 2009) conducted an experiment to compare the amount of salt dissolved in 50 mL of water at three different pH levels and three different temperatures. The three pH levels were 4, 7, and 10 (lower means more acidic). The three temperatures were 4°C, 22°C, and 37°C. The procedure was as follows. An amount of water equal to 150 mL was poured into a container. A pH level was randomly selected from the three and enough hydrochloric acid (HCl) or base sodium hydroxide (NaOH) was added to the container of 150 mL of water to obtain the desired pH level. The 150 mL of water in the container was then divided into three flasks each with 50 mL. The flasks were then cooled or heated to obtain the desired temperature. Once the water in a flask was cooled or heated to the desired temperature a fixed amount of salt was added and the solution thoroughly mixed. A lab procedure was then used to measure the response amount of salt (grams) dissolved in the flask. This procedure was followed 9 times, three times for each of the pH levels.

A table of mean amount of salt absorbed (grams) is given below:

| | Temperature (°C) | | |
|---|---|---|---|
| pH | 4 | 22 | 37 |
| 4 | 16.5 | 16.6 | 17.4 |
| 7 | 15.6 | 16.9 | 17.4 |
| 10 | 16.6 | 17.4 | 17.5 |

a. The design is a split plot design with whole units arranged in a completely randomized design.

i. What are the "whole units" and what is the whole unit factor?

ii. What are the "split units" and what is the split unit factor?

b. Below is a partial ANOVA table.

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Temperature | | 6.58 | | | 0.0058 |
| pH | | 1.29 | | | 0.6201 |
| Temperature*pH | | 1.40 | | | 0.5105 |

Suppose that $MSE$ associated with the whole units is 1.25 with 6 degrees of freedom and $MSE$ associated with the split units is 0.40 with 12 degrees of freedom.

In the ANOVA table above fill in the missing value for df, MS, and F corresponding to Temperature, pH, and Temperature*pH.

c. Note that the interaction between Temperature and pH is not significant at the 0.10 level. Note also that the marginal means for amount of salt at the different temperature are significantly different at the 0.05 level. Calculate the Tukey simultaneous confidence intervals for making pairwise comparisons of the marginal means of amount of salt dissolved for the temperature factor. Use a 95% experiment-wise confidence level. Interpret the endpoints of the intervals within the context of this study.

9.6 In the article "Impact of Chicken Manure and Sowing Methods on Alfalfa (*Medicago sativa L.*) Growth, Forage Yield and Some Quality Attributes" researchers describe an experiment to investigate the effects of manure level and sowing method on forage dry weight using a split plot design. There were four levels of chicken manure (0, 2.5, 5, and 10 tonne $ha^{-1}$) and two levels of sowing method (flat plots and ridged plots) studied in a factorial design. A level of chicken manure was randomly assigned and applied to a 4 x 12 meter rectangular plot of land. The rectangular plot was then subdivided into 2 subplots with the two subplots being randomly assigned to the two sowing methods of alfalfa seeds. After several weeks of growth the alfalfa was cut and several responses measured on each subplot. One response was forage dry weight (tonne $ha^{-1}$) which was the dry weight of the alfalfa plants cut from a small section in the center of a subplot. It was not feasible to measure the dry weight of all plants in the subplot.

a. What are the whole units? What is the whole unit factor?

b. What are the split units? What is the split unit factor?

c. Describe the blocks in this study.

d. Are the whole units arranged in the completely randomized design or in a randomized complete block design? Explain.

e. How many treatments are in this study? The authors describe 3 replications per treatment combination. How many total whole plots did the study use? How many total split plots did the study use?

f. Suppose that the researchers obtained plants from three different small sections of each subplot exposed to a particular manure level and sowing method and measured the dry weight for each of these 3 samples of plants. Explain why these 3 observations on dry weight for each subplot would not represent 3 replications of the treatment combination.

9.7 Emily Shrader, Ashley Sawyer, and Lisa Kleinschmidt (Fall 2009) investigated the effects of liquid type and cup type on the temperature of the liquid 10 minutes after having been heated to 160°F. Three different liquid types were used: water, water + lemon, and water + salt. Two cup types were used, paper and styrofoam. The experiment was carried out over nine liquid heating sessions. At a heating session a randomly selected liquid type was poured into a pot and heated to 160°F. After reaching this temperature the heated liquid was poured into two cups, one paper and one styrofoam, in equal amounts. The response variable was the temperature of the liquid in the cup after 10 minutes of cooling. The temperature data are given in the following table.

| | | Cup Type | |
|---|---|---|---|
| Liquid Type | | Styrofoam | Paper |
| | Session | | |
| Water | 1(7) | 139 | 136 |
| | 2(4) | 144 | 140 |
| | 3(1) | 148 | 141 |
| Water+Lemon | 1(9) | 140 | 138 |
| | 2(3) | 129 | 122 |
| | 3(6) | 126 | 126 |
| Water+Salt | 1(5) | 130 | 125 |
| | 2(8) | 130 | 126 |
| | 3(2) | 131 | 128 |

a. What is the whole plot factor? What is the whole plot experimental unit? Give some extraneous variables that contribute to whole plot experimental error.

b. Are the whole plot experimental units arranged in a completely randomized design or a block design? Explain.

c. What is the split plot factor? What is the split plot experimental unit? Give some extraneous variables that contribute to split plot experimental error.

d. Give the population effects model for this experiment and describe the terms in the model including the error terms.

e. Use a statistical software to obtain an ANOVA table for the temperatures.

f. Is there evidence of interaction between type of cup and liquid type? Use a 0.10 level of significance.

g. From part (f) there is no evidence of interaction between type of cup and liquid type. Thus test for main effects of liquid type and cup type. Use a significance level of 0.05 for each test. Make appropriate pairwise comparisons using the Tukey-Kramer multiple confidence intervals with an overall confidence level of 0.95.

h. Check the assumptions of normality and homogeneity of split plot error variance with appropriate plots. Comment.

9.8 Elizabeth Napoda and Jennifer Sherman (Spring 2009) investigated the effects of type of flour and temperature of cookie dough on the diameter of baked sugar cookies. Cookie dough was made from one of two different types of flour, wheat and rice. After making a batch of cookie dough with a randomly selected type of flour the dough was formed into twelve cookie dough balls. Four of the balls were then put in the freezer, four in the refrigerator, and four kept at room temperature, in all three cases for one hour, before baking. The twelve balls were randomly placed on a cookie sheet and baked for the suggested time on the recipe. The entire experiment was conducted over six mixing/baking sessions, three for each type of flour, with the type of flour used randomly selected. The response variable was the mean diameter (cm) of the four baked sugar cookies prepared at one session with one of the temperature environments. The experimental design is the split plot design. The mean diameters (cm) of cookies are given in the following table.

| | | Temperature | | |
|---|---|---|---|---|
| Type of Flour | | Room | Refrigerator | Frozen |
| | Session | | | |
| Wheat | 1 | 4.39 | 4.19 | 4.46 |
| | 2 | 4.40 | 4.54 | 4.54 |
| | 3 | 4.59 | 4.31 | 4.28 |
| Rice | 1 | 7.60 | 7.61 | 7.61 |
| | 2 | 7.19 | 7.33 | 7.53 |
| | 3 | 7.59 | 7.36 | 7.31 |

a. What is the whole unit factor? What are the whole units? Give some extraneous variables that contribute to whole unit error?

b. Are the whole units arranged in a completely randomized design or a block design? Explain.

c. What is the split unit factor? What are the split units? Give some extraneous variables that contribute to split unit error.

d. Why use the mean diameter of the four dough balls at a particular temperature for the response rather than the diameters of individual dough balls?

e. Give the population effects model for this experiment and describe the terms in the model including the error terms. Give the assumptions associated with the model.

f. Use a statistical software to obtain an ANOVA table for the data.

g. Is there evidence of interaction between type of flour and temperature? Use a 0.10 level of significance.

h. From part (f) there is no evidence of interaction between type of flour and temperature. Thus test for flour type and temperature main effects. Use a significance level of 0.05 for each test. Make appropriate pairwise comparisons using the Tukey-Kramer multiple confidence intervals with an overall confidence level of 0.95.

i. Check the assumptions of normality and homogeneity of split unit error variance with appropriate plots. Comment.

j. Explain how this experiment could have been conducted using a completely randomized design.

9.9 Stephen Clark and Jodie Tsou (Spring 2009) used a split plot design to investigate the effects of different cooling methods and types of container on the temperature of a beverage after 15 minutes of cooling. The cooling methods were ice, ice + water, and ice + water + salt which were put into a styrofoam container. The types of containers were glass bottle, aluminum can, and plastic bottle. The entire experiment was conducted over three days with one complete replication of all treatments in a day. On each of three days three styrofoam ice chests of the same type and size were stocked, one with ice, one with ice + water, and another with ice+water+salt. Within each ice chest three container types were randomly placed in 1/3 compartments of the chest. Each container had the same amount of liquid (beer) which had been stored at room temperature before the cooling. After being placed in the chests the liquid was allowed to cool for 15 minutes. After the 15 minute period of cooling the liquids in the three containers were measured for temperature. The temperature measurements (°F) are given in the table below.

| | | Cooling Method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Day | | Ice | | | Ice+Water | | | Ice+Water+Salt | | |
| | Container | A | P | G | A | P | G | A | P | G |
| 1 | | 50.1 | 61 | 59.8 | 46.2 | 48.1 | 49.8 | 48.8 | 49.1 | 46.6 |
| 2 | | 49.2 | 57 | 62.5 | 47.1 | 43.5 | 45.3 | 49.0 | 45.8 | 51.8 |
| 3 | | 51.2 | 60.2 | 56.1 | 45.2 | 45.9 | 46.9 | 47.9 | 46.3 | 45.9 |

a. What is the whole unit factor? What are the whole units and how many are there? Give some extraneous variables that contribute to whole unit error.

b. Are the whole units arranged in a completely randomized design or a block design? Explain.

c. What is the split plot factor? What are split units and how many are there? Give some extraneous variables that contribute to split unit error.

d. Describe all types of blocking in this experiment.

e. Give the population effects model for this experiment and describe the terms in the model including the error terms. What are the assumptions associated with the model?

f. Use statistical software to obtain an ANOVA table for the temperatures.

g. Is there evidence of interaction between type of coolant and container type? Use a 0.10 level of significance.

h. From part (g) there is evidence of interaction between type of coolant and container type. Using statistical software make three sets of pairwise comparisons of mean temperature across the different container types, one for each level of coolant. Use the Tukey-Kramer multiple confidence intervals with an overall confidence level of 0.95 for each level of coolant.

i. Check the assumptions of normality and homogeneity of split plot error variance with appropriate plots. Comment.

j. Explain how this experiment could have been conducted using a completely randomized design.

# Chapter 10

# Analysis of Designs with Subsampling and Repeated Measures

Designs with **subsampling and measurement units** were discussed briefly in Chapter 1. The measurement unit is the unit upon which the measurement of the response is made. In some cases the measurement unit and the experimental unit are the same. In other cases measurement units are subsamples of the experimental units. For example in animal health studies pens are usually experimental units since treatments are given in the feed or water of a pen and treatments are assigned to pens. There are some responses for which the measurement is made directly on the pen or experimental unit such as feed consumption, and in this case the measurement unit and experimental unit are the same. Other measures may be made on individual animals within the pen rather than at the pen level or experimental unit level, such as degree of sickness or weight gain of the animal. In this case experimental units are pens but measurement units are individual pigs.

Section 10.1 will investigate methods for analyzing the response when there is subsampling of the experimental unit in a one factor completely randomized design.

**Two factor repeated measures designs** are similar to two factor split plot designs from Chapter 9 in that there are two two kinds of experimental units. The levels of one factor, the whole unit factor, are assigned at random to whole units, such as persons, arranged in either a completely randomized or block design. Each whole unit is then measured repeatedly on the response variable over time or space. The second type of experimental unit is a time slot or unit of space corresponding to the repeated measures. The split unit factor is a time or space scale associated with the time slots, an observational factor rather than an experimental factor. Interest is in trends of the response variable over time or space for different levels of the whole unit factor and/or comparisons of trends across the levels of the whole unit factor. The split plot analysis may be used in some cases to analyze data from repeated measures designs. In other cases an alternative analysis must be used. More detail in provided in Section 10.2.

## 10.1 One Factor Studies with Subsampling

### 10.1.1 Example: Completely Randomized Design

Consider an experiment to compare four different mulching methods on moisture of the soil. Twelve available plots are assigned at random to the four different mulches, with 3 plots per mulch. The experimental units are the twelve plots. It is more convenient to take smaller soil samples (or sub-samples) from each plot and measure moisture than measuring moisture for the entire plot. Suppose that three soil samples are selected from each of the plots and the moisture content is measured for each sample. This is an example of a one factor completely randomized design with sub-sampling. The single factor is the type of mulch with four levels. The types of mulch are assigned completely at random to the experimental units, here plots. However the entire plots are not measured for moisture. Three samples of soil were measured for moisture rather moisture being measured on the the entire plot. The individual samples of soil are the measurement units, which are sub-samples of the experimental unit plot. The fact that the values of the response variable are measured from parts of the experimental unit has to be taken into account in the analysis.

### 10.1.2 Example: Completely Randomized Design

In a student project three different brands of cupcake mixes are compared for cupcake height. The experiment is conducted as follows. A brand of cupcake mix is randomly selected and dough is made from a box of mix of that brand. The box is capable of making 12 cupcakes, but only 3 cupcakes are baked at a time. The rest of the dough is frozen for later use. The three cupcakes are baked at a pre-selected temperature and amount of time. The heights of the centers of each of the three cupcakes is measured. This process is repeated for a total of four boxes per brand and 3 brands for a total of 36 cupcakes. Only one box (3 cupcakes from that box)are baked at an oven run. This is an example of a completely randomized design with sub-sampling. The experimental unit is batch of dough made from a box of cup cake baked at a particular oven run. The boxes are assigned completely at random to the 12 oven runs. The 3 cupcakes, with their heights, are a sub-sample or a part of the batch of cupcake mix that could have been baked with the particular box.

### 10.1.3 Example: Randomized Complete Block Design

Oehlert (2000) provides an example of sub-sampling of experimental units in a one factor randomized complete block design. The blocks are 5 Cycad plants.

Three branches on each plant are selected and randomly assigned to three treatments for mealybug. The treatments are 1) water (control), 2) horticultural oil, and 3) fungal spores in water. There are 15 experimental units with 3 per branch. This is an example of Type A blocking (natural grouping) from Chapter 7 on blocking. The response variable is change in number of mealybugs (before treatment - 3 days after treatment) obtained from each of two 3 cm by 3 cm patches on a branch. The two patches on each branch constitute a subsample of the experimental unit, branch, and the two patches are measurement units. Presumably the multiple measurements on a branch are taken for increasing precision of the comparison of the treatments.

### 10.1.4 Model for One Factor CRD with Subsampling

The population effects model for the one factor completely randomized design (CRD) with sub-sampling of the experimental unit is:

$$y_{ijk} = \mu + \alpha_i + \epsilon_{ij} + \eta_{ijk} \tag{10.1}$$

where $i = 1, ..., t$, with $t$ being the number of levels/treatments of the single factor factor A, $j = 1, ..., r$, with $r$ being the number of experimental units at each level of factor A, and $k = 1, ..., n$, with $n$ being the number of measurement units on the $j^{th}$ experimental unit for the $i^{th}$ treatment.

- $y_{ijk}$ is the value of the response variable for the $k^{th}$ measurement unit of the $j^{th}$ experimental unit receiving the $i^{th}$ treatment
- $\mu$ is the grand mean of the response variable averaged over a population of measurement units, all levels of factor A and a population of experimental units.
- $\alpha_i$ is the true effect of the $i^{th}$ level of the factor A on the response variable
- $\epsilon_{ij}$ is the error term associated with the $j^{th}$ experimental unit for the $i^{th}$ level of the factor A, representing the effect of extraneous variables associated with the experimental unit.
- $\eta_{ijk}$ is the sampling error for the $k^{th}$ measurement unit with the $j^{th}$ experimental unit associated with the $i^{th}$ level of A, representing the effect of extraneous variables associated with the measurement unit.

The model assumes that the experimental errors are independent normal random variables each with mean 0 and common variance $\sigma^2_\epsilon$ and that the measurement unit errors are independent normal random variables each with mean 0 and common variance $\sigma^2_\eta$. It is also assumed that experimental errors are independent of measurement unit errors.

The ANOVA table is derived in a manner similar to the derivation for other designs. The observed responses can be partitioned into parts representing the grand mean, the effect of the particular level of factor A, the error associated with the experimental unit, and the error associated with the measurement unit. The sum of squares of the deviations of the observed responses from the grand

Table 10.1: ANOVA Table for One Factor CRD with Subsampling

| Source of Variation | df | SS | MS | F | EMS |
|---|---|---|---|---|---|
| A | t - 1 | SSA | MSA | $MSA/MSE_\epsilon$ | $\sigma_\eta^2 + n\sigma_\epsilon^2 + rnQ_t$ |
| $Error_\epsilon$ | $t(r-1)$ | $SSE_\epsilon$ | $MSE_\epsilon$ | | $\sigma_\eta^2 + n\sigma_\epsilon^2$ |
| $Error_\eta$ | tr(n - 1) | $SSE_\eta$ | $MSE_\eta$ | | $\sigma_\eta^2$ |
| Total(C) | $trn-1$ | SSTotalC | | | |

mean can be partitioned into sums describing variability in the effects of A, experimental unit effects, and measurement unit effects.

$$SSTOT_C = SSA + SSE_\epsilon + SSE_\eta$$

The ANOVA table is given in Table 10.1. Mean squares, MS, are as usual, sums of squares, SS, divided by respective degrees of freedom. Note that the F ratio for testing for A treatment effects uses $MSE_\epsilon$, that is the variation associated with the experimental units rather than $MSE_\eta$, variation associated with the measurement units.

The mathematical derivations of the sums of squares are given below (see [14], p. 162):

$$\begin{aligned} SSA &= rn\sum_{i=1}^{t}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot\cdot\cdot})^2 \qquad (10.2) \\ SSE_\epsilon &= n\sum_{i=1}^{t}\sum_{j=1}^{r}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot})^2 \\ SSE_\eta &= \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij\cdot})^2 \\ SSTotalC &= \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{\cdot\cdot\cdot})^2 \end{aligned}$$

In the above for $SSE_\epsilon$, $\overline{y}_{ij\cdot}$ is the mean of response for the $j^{th}$ experimental unit of the $i^{th}$ treatment, averaged over the measurement units for that experimental unit, while the value $\overline{y}_{i\cdot\cdot}$ is the mean of the response for the $i^{th}$ treatment. Thus $SSE_\epsilon$ measures variation in the means of the experimental units from their respective treatment means, pooled across the different treatments. $SSE_\eta$ measures variation in the individual response values for the measurement units from the respective experimental unit mean pooled across all experimental units and treatments.

The Tukey-Kramer confidence intervals with overall confidence level $1-\alpha$ for making pairwise comparisons of the treatment means of the response is

$$(\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}) \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}}\sqrt{MSE_\epsilon}\sqrt{\frac{1}{rn} + \frac{1}{rn}}$$

where $\sqrt{MSE_\epsilon}\sqrt{\frac{1}{rn}+\frac{1}{rn}}$ is the standard error of the difference in two treatment means, $\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}$, and $q_{\alpha;\nu,t}$ is the upper $\alpha$ probability point from the Studentized range distribution. Here $\nu$ refers to the degrees of freedom associated with $MSE_\epsilon$, experimental unit mean squared error and $t$ refers to the number of levels (treatments) of factor A.

### 10.1.5 Example: Analysis of a CRD with subsampling

**Example 10.1** *This example is based on the cupcake baking experiment by Hillary Hayes and Tristin Millette (Fall 2012). Three flavors of cupcakes (Chocolate, Strawberry and Funfetti) were compared in terms of height of cupcake measured at the center of the cupcake. The experiment was conducted as follows. A box of cupcake mix of a randomly selected flavor was used to make a batch of dough. Enough dough was selected to make four cupcakes from the batch. The oven was preset to a predetermined heat level. The four cupcakes were randomly positioned on the middle rack of the oven. The cupcakes were cooked for* 23 *minutes. The cupcakes were removed from the oven and center of the cupcake was measured to the nearest millimeter. This process was repeated for a total of* 12 *oven runs,* 4 *runs for each of the flavors.*

*The design is a completely randomized design. The experimental units are batches of cupcakes of different flavors, the order of baking randomly assigned. Measurement units for each batch are the* 4 *cupcakes per batch measured for height. The heights are given in the following table.*

The population effects model for the cupcake heights is:

$$y_{ijk} = \mu + \alpha_i + \epsilon_{ij} + \eta_{ijk} \tag{10.3}$$

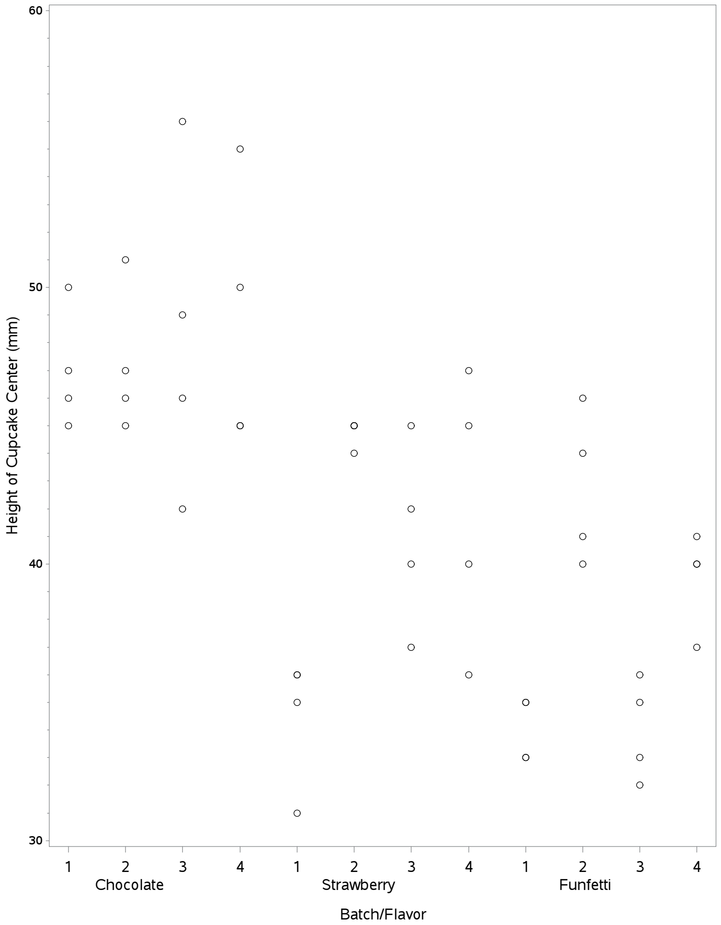
where $i = 1, ..., t = 3$, with $i$ being an index on the flavor, $j = 1, ..., r = 4$, with $j$ being an index on the replicate batch at each flavor, and $k = 1, ..., n = 4$, with $k$ being an index on the measurement unit (cupcake) for the $j^{th}$ replicate batch for the $i^{th}$ flavor.

- $y_{ijk}$ is the value of cupcake height for the $k^{th}$ cupcake with the $j^{th}$ replicate batch using the $i^{th}$ flavor.
- $\mu$ is the grand mean of cupcake heights averaged over a population of cupcakes, all flavor levels and a population of batches.
- $\alpha_i$ is the true effect of the $i^{th}$ flavor on cupcake height
- $\epsilon_{ij}$ is the error term associated with the $j^{th}$ replicate batch of the $i^{th}$ flavor, representing the effect of extraneous variables associated with the batch, box of cupcake mix, run of the oven.
- $\eta_{ijk}$ is the error for the $k^{th}$ cupcake with the $j^{th}$ replicate batch using the $i^{th}$ flavor, representing the effect of extraneous variables associated with the cupcake, such as variation within the cupcake mix, variation of temperature within the oven.

Table 10.2: Heights (mm) of Cupcakes

| | | Batch (j) | | | |
|---|---|---|---|---|---|
| Flavor(i) | | 1 | 2 | 3 | 4 |
| Chocolate (1) | Cupcake(k) | | | | |
| | 1 | 46 | 51 | 42 | 55 |
| | 2 | 50 | 46 | 56 | 50 |
| | 3 | 47 | 47 | 49 | 45 |
| | 4 | 45 | 45 | 46 | 45 |
| | $\overline{y}_{1j\cdot}$ | 47.0 | 47.3 | 48.3 | 48.8 |
| Strawberry (2) | Cupcake(k) | | | | |
| | 1 | 36 | 45 | 42 | 40 |
| | 2 | 31 | 44 | 37 | 36 |
| | 3 | 35 | 45 | 40 | 45 |
| | 4 | 36 | 45 | 45 | 47 |
| | $\overline{y}_{2j\cdot}$ | 34.5 | 44.8 | 41.0 | 42.0 |
| Funfetti (3) | Cupcake(k) | | | | |
| | 1 | 33 | 44 | 36 | 37 |
| | 2 | 35 | 46 | 35 | 41 |
| | 3 | 35 | 41 | 32 | 40 |
| | 4 | 33 | 40 | 33 | 40 |
| | $\overline{y}_{3j\cdot}$ | 34.0 | 42.8 | 34.0 | 39.5 |

Figure 10.1: Plot of Cupcake Heights versus Batch and Flavor



The model assumes that the experimental errors are independent normal random variables each with mean 0 and common variance $\sigma_\epsilon^2$ and that the cupcake errors are independent normal random variables each with mean 0 and common variance $\sigma_\eta^2$. It is also assumed that the batch/oven run errors are independent of cupcake errors.

A dot plot is given in Figure 10.1. It appears that heights for chocolate flavored cupcakes are greater than heights for the other two flavors.

The ANOVA table is given in Table 10.3. The results of the F test ($F = 8.72, \text{P-value} = 0.0078$) indicate statistically significant differences in mean heights of cupcakes among the three flavors at the 0.05 level of significance.

Table 10.4 gives the Tukey-Kramer pairwise comparisons of the flavors on cupcake height. The differences in mean heights between the chocolate flavor

Table 10.3: ANOVA Table for Cupcake Example

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Flavor | 2 | 888.7 | 444.3 | 8.72 | 0.0078 |
| Error (Batch) | 9 | 458.6 | 4.8 | | |
| Error (Cupcake) | 36 | 379.8 | 10.5 | | |
| Total(C) | 47 | 1727.0 | | | |

Table 10.4: Tukey Pairwise Comparisons of Flavors

| Flavor | Flavor | Mean Difference | Std.Error | DF | t Value | P-value | LCL | UCL |
|---|---|---|---|---|---|---|---|---|
| Chocolate | Funfetti | 10.25 | 2.5 | 9 | 4.06 | 0.0072 | 3.20 | 17.30 |
| Chocolate | Strawberry | 7.25 | 2.5 | 9 | 2.87 | 0.0441 | 0.20 | 14.30 |
| Funfetti | Strawberry | -3.00 | 2.5 | 9 | -1.19 | 0.4884 | -10.05 | 4.05 |

and each of the Funfetti and Strawberry flavors are statistically significant. There is not enough evidence of a difference in mean height between the Funfetti and Strawberry flavors.

### 10.1.6 Note on Subsampling

1. The design for this example is balanced. That is there are the same number of experimental units per treatment and the same number of measurement units per experimental unit. For balanced designs the same results regarding the comparison of the treatments can be obtained by using the methods of Chapter 4 with the mean of the response averaged over all measurement units.

2. If measurement units are treated as experimental units and the data analyzed using the method of Chapter 4 then mean squared error will generally be underestimated and results of the F test and the multiple comparison procedures will be invalid. In the cupcake example this means that an incorrect analysis would be conducted if the individual cupcakes were treated as experimental units (16 experimental units per flavor) and cupcake heights analyzed using the method of Chapter 4.

## 10.2 Repeated Measures Designs

In a **two factor repeated measures design** subjects are assigned at random to treatments (levels of one factor, say A), and then the response variable is observed for each of the subjects on several time slots or occasions after treatment. Some measure of time associated with the time slots or occasions form the second factor, B. The purpose of the study is to study the nature of the trend in the values of the response variable across time for the individual treatments and to compare trends for the different treatments.

The design is similar to the split plot design of Chapter 9 with whole units being the subjects or objects receiving the treatments and with split units being time slots or occasions associated the subjects when the repeated measures on the response are taken. In the standard split plot design of Chapter 9 the split units are randomly assigned to levels of the split unit factor. In the repeated measures design the levels of the split unit factor are measures of time, such as days, hours, weeks, etc., which cannot be assigned at random to time slots or occasions. The measure of time are inherent characteristics of the time slots. This fact has ramifications on when a split plot design analysis can or cannot be used to analyze repeated measures data.

It should be noted that some textbooks (see [15], [25], [29]) use the term repeated measures not only for studies where subjects are repeatedly measured on the response over time after receiving a treatment but also designs where subjects or objects are reused over different time slots, receiving **different** treatments at different time slots over time. Chapter 10 of this text concentrates on the former type of study. The latter type of study was considered in Chapter 9.

### 10.2.1 Example of Repeated Measures Study - Whole Units in CRD

This example illustrates the use of a repeated measures design where the whole units are assigned completely at random to the levels of the whole unit factor. In the study described in the article "Efficacy and safety of eperisone in patients with low back pain: a double blind randomized study" (*European Review for Medical and Pharmacological Sciences*, 2008; 12: 229-235) researchers assigned 160 patients with low back pain to one of two medications, eperisone 100 mg three times daily or thiocolchicoside 8 mg twice daily for 12 consecutive days, with 80 patients per medication group. Analgesic activity of the two medications was evaluated at Day 0 before start of medication and on Days 3, 7, and 12 days while on treatment using a 100-mm visual analogue scale (VAS). Whole units are the 160 patients randomly assigned to the two medication groups. Split units are the time slots/days at which observations on the VAS scale were determined. The split unit factor is a time variable with levels of 0, 3, 7, and 12 days after medication started.

### 10.2.2 Example of Repeated Measures Study - Whole Units Blocked

This example illustrates a repeated measures study where the whole units are blocked first and then treatments are assigned at random within blocks. In the article "Acute Effects of a Caffeine-Taurine Energy Drink on Repeated Sprint Performance of American College Football Players (*International Journal of Sport Nutrition and Exercise Metabolism*, 2012, 22, 109-116) each of twenty football players ran two sets of 6 35(m) sprints on each of two separate days, with a 10 second recovery between sprints. One one day the player drank an energy drink (AdvoCare Spark) before the sprints and on the other day a placebo drink. The order of the drinks on the two days was random. One of the response variables was time (seconds)to complete the sprint. The design is similar to a split plot design with whole units blocked. The whole units are the larger time slots, days, associated with the two sets of 6sprints. These are blocked by sprinter with each sprinter providing two days of sprinting. The split units for each whole unit are the 6 smaller time slots that correspond to the 6 individual sprints which are labelled by the order in which they occur (1, 2, 3, 4, 5, 6). The whole unit factor is type of drink (energy or placebo) assigned at random to the two whole units or days for each sprinter. The split unit factor is the time order of the individual sprints. This is a repeated measures design since measurements (sprint times) are collected repeatedly over time for each sprinter after being treated with a type of drink.

### 10.2.3 Example of Repeated Measures Study over Space

Rather than repeated measures being taken over time in some studies the repeated measures are over space. Kutner, Nachtsheim, Neter, and Li ([15], page 1149), describe a study to compare blood flow in rats at five different parts of the body (Bone, Brain, Skin, Muscle, Heart) without and with exercise. All eight rats were injected intravenously with radioactive microspheres to determine blood flow. Four rats, randomly selected from the eight were exercised on a treadmill (exercise) for 15 minutes. The other four rats were put on the treadmill but it was not turned on. After the exercise the rats were sacrificed, tissue from the five parts of the body harvested, and blood flow determined based on the microspheres. The whole units are the rats who were assigned at random to one of the levels (exercise, not exercise) of the whole unit factor. The split units are spatial units, that is the tissues for each of the rats. The split unit factor is the type of tissue. The levels of tissue type are not assigned at random to the harvested tissues. There are 5 repeated measures on each rat corresponding to the five harvested tissues.

### 10.2.4 Example: Split Plot Design, Not Repeated Measures

Recall the split plot design for the class project of John Szarka and Zamda Lumbi (Fall 2004) used in Chapter 9. They were interested in investigating the effects of type of flour (white, wheat, bread) and length of time in oven (5, 10, 15 minutes) on the change in height of dough after baking. Three rolls of dough were made from each type of flour for a total of nine rolls. Each roll was made using the same ingredients except for the type of flour. Each roll was divided into 3 equal parts and the 3 parts put into an oven. One part was baked at 5 minutes, another part at 10 minutes, and another for 15 minutes. Thus one run of the oven involved one roll (3 parts). The type of flour used for a particular roll and run of the oven was selected at random. The 3 parts of the roll were assigned at random to locations in the oven and time of baking. At the end of the 5, 10, and 15 minute periods, the appropriate parts were taken out of the oven and measured for height change. The split unit factor was elapsed time in the oven but the elapsed times corresponded to different split units, here parts, not the same experimental unit. Thus the original experiment was not a repeated measures study. A modification of this experiment for repeated measures would be where each part is left in the oven the entire 15 minutes, with repeated measures of change in height on each part taken at 5, 10, and 15 minutes after the start of baking.

### 10.2.5 Equal Correlation Between Response Among Pairs of Repeated Measures

The assumptions for the split plot design models of Chapter 9 are that 1) responses have equal variances, 2) responses are independent for two split units from different whole units, and 3) responses are correlated for two responses from the same whole unit, with that correlation being the same for different pairs of split units. One justification for the equal correlation assumption is the fact that the levels of the split unit factor are assigned at random to the split units. The assumption of equal variance can be checked with plots.

In the repeated measures design in this chapter the levels of the split unit factor (time marker) are not assigned at random to the split units or time slots. Correlation may exist on the response variable between pairs of time slots and the correlation may differ depending on the pair of time slots. In addition variances of the response at the time slots may depend on time. The following repeated measures example illustrates.

**Example 10.2** *In the article "Problems in the Analysis of Growth and Wear Curves" (*Biometrics 6, 262-289*) the author gives growth rate data for three groups of rats: control, Thyroxin, and Thioruacil in the drinking water.*

*The weights of the rats at Weeks 0, 1, 2, 3, 4 after treatment are given in Table 10.5*

Line plots are given in Figure 10.2, Figure 10.3, and Figure 10.4. Lines connect the weights over time for individual rats. It is evident that rats weights increase over time with perhaps greater variability in the latter weeks.

An interaction plot is given in Figure 10.5. There is evidence of interaction between Treatment and Week, with control and thyroxin rats having similar growth rates, but thiourac rats having a lower growth rate.

Treatment, marginal, and grand mean for weights are provided in Table 10.6.

Standard deviations for the weights at each time period for each treatment are given in Table 10.7.

Clearly variation in weights is increasing over time regardless of the group. Variation across groups is fairly similar at each time point.

Table 10.5: Rat weights for Three Groups

| | | | | Week | | |
|---|---|---|---|---|---|---|
| Group | | 0 | 1 | 2 | 3 | 4 |
| | Rat | | | | | |
| Control | 1 | 57 | 86 | 114 | 139 | 172 |
| | 2 | 60 | 93 | 123 | 146 | 177 |
| | 3 | 52 | 77 | 111 | 144 | 185 |
| | 4 | 49 | 67 | 100 | 129 | 164 |
| | 5 | 56 | 81 | 104 | 121 | 151 |
| | 6 | 46 | 70 | 102 | 131 | 153 |
| | 7 | 51 | 71 | 94 | 110 | 141 |
| | 8 | 63 | 91 | 112 | 130 | 154 |
| | 9 | 49 | 67 | 90 | 112 | 140 |
| | 10 | 57 | 82 | 110 | 139 | 169 |
| Thyroxin | 1 | 59 | 85 | 121 | 146 | 181 |
| | 2 | 54 | 71 | 90 | 110 | 138 |
| | 3 | 56 | 75 | 108 | 151 | 189 |
| | 4 | 59 | 85 | 116 | 148 | 177 |
| | 5 | 57 | 72 | 97 | 120 | 144 |
| | 6 | 52 | 73 | 97 | 116 | 140 |
| | 7 | 52 | 70 | 105 | 138 | 171 |
| Thiouracil | 1 | 61 | 86 | 109 | 120 | 129 |
| | 2 | 59 | 80 | 101 | 111 | 122 |
| | 3 | 53 | 79 | 100 | 106 | 133 |
| | 4 | 59 | 88 | 100 | 111 | 122 |
| | 5 | 51 | 75 | 101 | 123 | 140 |
| | 6 | 51 | 75 | 92 | 100 | 119 |
| | 7 | 56 | 78 | 95 | 103 | 108 |
| | 8 | 58 | 69 | 93 | 116 | 140 |
| | 9 | 46 | 61 | 78 | 90 | 107 |
| | 10 | 53 | 72 | 89 | 104 | 122 |

Figure 10.2: Plot of Rat Weight versus Week: Control

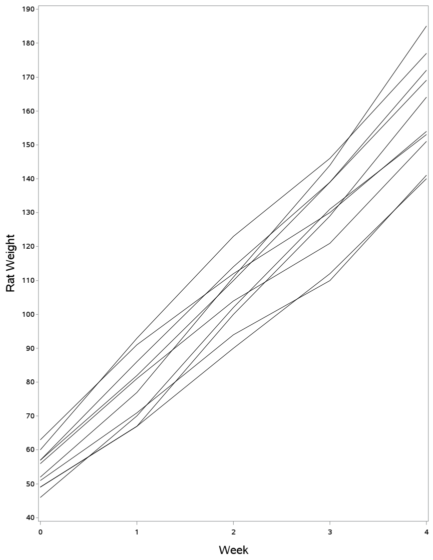


Figure 10.3: Plot of Rat Weight versus Week: Thyroxin

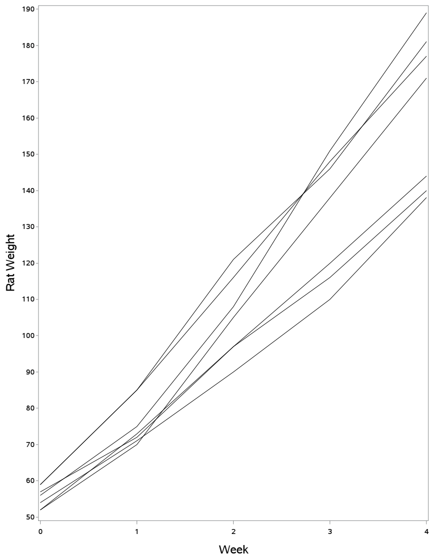


Figure 10.4: Plot of Rat Weight versus Week: Thyroxin

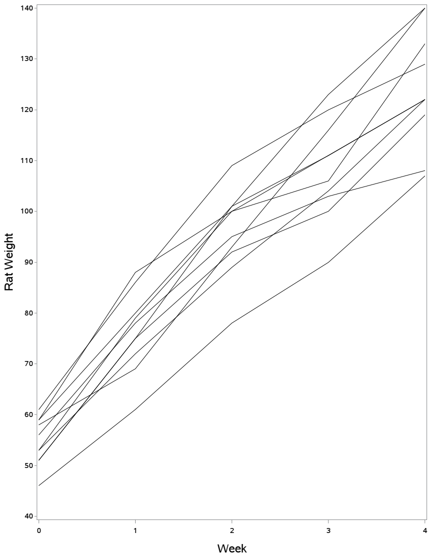


Figure 10.5: InteractionPlot

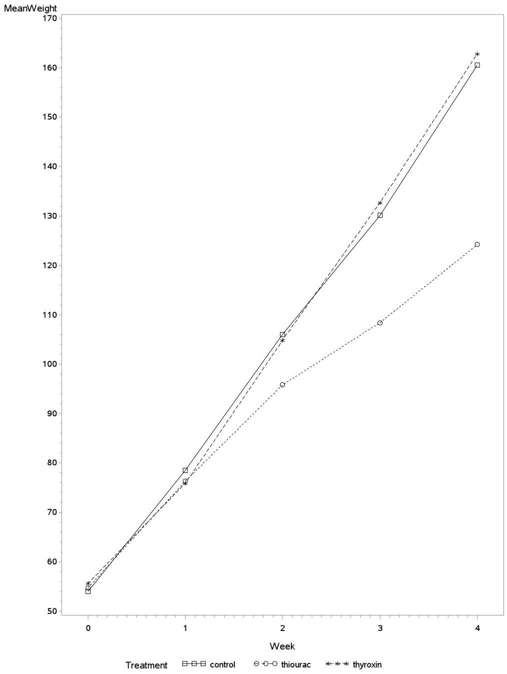


Table 10.6: Weight Means: Rat Example

| | | | Week | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | $\overline{y}_{i..}$ |
| Group | | | | | | |
| Control | 54.0 | 78.5 | 106.0 | 130.1 | 160.6 | 105.8 |
| Thiourac | 54.7 | 76.3 | 95.8 | 108.4 | 124.2 | 91.9 |
| Thyroxin | 55.6 | 75.9 | 104.9 | 132.7 | 162.9 | 106.4 |
| $\overline{y}_{.j.}$ | 54.7 | 77.0 | 102.9 | 125.4 | 152.0 | |
| | | | | | | $\overline{y}_{...} = 49.8$ |

Table 10.7: Rat Weight Standard Deviations

| | | | Week | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| Group | | | | | |
| Control | 5.4 | 9.6 | 9.9 | 12.6 | 15.2 |
| Thiourac | 4.7 | 7.9 | 8.5 | 9.9 | 11.5 |
| Thyroxin | 3.0 | 6.4 | 11.1 | 17.0 | 21.5 |

Correlation coefficient for weights at two different time periods are provided in Table 10.8 for each group. Thus, for example, the correlation between the weights at Week 1 and 3 for the control rats is 0.59.

An inspection of Table 10.8 indicates that the correlations are decreasing the further apart in time. Figure 10.6 is a plot of the correlations versus lag number. Lag refers to the difference in number of weeks on which the two correlations are calculated. For example the correlation coefficient between the weights at Week 1 and Week 3 for the control rats has lag 2. The plot vividly indicates that correlations generally decrease as lag increases.

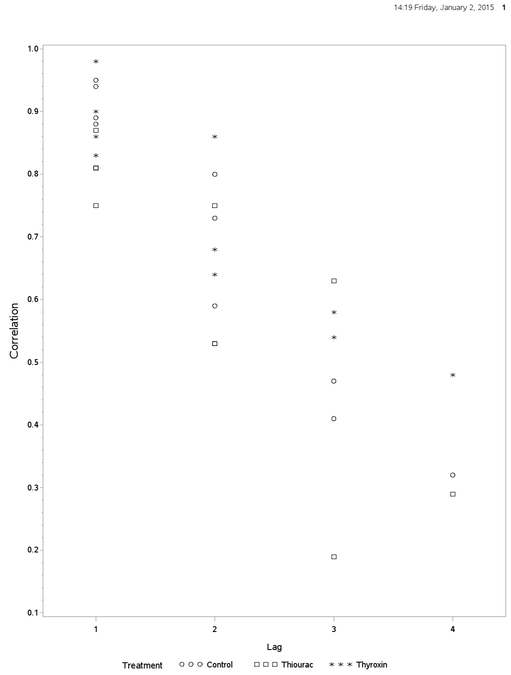
If the split plot assumptions of equal variance and equal correlation holds, called **compound symmetry**, then the analysis of data from a repeated measures design can be validly carried out using the split plot analysis method from Chapter 9. Example 10.3 will illustrate. The conditions of compound symmetry obviously do not hold for the rat weight data and therefore a split plot analysis is not valid.

A condition, called the Huynh-Feldt Condition, less stringent than compound symmetry for the validity of the split plot analysis for repeated measures, is that differences in the response between pairs of repeated measures have the same standard deviation. The condition of compound symmetry satisfies the Huynh-Feldt condition. An informal approach to checking for validity of the split plot approach is to calculate differences in repeated measures for each pair

Table 10.8: Pearson Correlations: Rat Weight Example

| | | | | Week | | |
|---|---|---|---|---|---|---|
| Group | | | | | | |
| Control | Week | 0 | 1 | 2 | 3 | 4 |
| | 0 | 1.00 | 0.95 | 0.73 | 0.41 | 0.32 |
| | 1 | | 1.00 | 0.88 | 0.59 | 0.47 |
| | 2 | | | 1.00 | 0.89 | 0.80 |
| | 3 | | | | 1.00 | 0.94 |
| | 4 | | | | | 1.00 |
| Thiourac | Week | 0 | 1 | 2 | 3 | 4 |
| | 0 | 1.00 | 0.75 | 0.75 | 0.63 | 0.29 |
| | 1 | | 1.00 | 0.87 | 0.53 | 0.19 |
| | 2 | | | 1.00 | 0.81 | 0.53 |
| | 3 | | | | 1.00 | 0.81 |
| | 4 | | | | | 1.00 |
| Thiourac | Week | 0 | 1 | 2 | 3 | 4 |
| | 0 | 1.00 | 0.83 | 0.68 | 0.54 | 0.48 |
| | 1 | | 1.00 | 0.86 | 0.64 | 0.58 |
| | 2 | | | 1.00 | 0.90 | 0.86 |
| | 3 | | | | 1.00 | 0.99 |
| | 4 | | | | | 1.00 |

Figure 10.6: Plot of Rat Weight Correlations versus Lag



within each level of the whole unit factor and compare standard deviations. If the standard deviations are similar then the split plot approach is valid. Statistical programs also have results of statistical tests for the Huynh-Feldt Condition. The test is beyond the scope of this text.

### 10.2.6 An Example of a Split Plot Analysis of a Repeated Measures Study

**Example 10.3** *The data in this example is fictitious but is loosely based on the study "Sugars and satiety: does the type of sweetener make a difference?" (American Journal of Clinical Nutrition [2007]: Vol. 86, pages 116-23). Three drinks are used here whereas there were five drinks in the study. Times of observations are altered. The hunger levels (VAS 100 mm scale) were simulated based on relationships observed in the article. Suppose that* 30 *subjects are assigned at random to drink one of three drinks: milk, diet cola, or no drink at 9:30 a.m. in the morning, approximately* 1 *hour after consuming breakfast. The subjects then complete a hunger level VAS scale four times, starting at 10:30 and ending at noon, before lunch.*

*The hunger levels are given in Table 10.9*

This is an example of a split plot design where whole units (subjects) are assigned completely at random to the levels of the whole unit factor, beverage

Table 10.9: Hunger Levels for Beverage Experiment

| | | Time | | | |
|---|---|---|---|---|---|
| Beverage | | 1 | 2 | 3 | 4 |
| | Subject | | | | |
| Milk | 1 | 28 | 39 | 49 | 56 |
| | 2 | 26 | 37 | 49 | 51 |
| | 3 | 27 | 39 | 51 | 54 |
| | 4 | 24 | 40 | 51 | 55 |
| | 5 | 27 | 40 | 51 | 53 |
| | 6 | 27 | 39 | 48 | 55 |
| | 7 | 28 | 39 | 50 | 57 |
| | 8 | 26 | 40 | 53 | 56 |
| | 9 | 26 | 38 | 46 | 54 |
| | 10 | 26 | 38 | 50 | 56 |
| Coke | 11 | 26 | 42 | 54 | 63 |
| | 12 | 26 | 41 | 54 | 60 |
| | 13 | 26 | 43 | 55 | 64 |
| | 14 | 27 | 43 | 55 | 62 |
| | 15 | 29 | 44 | 56 | 64 |
| | 16 | 30 | 43 | 55 | 65 |
| | 17 | 28 | 42 | 54 | 62 |
| | 18 | 27 | 40 | 55 | 62 |
| | 19 | 27 | 41 | 55 | 62 |
| | 20 | 28 | 42 | 56 | 63 |
| Control | 21 | 50 | 55 | 63 | 70 |
| | 22 | 50 | 54 | 63 | 67 |
| | 23 | 50 | 56 | 64 | 71 |
| | 24 | 51 | 56 | 64 | 69 |
| | 25 | 53 | 57 | 65 | 71 |
| | 26 | 54 | 56 | 64 | 72 |
| | 27 | 52 | 55 | 63 | 69 |
| | 28 | 51 | 53 | 64 | 69 |
| | 29 | 51 | 54 | 64 | 69 |
| | 30 | 52 | 55 | 65 | 70 |

type.

The split unit/plot factor is is a time scale, here denoted by 1, 2, 3, and 4, for the occasions when the hunger levels were measured. The time variable is an observational factor whose levels are not assigned to occasions.

The model for this data is the same model from Chapter 9:

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{10.4}$$

with $i = 1(\text{milk}), i = 2(\text{diet cola}), i = a = 3(\text{control})$ indexing type of beverage, $j = 1(10{:}30), j = 2(11{:}00), j = 3(11{:}30), j = b = 4(12{:}00)$ indexing time, and $k = 1, 2, 3..., 10(n)$ indexing the subject associated with a particular beverage, and

- $y_{ijk}$ is the observation on hunger (VAS scale, 100mm) at the $i^{th}$ level of beverage type, $k^{th}$ subject nested within the $i^{th}$ level of beverage type, and $j^{th}$ level of time of measurement.
- $\mu$ is the grand mean of hunger averaged over a population of subjects, all levels of beverage type, and all levels of times of observation.
- $\alpha_i$ is the true effect of the $i^{th}$ level of the beverage type on hunger.
- $\epsilon^w_{k(i)}$ is the error term for the the $k^{th}$ subject nested within the $i^{th}$ level of the beverage type, representing the effect of extraneous variables associated with the subject, such as varying degrees of initial hunger, metabolism, etc.
- $\beta_j$ is the true effect of the $j^{th}$ level of time on hunger
- $\alpha\beta_{ij}$ is the true interaction effect on hunger of the $i^{th}$ level of beverage type and the $j^{th}$ level of time.
- $\epsilon^s_{ijk}$ is the error term for the split unit, here time slot, associated with the $i^{th}$ level of beverage type, $k^{th}$ subject nested under the $i^{th}$ level of beverage type, and the $j^{th}$ level of time, representing the effect of extraneous variables for a particular time slot, such as distractions, etc.

The model assumes that the "subject" errors, $\epsilon^w_{k(i)}$, are independent normal random variables each with mean 0 and common variance $\sigma^2_w$ and that the "time slot" errors, $\epsilon^s_{ijk}$, are independent normal random variables each with mean 0 and common variance $\sigma^2_s$. It is also assumed that a "subject" error is independent of a "time slot" error.

While the errors are all independent of one another the model hypothesizes that the observations on hunger for the time slots of a particular subject are equally correlated across all pairs of observations. This condition will be checked with correlation coefficients.

Line plot of hunger level across time for the different groups are given in Figure 10.7, Figure 10.8, and Figure 10.9. As expected hunger levels increase over time. There is no evidence that variation in hunger level changes over time.

Treatment, marginal, and grand mean for hunger are provided in Table 10.10.

Standard deviations for the hunger levels at each time period for each beverage are given in Table 10.11. There is no evidence that variation depends on time or beverage group.

Figure 10.7: Hunger Level versus Time: Milk Group

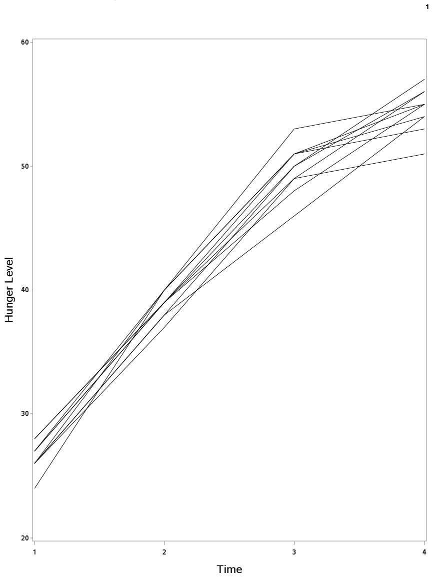


Figure 10.8: Hunger Level versus Time: Cola Group

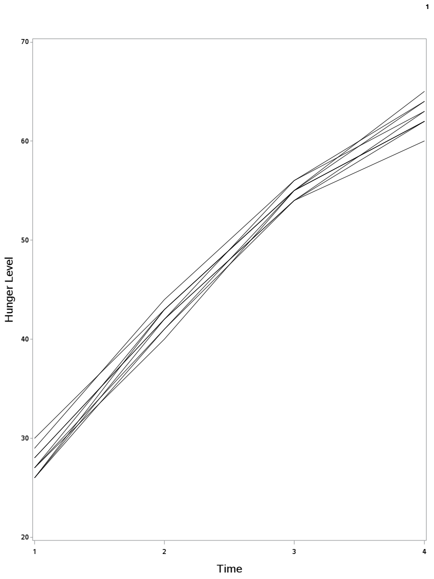


Figure 10.9: Hunger Level versus Time: Control Group

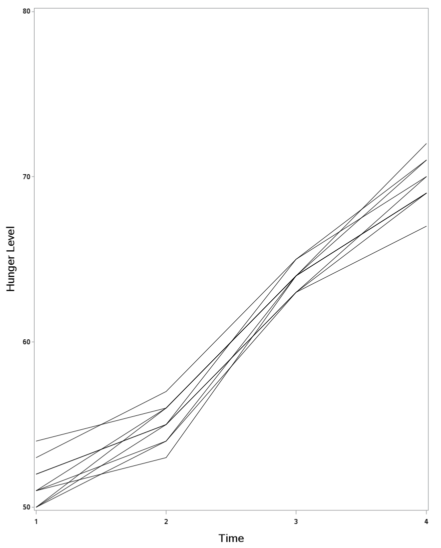


Table 10.10: Hunger Means: Satiety Example

| | Time | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | $\overline{y}_{i..}$ |
| Beverage | | | | | |
| Milk | 26.5 | 38.9 | 49.8 | 54.6 | 42.5 |
| Coke | 27.4 | 42.1 | 54.9 | 62.7 | 46.8 |
| Control | 51.4 | 55.1 | 63.9 | 69.7 | 60.0 |
| $\overline{y}_{.j.}$ | 41.7 | 43.1 | 44.0 | 62.3 | |
| | | | | | $\overline{y}_{...} = 49.8$ |

Table 10.11: Hunger Standard Deviations: Satiety Example

| | Time | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Beverage | | | | |
| Milk | 1.2 | 1.0 | 1.9 | 1.7 |
| Coke | 1.3 | 1.2 | 0.7 | 1.4 |
| Control | 1.3 | 1.2 | 0.7 | 1.4 |

Table 10.12: Pearson Correlations: Satiety Example

| | | Correlation | | | |
|---|---|---|---|---|---|
| Beverage | | | | | |
| Milk | | Time1 | Time2 | Time3 | Time4 |
| | Time1 | 1.00 | -0.05 | 0.15 | 0.28 |
| | Time2 | | 1.00 | 0.62 | 0.37 |
| | Time3 | | | 1.00 | 0.11 |
| | Time4 | | | | 1.00 |
| Coke | | Time1 | Time2 | Time3 | Time4 |
| | Time1 | 1.00 | 0.45 | 0.49 | 0.59 |
| | Time2 | | 1.00 | 0.39 | 0.67 |
| | Time3 | | | 1.00 | 0.50 |
| | Time4 | | | | 1.00 |
| Control | | Time1 | Time2 | Time3 | Time4 |
| | Time1 | 1.00 | 0.45 | 0.49 | 0.59 |
| | Time2 | | 1.00 | 0.39 | 0.67 |
| | Time3 | | | 1.00 | 0.50 |
| | Time4 | | | | 1.00 |

Pearson correlation coefficients are provided in Table 10.12 and a plot of correlations versus lag number is given in Figure 10.10. There is no evidence of correlation depending on lag.

An interaction plot is given in Figure 10.11. There is evidence of interaction between beverage and time.

The ANOVA table for the Satiety/Beverage experiment is given in Table 10.13. Note that there is evidence of interaction between type of beverage and time ($F = 137.7, \text{P-value} < 0.0001$) at the 0.10 level. Thus the rates of increase in hunger across time are not the same for the three beverage types.

Tukey-Kramer pairwise comparisons of the beverages at each of the times is given in Table 10.14.

Table 10.13: ANOVA Table for Satiety Experiment

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Beverage | 2 | 6708.7 | 3354.3 | 940.0 | < 0.0001 |
| Error (Subject(Beverage)) | 27 | 96.4 | 3.6 | | |
| Time | 3 | 13013.0 | 4337.8 | 3919.3 | < 0.0001 |
| Beverage*Time | 6 | 914.5 | 152.4 | 137.7 | < 0.0001 |
| Error (Time slot) | 81 | 89.7 | 1.1 | | |

Figure 10.10: Hunger Correlation Plot

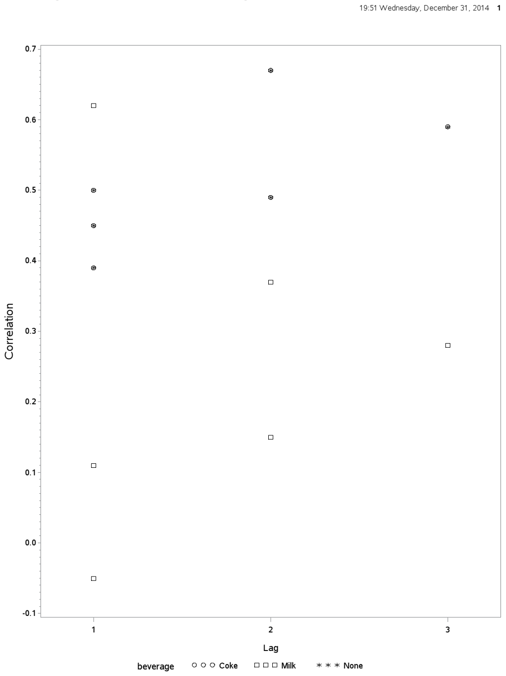
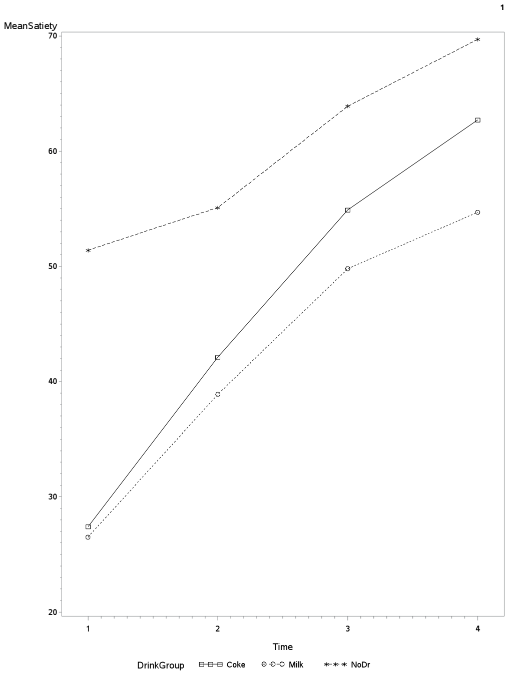


Table 10.14: Tukey-Kramer Pairwise Comparisons of Drinks at each Time

| Time | Drink | Drink | Mean Difference | Std.Error | df | t Value | P-value | LCL | UCL |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Coke | Milk | 0.9 | 0.59 | 78.1 | 1.53 | 0.2808 | -0.50 | 2.31 |
| 1 | Coke | Control | -24.0 | 0.59 | 78.1 | -40.9 | <.0001 | -25.4 | -22.6 |
| 1 | Milk | Control | -24.9 | 0.59 | 78.1 | -42.4 | <.0001 | -26.3 | -23.5 |
| 2 | Coke | Milk | 3.2 | 0.59 | 78.1 | 5.45 | <.0001 | 1.8 | 4.6 |
| 2 | Coke | Control | -13.0 | 0.59 | 78.1 | -22.2 | <.0001 | -14.4 | -11.6 |
| 2 | Milk | Control | -16.2 | 0.59 | 78.1 | -27.6 | <.0001 | -17.6 | -14.8 |
| 3 | Coke | Milk | 5.1 | 0.59 | 78.1 | 8.69 | <.0001 | 3.70 | 6.5 |
| 3 | Coke | Control | -9.0 | 0.59 | 78.1 | -15.53 | <.0001 | -10.4 | -7.6 |
| 3 | Milk | Control | -14.1 | 0.59 | 78.1 | -24.02 | <.0001 | -15.5 | -12.7 |
| 4 | Coke | Milk | 8.1 | 0.59 | 78.1 | 13.80 | <.0001 | 6.7 | 9.5 |
| 4 | Coke | Control | -7.0 | 0.59 | 78.1 | -11.93 | <.0001 | -8.4 | -5.6 |
| 4 | Milk | Control | -15.1 | 0.59 | 78.1 | -25.73 | <.0001 | -16.5 | -13.7 |

Figure 10.11: InteractionPlot



## 10.2.7 General Conditions Necessary for Valid Split Plot Analysis of Repeated Measures

Compound symmetry is a special case of a more general condition on repeated measures correlations and variances that will guarantee that the split plot analysis is valid. Huynh and Feldt in 1970 ("Conditions under which mean square ratios in repeated measures designs have exact F-distributions," *Journal of the American Statistical Association* Vol. 65, 1582-1589) showed that the split plot analysis is valid if variances of all possible pairs of the response variable taken at different times, say $y_i$ and $y_j$, are the same within and between all treatment groups. This condition, called the Huynh-Feldt condition, holds under the compound symmetry structure noted earlier. It does not hold if correlations are decreasing with increasing lag as with the rat weight data.

The Huynh-Feldt condition can be checked informally by calculating sample variances or standard deviations of the difference in the response for all possible time pairs and comparing. Sample standard deviations for the Satiety example are given in Table 10.15.

Since we have already informally decided that the compound symmetry condition is reasonable for the satiety example then it should not be surprising that the variances of differences are similar across pairs of time points and treatments. The largest standard deviation is 2.4 with the smallest of 1.1, differing

Table 10.15: Standard Deviations for Pairwise Differences in Hunger Levels

| | Time Pair | | | | | |
|---|---|---|---|---|---|---|
| Beverage | 1 vs 2 | 1 vs 3 | 1 vs 4 | 2 vs 3 | 2 vs 4 | 2 vs 5 |
| Milk | 1.6 | 2.4 | 1.8 | 1.5 | 1.6 | 2.4 |
| Coke | 1.3 | 1.2 | 1.3 | 1.1 | 1.1 | 1.2 |
| Control | 1.3 | 1.2 | 1.3 | 1.1 | 1.1 | 1.2 |

Table 10.16: Standard Deviations of Pairwise Differences of Rat Weights

| | Time Pair | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group | 0 vs 1 | 0 vs 2 | 0 vs 3 | 0 vs 4 | 1 vs 2 | 1 vs 3 | 1 vs 4 | 2 vs 3 | 2 vs 4 | 3 vs 4 |
| Control | 4.8 | 7.0 | 11.5 | 14.4 | 4.7 | 10.3 | 13.6 | 5.9 | 9.5 | 5.4 |
| Thiourac | 5.4 | 5.9 | 7.9 | 11.1 | 4.2 | 8.8 | 12.7 | 5.8 | 10.0 | 6.8 |
| Thyroxin | 4.3 | 9.3 | 15.6 | 20.2 | 6.4 | 13.8 | 18.5 | 8.6 | 13.2 | 5.4 |

by a factor of only about 2.

Standard deviations for differences in rat weights for all possible pairs and treatments are given in Table **??**.

Note that there is considerable difference among the standard deviations. The largest standard deviation of 20.2 is almost 5 times the smallest standard deviation of 4.2, confirming the validity of using a split plot model analysis for this data.

The Mauchly W test is a formal hypothesis test of the Huynh-Feldt condition developed by Mauchly in 1940 ("Significance test for sphericity of a normal n-variate distribution," *Annals of Mathematical Statistics* 11, 204-209). The null hypothesis is that the Huyhn-Feldt condition for the variances and covariances of the repeated measures holds and the alternative hypothesis is that the condition does not hold. The test statistic has an approximate large sample chi-square distribution under the null hypothesis. The value of the test statistic and P-value is reported in software with repeated measures analysis. A significant result (P-value $< \alpha$) indicates that the condition does not hold and therefore the split plot analysis is not valid.

The results of the Mauchly test for the rat weights data are W $=$ 0.852, Chi-square $=$ 101.79003 with 5 degrees of freedom, P-value $<$ 0.0001. There is evidence at the 0.05 level of significance that the Huynh-Feldt condition does not hold. Thus the split plot analysis is inappropriate.

The results of the Mauchly test for the satiety data are W = 0.787, Chi-square = 6.16 with 5 degrees of freedom, P-value = 0.2910. There is no evidence at the 0.05 level of significance that the Huynh-Feldt condition is violated. Thus the split plot analysis is a reasonable approach to analyzing the data.

Informal and formal procedures, along with a special knowledge of the variables, should be used to assess whether the split plot analysis can be used to analyze repeated measures data.

### 10.2.8 Alternative Analyses if Huynh-Feldt Conditions Do Not Hold

When the Huyn-Feldt condition is not satisfied there are several types of analysis that can be employed. These are described below.

**Multivariate methods**

This approach treats the repeated measures for each subject as a vector and then analyzes the vectors within and between treatments using multivariate analysis of variance (MANOVA). The variances and covariances of the repeated measures can take on any form. If a subject has at least one missing repeated observation then the rest of the data for that subject is not used. MANOVA is beyond the scope of this text. Interested readers may refer to Milliken and Johnson ([21], page 537) for a detailed discussion of this approach.

**Adjust P-values in Split Plot Analysis**

This approach uses the split plot analysis but adjusts degrees of freedom, and thus P-values, associated with the split plot test for time main effects and the interaction between Treatment and Time. (The test for main effects of treatment is valid even if the Huynh-Feldt condition is not satisfied). The adjustment is to multiply the usual degrees of freedom for the two tests, resulting in lower degrees of freedom and higher P-values. There are two different adjustments that are usually seen in software output. One is called the Greenhouse and Geisser adjustment (”On methods in the analysis of profile data,” 1959, Psychometrika, 24, 95-11). The other is the Huynh-Feldt adjustment (”Estimation of the Box correction for degrees of freedom from sample data in the randomized block design and split-plot designs,” Journal of Educational Statistics, 1976, 1, 69-82).

This approach fixes the problem by altering the usual split plot analysis. Schabenberger and Pierce ([27], page 464) argue that rather than forcing the usual split plot analysis on inappropriate data by employing *fudge* factors the analyst should use a direct approach employing modern statistical methods. That is use procedures that recognize and models variance and correlation structures which might be appropriate for the data at hand. This author agrees with the philosophy of Schabenberger and Pierce and will not further explore the P-value adjustment methods.

### Reduce/summarize repeated measures with single value

The basic idea of this approach is to reduce the repeated measures on each subject to a single measure and then compare the single measure using the analysis for a one-factor completely randomized design of Chapter 4. For example the repeated measures for each subject might be linearly related to the time variable. If so then a linear regression analysis could be conducted for each subject giving a slope measuring the rate of change of the response for each. These slopes could then be compared using the methods of Chapter 4.

### Mixed Model Methods

Mixed models methodology offers possible variance/covariance structures to be selected in the analysis. Scientific knowledge may offer suggestions as to one of a few appropriate structure. For example it may be hypothesized that correlations decay with time or that variances differ across time. The few structures may be estimated based on the variation in the data and then one is selected for the final analysis based on model selection criteria. Treatment, time, and interaction effects with standard errors are then estimated based on the chosen variance covariance structure and estimated degrees of freedom. Tests for various effects and multiple comparison procedures can then be employed as in ANOVA. The specifics of mixed model methods are beyond the scope of this text. Interested readers with a background in matrix algebra may refer to Milliken and Johnson ([21], page 553.

## 10.3 SAS Code

### 10.3.1 Example 10.1

```
*  SAS Code for Example 10.1;


*  Input data;
data Cupcakes;
input Flavor $ Batch $  Cupcake Height BatchCupcake;
datalines;
C 1 1 46 C1
C 1 2 50 C1
C 1 3 47 C1
C 1 4 45 C1
C 2 1 51 C2
C 2 2 46 C2
C 2 3 47 C2
C 2 4 45 C2
C 3 1 42 C3
C 3 2 56 C3
C 3 3 49 C3
C 3 4 46 C3
C 4 1 55 C4
C 4 2 50 C4
C 4 3 45 C4
C 4 4 45 C4
S 1 1 36 S1
S 1 2 31 S1
S 1 3 35 S1
S 1 4 36 S1
S 2 1 45 S2
S 2 2 44 S2
S 2 3 45 S2
S 2 4 45 S2
S 3 1 42 S3
S 3 2 37 S3
S 3 3 40 S3
S 3 4 45 S3
S 4 1 40 S4
S 4 2 36 S4
S 4 3 45 S4
S 4 4 47 S4
F 1 1 33 F1
F 1 2 35 F1
F 1 3 35 F1
```

```
F 1 4 33 F1
F 2 1 44 F2
F 2 2 46 F2
F 2 3 41 F2
F 2 4 40 F2
F 3 1 36 F3
F 3 2 35 F3
F 3 3 32 F3
F 3 4 33 F3
F 4 1 37 F4
F 4 2 41 F4
F 4 3 40 F4
F 4 4 40 F4
;
run;


* Compute ANOVA table and construct Tukey-Kramer pairwise comparisons;
proc glm data = Cupcakes;
  class Flavor Batch;
  model Height = Flavor Batch(Flavor);
  random Batch(Flavor) / test;
  lsmeans Flavor / pdiff tdiff adjust = tukey e = Flavor(Batch);
run;


*  SAS Code for Example 10.3;


*  Input data;
data Satiety;
input Beverage $ Subject Time;
datalines;
Milk 1 1 28
Milk 1 2 39
Milk 1 3 49
Milk 1 4 56
Milk 2 1 26
Milk 2 2 37
Milk 2 3 49
Milk 2 4 51
Milk 3 1 27
Milk 3 2 39
Milk 3 3 51
Milk 3 4 54
Milk 4 1 24
Milk 4 2 40
Milk 4 3 51
Milk 4 4 55
Milk 5 1 27
Milk 5 2 40
Milk 5 3 51
Milk 5 4 53
Milk 6 1 27
Milk 6 2 39
Milk 6 3 48
Milk 6 4 55
Milk 7 1 28
Milk 7 2 39
Milk 7 3 50
Milk 7 4 57
Milk 8 1 26
Milk 8 2 40
Milk 8 3 53
Milk 8 4 56
Milk 9 1 26
Milk 9 2 38
Milk 9 3 46
Milk 9 4 54
Milk 10 1 26
Milk 10 2 38
Milk 10 3 50
Milk 10 4 56
Coke 11 1 26
Coke 11 2 42
Coke 11 3 54
Coke 11 4 63
Coke 12 1 26
Coke 12 2 41
Coke 12 3 54
Coke 12 4 60
Coke 13 1 26
Coke 13 2 43
Coke 13 3 55
Coke 13 4 64
Coke 14 1 27
Coke 14 2 43
Coke 14 3 55
Coke 14 4 62
Coke 15 1 29
Coke 15 2 44
Coke 15 3 56
Coke 15 4 64
Coke 16 1 30
Coke 16 2 43
Coke 16 3 55
Coke 16 4 65
Coke 17 1 28
Coke 17 2 42
Coke 17 3 54
Coke 17 4 62
Coke 18 1 27
Coke 18 2 40
Coke 18 3 55
Coke 18 4 62
Coke 19 1 27
Coke 19 2 41
Coke 19 3 55
Coke 19 4 62
Coke 20 1 28
Coke 20 2 42
Coke 20 3 56
Coke 20 4 63
Control 21 1 50
Control 21 2 55
Control 21 3 63
Control 21 4 70
Control 22 1 50
Control 22 2 54
Control 22 3 63
Control 22 4 67
Control 23 1 50
Control 23 2 56
Control 23 3 64
Control 23 4 71
Control 24 1 51
Control 24 2 56
Control 24 3 64
Control 24 4 69
Control 25 1 53
Control 25 2 57
Control 25 3 65
Control 25 4 71
Control 26 1 54
Control 26 2 56
Control 26 3 64
Control 26 4 72
Control 27 1 52
Control 27 2 55
Control 27 3 63
Control 27 4 69
Control 28 1 51
Control 28 2 53
Control 28 3 64
Control 28 4 69
Control 29 1 51
Control 29 2 54
Control 29 3 64
Control 29 4 69
Control 30 1 52
Control 30 2 55
Control 30 3 65
Control 30 4 70
;

;
run;

* Proc glm to obtain ANOVA table;
proc glm data = Satiety;
  class Beverage Subject Time;
  model Satiety = Beverage Subject(Beverage) Time Beverage*Time;
  random Subject(Beverage) / test;
run;


* Use Proc glimmix to Compute Tukey-Kramer Pairwise comparisons;
proc glimmix data = Satiety;
   class Beverage Subject Time;
   model Satiety = Beverage Time Beverage*Time / ddfm = kr;
   random Subject(Beverage);
   lsmeans Beverage*Time / slicediff = (Beverage Time) cl adjust = tukey;
run;


* Create data set with hunger levels in wide format for
correlations and Mauchly test results.

Calculate differences between hunger levels for pairs of times
for calculation of standard deviations of differences;

data Satiety_Wide;
  input Beverage $ Subject Time1 Time2 Time3 Time4;
  diff12 = Time1 - Time2;
  diff13 = Time1 - Time3;
  diff14 = Time1 - Time4;
  diff23 = Time2 - Time3;
  diff24 = Time2 - Time4;
  diff34 = Time3 - Time4;
datalines;
Milk 1 28 39 49 56
Milk 2 26 37 49 51
Milk 3 27 39 51 54
Milk 4 24 40 51 55
Milk 5 27 40 51 53
Milk 6 27 39 48 55
Milk 7 28 39 50 57
Milk 8 26 40 53 56
Milk 9 26 38 46 54
Milk 10 26 38 50 56
Coke 11 26 42 54 63
Coke 12 26 41 54 60
Coke 13 26 43 55 64
Coke 14 27 43 55 62
Coke 15 29 44 56 64
Coke 16 30 43 55 65
Coke 17 28 42 54 62
Coke 18 27 40 55 62
Coke 19 27 41 55 62
Coke 20 28 42 56 63
Control 21 50 55 63 70
Control 22 50 54 63 67
Control 23 50 56 64 71
Control 24 51 56 64 69
Control 25 53 57 65 71
Control 26 54 56 64 72
Control 27 52 55 63 69
Control 28 51 53 64 69
Control 29 51 54 64 69
Control 30 52 55 65 70
;


* Code to obtain correlations of pairs of hunger levels
proc sort data = Satiety_Wide;
  by Group;
run;

proc corr data = Satiety_Wide;
 by Group;
 var Time1 Time2 Time3 Time4;
run;



* Proc means to obtain standard deviations
for differences of hunger levels for pairs of
times;

proc means data = Satiety_Wide;
   var diff12 diff13 diff14 diff23 diff24 diff34;
   class group;
run;



* Proc glm to obtain results of
Mauchly test of sphericity;

proc glm data = Satiety_Wide;
  class Group;
  model Time1-Time4 = Group / nouni;
  repeated time 4 / printe;
run;
```

# Problems for Chapter 10

10.1* This example is a modification of a project conducted by Zachary Buchin and and Frank Galante (Spring 2013). Three paper airplane designs (Basic Dart, Lightning, Thunder) were compared on flight distance (cm). The experiment was conducted as follows. A design was randomly selected. A paper airplane of the given design was constructed using standard printer paper. The plane was launched by hand down the hallway of a dormitory. Each plane constructed was thrown three times to increase precision for the particular plane. The design is a one factor completely randomized design with subsampling. The distances (cm) are given in the table below.

| | Plane Design | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Basic Dart | | | Lightning | | | Thunder | | |
| Observation Unit | Plane 1 | Plane 2 | Plane 3 | Plane 4 | Plane 5 | Plane 6 | Plane 7 | Plane 8 | Plane 9 |
| 1 | 586 | 590 | 560 | 450 | 482 | 501 | 151 | 170 | 157 |
| 2 | 592 | 565 | 582 | 447 | 459 | 485 | 159 | 201 | 174 |
| 3 | 561 | 601 | 540 | 482 | 467 | 495 | 168 | 188 | 179 |

a. What are the experimental units? What are some extraneous variables that contribute to experimental error?

b. What are the measurement units? What are some extraneous variables that contribute to error among measurement units?

c. Give a model for this experiment and describe the terms in the model including the error terms.

d. Use software to determine if the differences in flight time are statistically significant.

e. Can the flight distances be analyzed differently by summarizing the 3 distances for each plane? Explain.

10.2* In the article "Performance of broiler finishing hens as affected by fermentation duration of Bambara groundnut ( *Vigna subterranean*) meal" ([2014]: 2(2): pgs 29-34), researchers compared five test diets on weight gain of broiler hens. Two hundred 21-day old female Arbor acre broiler chicks were randomly assigned to 25 pens each with 8 broiler hens. The 25 pens were randomly assigned to one of five test diets (five pens per diet) given below:

1. C0: control, no Bambura groundnut (BGN)
2. F0: 15 % BGN, socked in water for 0 hours

3. F12: 15 % BGN, socked in water for 12 hours
4. F24: 15 % BGN, socked in water for 24 hours
5. F36: 15 % BGN, socked in water for 36 hours

The design is a one factor completely randomized design with subsampling.

a. What are the experimental units? Give some extraneous variables that contribute to variation in weight gain among the experimental units.
b. What are the measurement units? Give some extraneous variables that contribute to variation in weight gain among the measurement units.
c. Weight gains for the 8 chicks in each pen were summarized with the response pen mean weight gain with 5 values per group. The following are summaries of pen mean weight gain for the 5 treatment groups ($mean \pm std.dev(grams)$):
   - C0: $1852.50 \pm 94.01$
   - F0: $1532.62 \pm 72.44$
   - F12: $1579.37 \pm 90.34$
   - F24: $1571.90 \pm 33.19$
   - F36: $1578.00 \pm 21.29$

   With this information could you calculate MSE for testing for diet effects? Yes or no and explain.

10.3* This is a new look at the balloon inflation example from Dean and Voss ([6], page 62). See also Example 8.2. The purpose of the study was to determine if color of a balloon had an effect on the amount of time to blow up the balloon. One person blew up 20 balloons of 4 different colors, 5 per color. The colors were pink, yellow, orange, and blue. One person blew up all 20 balloons in a random order. The design is a one factor completely randomized with experimental unit being time slots, or the blowing up of a balloon at a time slot. The response variable was the amount of time to blew up a balloon recorded to the nearest 1/10 of a second.

Suppose that we want to also study the effect that number of inflations on the same balloon has on the amount of time to blow up the balloon. Suppose that each of the 20 balloons is blown up 3 times in succession by the same person at 5 minute intervals. After a particular balloon is blown up once the air is let out. Five minutes later it is blown up a second time and then the air is let out again. Five minutes later it is blown up a third time and then the air is let out. Thus there are three repeated measurements on amount of time to inflate (seconds) for each balloon In addition to the effect of the color of the balloon on amount of time to inflate a balloon we also wish to study the effect of the number of inflations of the balloon.

The resulting design is a repeated measures design.

a. What are the two types of experimental units - whole unit and split unit?

b. What are the two factors?

c. Are the whole units arranged in a completely randomized design or in blocks? Explain.

d. What are the block(s)?

10.4* Justin Bell and Josh Doorly (Spring 2014) investigated the effects of brand of gum (Big Red, Wrigley's Juice Fruit) and chewing time (30, 45, 60, 75, 90, 105 seconds) of gum on how long the chewed gum could be stretched before breaking. Five sticks of gum were used for each of the 12 combinations of brand and chewing time for a total of 60 sticks of gum. The 60 sticks of gum were chewed and stretched one at a time. At a particular chewing/stretching session a combination of brand and chewing time was randomly selected. Then a stick was randomly selected corresponding to the selected brand. This was repeated for 60 sessions. The same person did all of the chewing of the sticks of gum.

a. What are the factors of interest in this study?

b. Are there one or two types of experimental units?

c. Is the design a repeated measures design? Explain.

10.5* Consider the paper airplane study of Exercise 10.1. Since there are three distances collected over time for each plane there is a potential time effect in that repeated handling of the plane might result in a deterioration of the structure and loss in flight distances. Explain how the design might be regarded as a split plot design. Give the two types of experimental units. Give the two factors, whole and split unit factors. What are the blocks?

10.6* Consider the mealybug example in this chapter (Section 10.3.1). Suppose that there is only one patch per branch. The number of mealybugs on a patch is measured right before treatment. After the treatment there are 7 daily observations on number of mealybugs on the patch. The response variable is percentage decrease in the number of mealybugs on a day compared to the initial number before treatment.

This new experiment can be viewed as a split plot design with repeated measures.

a. Give the two types of experimental units, whole units and split units.

b. Give the two factors, whole and split unit factors.

c. Are the whole units blocked before assignment to treatments or are they assigned to treatments in a completely randomized manner?

d. Describe all types of blocking in this repeated measures design.

10.7* Se Chang, Matthew Harris, and Kevin Tomlinson (Spring 2015) investigated the cooling rates of different liquids (cola, deionized water, olive oil). They conducted their experiment over a period of 5 days. On the first day three beakers were each filled with 50 mL of liquid, one with cola, one with deionized water, and one with olive oil. The three beakers were placed at randomly located positions on a burner with a digital thermometer placed in each beaker/liquid. The burner was then turned on and each of the liquids was in the beakers were allowed to heat until reaching 50°C. Upon reaching the desired temperature a beaker with its liquid was removed and then the temperature of the liquid was measured at 1, 2, 3, 4, and 5 minutes after removal from the heat. This procedure of testing the three liquids at a time was repeated on 4 other days. Interest was in investigating the time profile of the temperature for each liquid and comparing the temperature of the three liquids at each time point.

This is an example of split plot/repeated measures design.

a. What are the whole units and give an extraneous variable associated with the whole units?
b. What are the split units and give an extraneous variable associated with the split units?
c. Give the two factors, whole and split unit factors.
d. Are the whole units assigned completely at random to the levels of the whole unit factor or are the whole units blocked first and then assigned to levels of the whole unit factor? Explain.
e. Describe all blocking that was used in this experiment.

10.8* This example is a modification of an example from Dowdy and Reardon ([9], page 376). Five different methods are being considered for packaging soda crackers to protect them from humidity. Crackers lose their crispness in humid conditions. The packaging methods are: 1) Control Box (cardboard), 2) Wax Paper Box, 3) Metal Foil Box, 4) Plastic Box, and 5) Metal Foil and Plastic Box. A randomly selected box of each type of packaging is selected and the four boxes placed in random locations in a chamber where the humidity is maintained at 80% humidity for 24 hours. After 24 hours the boxes are opened and three crackers are selected at random from each of the 5 boxes and measured for moisture content. This testing procedure is repeated for four other different chambers, each chamber being maintained at 80% humidity for 24 hours.

This study has subsampling.

a. What is the factor of interest? What are the treatments?
b. What are the experimental units? How many experimental units are there?
c. What are the measurement units? How many measurement units are there?
d. Is this a completely randomized design or a randomized complete block design? Explain.

10.9* Angie Davenport, Taryn McLaughlin, and Charlie Kim (Spring 2015) investigated the melting rate of ice using different salt treatments. They conducted their experiment over a period of 5 days. On each of 5 days three trials were conducted, one for each of the three salt treatments(Morton kosher salt, Road Runner calcium chloride, no salt), randomly assigned through time. On a particular trial a small ice mass was laid on top of a wire mesh which was situated above a graduated beaker so that water from the melting ice could be measured. Ice masses were formed by freezing 75 mL of water in a Dixie cup. The randomly selected salt treatment was applied to the ice mass and then the cumulative amount of water from the melting ice (mL) mass was measured at 30, 60, 90, 120, and 150 minutes after the salt treatment was applied.

| | | Time Interval(min) | | | | |
|---|---|---|---|---|---|---|
| Type of Salt | | 30 | 60 | 90 | 120 | 150 |
| | Block | | | | | |
| Kosher | 1 | 12.0 | 21.0 | 32.0 | 42.0 | 49.0 |
| | 2 | 15.0 | 24.0 | 38.0 | 48.0 | 54.0 |
| | 3 | 12.0 | 20.0 | 30.0 | 42.0 | 50.0 |
| | 4 | 17.0 | 24.0 | 39.0 | 49.0 | 57.0 |
| | 5 | 13.0 | 22.0 | 35.0 | 47.0 | 52.0 |
| Calcium chloride | 1 | 15.2 | 28.8 | 43.0 | 53.0 | 60.0 |
| | 2 | 14.0 | 27.0 | 40.0 | 50.0 | 59.0 |
| | 3 | 15.0 | 20.0 | 30.0 | 42.0 | 52.0 |
| | 4 | 16.0 | 27.0 | 42.0 | 51.0 | 60.0 |
| | 5 | 15.0 | 30.0 | 44.0 | 54.0 | 60.0 |
| None | 1 | 10.5 | 16.5 | 25.0 | 32.0 | 40.0 |
| | 2 | 6.0 | 15.0 | 23.0 | 31.0 | 40.0 |
| | 3 | 6.0 | 13.0 | 20.0 | 28.0 | 32.0 |
| | 4 | 7.0 | 15.0 | 25.0 | 34.0 | 41.0 |
| | 5 | 7.0 | 17.0 | 27.0 | 35.0 | 43.0 |

This is an example of split plot/repeated measures design.

a. What are the whole units and give an extraneous variable associated with the whole units?

b. What are the split units and give an extraneous variable associated with the split units?

c. Give the two factors, whole and split unit factors.

d. Are the whole units assigned completely at random to the levels of the whole unit factor or are the whole units blocked first and then assigned to levels of the whole unit factor? Explain.

e. Describe all blocking that was used in this experiment.

f. Use software to determine if there are any trends in the correlations among pairs of the response variable.

g. Use software to calculate standard deviations of differences in the response among all possible pairs of time points to informally check on the Huyn-Feldt condition. Interpret.

h. Use software to obtain results of the Mauchly test of the Huyn-Feldt condition. Is there any evidence that the condition is not satisfied for this data? Explain.

# Appendix A

# Tables

Table A.1: Standard Normal Right Tail Probabilities

| | Table entries are areas under standard normal curve to the right of z | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
| 0.00 | .5000 | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |
| 0.10 | .4602 | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| 0.20 | .4207 | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| 0.30 | .3821 | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| 0.40 | .3446 | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| 0.50 | .3085 | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| 0.60 | .2743 | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| 0.70 | .2420 | .2389 | .2358 | .2327 | .2296 | .2266 | .2236 | .2206 | .2177 | .2148 |
| 0.80 | .2119 | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| 0.90 | .1841 | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| 1.00 | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| 1.10 | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| 1.20 | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| 1.30 | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| 1.40 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| 1.50 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| 1.60 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| 1.70 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| 1.80 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| 1.90 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| 2.00 | .0228 | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| 2.10 | .0179 | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| 2.20 | .0139 | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| 2.30 | .0107 | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| 2.40 | .0082 | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| 2.50 | .0062 | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| 2.60 | .0047 | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| 2.70 | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |
| 2.80 | .0026 | .0025 | .0024 | .0023 | .0023 | .0022 | .0021 | .0021 | .0020 | .0019 |
| 2.90 | .0019 | .0018 | .0018 | .0017 | .0016 | .0016 | .0015 | .0015 | .0014 | .0014 |
| 3.00 | .0013 | .0013 | .0013 | .0012 | .0012 | .0011 | .0011 | .0011 | .0010 | .0010 |

Table A.2: Upper $\alpha$ probability points for the Student t distribution

| | Table entries are $t_{\alpha;\nu}$, where $P[t > t_{\alpha;\nu}] = \alpha$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$ | | | | | | | |
| $\nu$ | 0.40 | 0.25 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0005 |
| 1 | .325 | 1.000 | 3.078 | 6.314 | 12.706 | 31.820 | 63.657 | 636.619 |
| 2 | .289 | .816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 31.599 |
| 3 | .277 | .765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 12.924 |
| 4 | .271 | .741 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 8.610 |
| 5 | .267 | .727 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 6.869 |
| 6 | .265 | .718 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959 |
| 7 | .263 | .711 | 1.415 | 1.895 | 2.365 | 2.998 | 3.500 | 5.408 |
| 8 | .262 | .706 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041 |
| 9 | .261 | .703 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781 |
| 10 | .260 | .700 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587 |
| 11 | .260 | .697 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.437 |
| 12 | .259 | .695 | 1.356 | 1.782 | 2.179 | 2.681 | 3.054 | 4.318 |
| 13 | .259 | .694 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 4.221 |
| 14 | .258 | .692 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 4.140 |
| 15 | .258 | .691 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 4.073 |
| 16 | .258 | .690 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 4.015 |
| 17 | .257 | .689 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.965 |
| 18 | .257 | .688 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.922 |
| 19 | .257 | .688 | 1.328 | 1.729 | 2.093 | 2.540 | 2.861 | 3.883 |
| 20 | .257 | .687 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.850 |
| 21 | .257 | .686 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.819 |
| 22 | .256 | .686 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.792 |
| 23 | .256 | .685 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.768 |
| 24 | .256 | .685 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.745 |
| 25 | .256 | .684 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.725 |
| 26 | .256 | .684 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.707 |
| 27 | .256 | .684 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.690 |
| 28 | .256 | .683 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.674 |
| 29 | .256 | .683 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.659 |
| 30 | .256 | .683 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.646 |
| 40 | .255 | .681 | 1.303 | 1.683 | 2.021 | 2.423 | 2.704 | 3.551 |
| 60 | .254 | .678 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.460 |
| 120 | .254 | .677 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.373 |
| $\infty$ | .253 | .674 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.291 |

Table A.3: Upper $0.05/2m$ Bonferroni probability point for the Student t distribution

Table entries are $t_{0.05/2m;\nu}$, where $P[t > t_{0.05/2m;\nu}] = 0.05/2m$

| $\nu^{\backslash m}$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 25.4 | 38.2 | 50.9 | 63.7 | 76.4 | 89.1 | 101. | 115. | 127. | 191. |
| 2 | 6.21 | 7.65 | 8.86 | 9.92 | 10.9 | 11.8 | 12.6 | 13.4 | 14.1 | 17.3 |
| 3 | 4.18 | 4.86 | 5.39 | 5.84 | 6.23 | 6.58 | 6.90 | 7.18 | 7.45 | 8.58 |
| 4 | 3.50 | 3.96 | 4.31 | 4.60 | 4.85 | 5.07 | 5.26 | 5.44 | 5.60 | 6.25 |
| 5 | 3.16 | 3.53 | 3.81 | 4.03 | 4.22 | 4.38 | 4.53 | 4.66 | 4.77 | 5.25 |
| 6 | 2.97 | 3.29 | 3.52 | 3.71 | 3.86 | 4.00 | 4.12 | 4.22 | 4.32 | 4.70 |
| 7 | 2.84 | 3.13 | 3.34 | 3.50 | 3.64 | 3.75 | 3.86 | 3.95 | 4.03 | 4.36 |
| 8 | 2.75 | 3.02 | 3.21 | 3.36 | 3.48 | 3.58 | 3.68 | 3.76 | 3.83 | 4.12 |
| 9 | 2.69 | 2.93 | 3.11 | 3.25 | 3.36 | 3.46 | 3.55 | 3.62 | 3.69 | 3.95 |
| 10 | 2.63 | 2.87 | 3.04 | 3.17 | 3.28 | 3.37 | 3.45 | 3.52 | 3.58 | 3.83 |
| 11 | 2.59 | 2.82 | 2.98 | 3.11 | 3.21 | 3.29 | 3.37 | 3.44 | 3.50 | 3.73 |
| 12 | 2.56 | 2.78 | 2.93 | 3.05 | 3.15 | 3.24 | 3.31 | 3.37 | 3.43 | 3.65 |
| 13 | 2.53 | 2.75 | 2.90 | 3.01 | 3.11 | 3.19 | 3.26 | 3.32 | 3.37 | 3.58 |
| 14 | 2.51 | 2.72 | 2.86 | 2.98 | 3.07 | 3.15 | 3.21 | 3.27 | 3.33 | 3.53 |
| 15 | 2.49 | 2.69 | 2.84 | 2.95 | 3.04 | 3.11 | 3.18 | 3.22 | 3.29 | 3.48 |
| 16 | 2.47 | 2.67 | 2.81 | 2.92 | 3.01 | 3.08 | 3.15 | 3.20 | 3.25 | 3.44 |
| 17 | 2.46 | 2.65 | 2.79 | 2.90 | 2.98 | 3.06 | 3.12 | 3.17 | 3.22 | 3.41 |
| 18 | 2.45 | 2.64 | 2.77 | 2.88 | 2.96 | 3.03 | 3.09 | 3.15 | 3.20 | 3.38 |
| 19 | 2.43 | 2.63 | 2.76 | 2.86 | 2.94 | 3.01 | 3.07 | 3.13 | 3.17 | 3.35 |
| 20 | 2.42 | 2.61 | 2.74 | 2.85 | 2.93 | 3.00 | 3.06 | 3.11 | 3.15 | 3.33 |
| 21 | 2.41 | 2.60 | 2.73 | 2.83 | 2.91 | 2.98 | 3.04 | 3.09 | 3.14 | 3.31 |
| 22 | 2.41 | 2.59 | 2.72 | 2.82 | 2.90 | 2.97 | 3.02 | 3.07 | 3.12 | 3.29 |
| 23 | 2.40 | 2.58 | 2.71 | 2.81 | 2.89 | 2.95 | 3.01 | 3.06 | 3.10 | 3.27 |
| 24 | 2.39 | 2.57 | 2.70 | 2.80 | 2.88 | 2.94 | 3.00 | 3.05 | 3.09 | 3.26 |
| 25 | 2.38 | 2.57 | 2.69 | 2.79 | 2.86 | 2.93 | 2.99 | 3.03 | 3.08 | 3.24 |
| 26 | 2.38 | 2.56 | 2.68 | 2.78 | 2.86 | 2.92 | 2.98 | 3.02 | 3.07 | 3.23 |
| 27 | 2.37 | 2.55 | 2.68 | 2.77 | 2.85 | 2.91 | 2.97 | 3.01 | 3.06 | 3.22 |
| 28 | 2.37 | 2.55 | 2.67 | 2.76 | 2.84 | 2.90 | 2.96 | 3.00 | 3.05 | 3.21 |
| 29 | 2.36 | 2.54 | 2.66 | 2.76 | 2.83 | 2.89 | 2.95 | 3.00 | 3.04 | 3.20 |
| 30 | 2.36 | 2.54 | 2.66 | 2.75 | 2.82 | 2.89 | 2.94 | 2.99 | 3.03 | 3.19 |
| 40 | 2.33 | 2.50 | 2.62 | 2.70 | 2.78 | 2.84 | 2.89 | 2.93 | 2.97 | 3.12 |
| 60 | 2.30 | 2.46 | 2.58 | 2.66 | 2.73 | 2.79 | 2.83 | 2.88 | 2.91 | 3.06 |
| 120 | 2.27 | 2.43 | 2.54 | 2.62 | 2.68 | 2.74 | 2.78 | 2.82 | 2.86 | 3.00 |
| $\infty$ | 2.24 | 2.39 | 2.50 | 2.58 | 2.64 | 2.69 | 2.73 | 2.77 | 2.81 | 2.94 |

Table A.4: Upper $0.01/2m$ Bonferroni probability point for the Student t distribution

Table entries are $t_{0.01/2m;\nu}$, where $P[t > t_{0.01/2m;\nu}] = 0.01/2m$

| $\nu \backslash m$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 127. | 191. | 255. | 318. | 382. | 446. | 509. | 573. | 624. | 955. |
| 2 | 14.1 | 17.3 | 20.0 | 22.3 | 24.5 | 26.4 | 28.3 | 30.0 | 31.6 | 38.7 |
| 3 | 7.45 | 8.58 | 9.46 | 10.2 | 10.9 | 11.4 | 12.0 | 12.5 | 12.9 | 14.8 |
| 4 | 5.60 | 6.25 | 6.76 | 7.17 | 7.53 | 7.84 | 8.12 | 8.38 | 8.61 | 9.57 |
| 5 | 4.77 | 5.25 | 5.60 | 5.89 | 6.14 | 6.35 | 6.54 | 6.71 | 6.87 | 7.50 |
| 6 | 4.32 | 4.70 | 4.98 | 5.21 | 5.40 | 5.56 | 5.71 | 5.84 | 5.96 | 6.43 |
| 7 | 4.03 | 4.36 | 4.59 | 4.79 | 4.94 | 5.08 | 5.20 | 5.31 | 5.41 | 5.80 |
| 8 | 3.83 | 4.12 | 4.33 | 4.50 | 4.64 | 4.76 | 4.86 | 4.96 | 5.04 | 5.37 |
| 9 | 3.69 | 3.95 | 4.15 | 4.30 | 4.42 | 4.53 | 4.62 | 4.71 | 4.78 | 5.08 |
| 10 | 3.58 | 3.83 | 4.00 | 4.14 | 4.26 | 4.36 | 4.44 | 4.52 | 4.59 | 4.85 |
| 11 | 3.50 | 3.73 | 3.89 | 4.02 | 4.13 | 4.22 | 4.30 | 4.37 | 4.44 | 4.68 |
| 12 | 3.43 | 3.65 | 3.81 | 3.93 | 4.03 | 4.12 | 4.19 | 4.26 | 4.32 | 4.55 |
| 13 | 3.37 | 3.58 | 3.73 | 3.85 | 3.95 | 4.03 | 4.10 | 4.16 | 4.22 | 4.44 |
| 14 | 3.33 | 3.53 | 3.67 | 3.79 | 3.88 | 3.96 | 4.03 | 4.09 | 4.14 | 4.35 |
| 15 | 3.29 | 3.48 | 3.62 | 3.73 | 3.82 | 3.90 | 3.96 | 4.02 | 4.07 | 4.27 |
| 16 | 3.25 | 3.44 | 3.58 | 3.69 | 3.77 | 3.85 | 3.91 | 3.96 | 4.01 | 4.21 |
| 17 | 3.22 | 3.41 | 3.54 | 3.65 | 3.73 | 3.80 | 3.86 | 3.92 | 3.97 | 4.15 |
| 18 | 3.20 | 3.38 | 3.51 | 3.61 | 3.69 | 3.76 | 3.82 | 3.87 | 3.92 | 4.10 |
| 19 | 3.17 | 3.35 | 3.48 | 3.58 | 3.66 | 3.73 | 3.79 | 3.84 | 3.88 | 4.06 |
| 20 | 3.15 | 3.33 | 3.46 | 3.55 | 3.63 | 3.70 | 3.75 | 3.80 | 3.85 | 4.02 |
| 21 | 3.14 | 3.31 | 3.43 | 3.53 | 3.60 | 3.67 | 3.73 | 3.78 | 3.82 | 3.99 |
| 22 | 3.12 | 3.29 | 3.41 | 3.50 | 3.58 | 3.64 | 3.70 | 3.75 | 3.79 | 3.96 |
| 23 | 3.10 | 3.27 | 3.39 | 3.48 | 3.56 | 3.62 | 3.68 | 3.72 | 3.77 | 3.93 |
| 24 | 3.09 | 3.26 | 3.38 | 3.47 | 3.54 | 3.60 | 3.66 | 3.70 | 3.75 | 3.91 |
| 25 | 3.08 | 3.24 | 3.36 | 3.45 | 3.52 | 3.58 | 3.64 | 3.68 | 3.73 | 3.88 |
| 26 | 3.07 | 3.23 | 3.35 | 3.43 | 3.51 | 3.57 | 3.62 | 3.67 | 3.71 | 3.86 |
| 27 | 3.06 | 3.22 | 3.33 | 3.42 | 3.49 | 3.55 | 3.60 | 3.65 | 3.69 | 3.84 |
| 28 | 3.05 | 3.21 | 3.32 | 3.41 | 3.48 | 3.54 | 3.59 | 3.63 | 3.67 | 3.83 |
| 29 | 3.04 | 3.20 | 3.31 | 3.40 | 3.47 | 3.52 | 3.58 | 3.62 | 3.66 | 3.81 |
| 30 | 3.03 | 3.19 | 3.30 | 3.39 | 3.45 | 3.51 | 3.56 | 3.61 | 3.65 | 3.80 |
| 40 | 2.97 | 3.12 | 3.23 | 3.31 | 3.37 | 3.43 | 3.47 | 3.51 | 3.55 | 3.69 |
| 60 | 2.91 | 3.06 | 3.16 | 3.23 | 3.29 | 3.34 | 3.39 | 3.43 | 3.46 | 3.59 |
| 120 | 2.86 | 3.00 | 3.09 | 3.16 | 3.22 | 3.26 | 3.31 | 3.34 | 3.37 | 3.49 |
| $\infty$ | 2.81 | 2.94 | 3.02 | 3.09 | 3.14 | 3.19 | 3.23 | 3.26 | 3.29 | 3.40 |

Table A.5: Upper $\alpha = 0.01$ probability point for the Studentized Range Distribution

Table entries are $q_{0.01;\nu,t}$, where $P[q > q_{0.01;\nu,t}] = 0.01$

| $\nu\backslash t$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 14.0 | 19.0 | 22.3 | 24.7 | 26.6 | 28.2 | 29.5 | 30.7 | 31.7 | 32.6 | 33.4 | 34.1 | 34.8 | 35.4 |
| 3 | 8.26 | 10.6 | 12.2 | 13.3 | 14.2 | 15.0 | 15.6 | 16.2 | 16.7 | 17.1 | 17.5 | 17.9 | 18.2 | 18.5 |
| 4 | 6.51 | 8.12 | 9.17 | 9.96 | 10.6 | 11.1 | 11.5 | 11.9 | 12.3 | 12.6 | 12.8 | 13.1 | 13.3 | 13.5 |
| 5 | 5.70 | 6.98 | 7.81 | 8.42 | 8.91 | 9.32 | 9.67 | 9.97 | 10.2 | 10.5 | 10.7 | 10.9 | 11.1 | 11.2 |
| 6 | 5.24 | 6.33 | 7.03 | 7.56 | 7.97 | 8.32 | 8.61 | 8.87 | 9.10 | 9.30 | 9.48 | 9.65 | 9.81 | 9.95 |
| 7 | 4.95 | 5.92 | 6.54 | 7.01 | 7.37 | 7.68 | 7.94 | 8.17 | 8.37 | 8.55 | 8.71 | 8.86 | 9.00 | 9.12 |
| 8 | 4.74 | 5.64 | 6.20 | 6.63 | 6.96 | 7.24 | 7.47 | 7.68 | 7.86 | 8.03 | 8.18 | 8.31 | 8.44 | 8.55 |
| 9 | 4.60 | 5.43 | 5.96 | 6.35 | 6.66 | 6.91 | 7.13 | 7.33 | 7.50 | 7.65 | 7.78 | 7.91 | 8.03 | 8.13 |
| 10 | 4.48 | 5.27 | 5.77 | 6.14 | 6.43 | 6.67 | 6.88 | 7.05 | 7.21 | 7.36 | 7.49 | 7.60 | 7.71 | 7.81 |
| 11 | 4.39 | 5.15 | 5.62 | 5.97 | 6.25 | 6.48 | 6.67 | 6.84 | 6.99 | 7.13 | 7.25 | 7.36 | 7.46 | 7.56 |
| 12 | 4.32 | 5.05 | 5.50 | 5.84 | 6.10 | 6.32 | 6.51 | 6.67 | 6.81 | 6.94 | 7.06 | 7.17 | 7.26 | 7.36 |
| 13 | 4.26 | 4.96 | 5.40 | 5.73 | 5.98 | 6.19 | 6.37 | 6.53 | 6.67 | 6.79 | 6.90 | 7.01 | 7.10 | 7.19 |
| 14 | 4.21 | 4.89 | 5.32 | 5.63 | 5.88 | 6.08 | 6.26 | 6.41 | 6.54 | 6.66 | 6.77 | 6.87 | 6.96 | 7.05 |
| 15 | 4.17 | 4.84 | 5.25 | 5.56 | 5.80 | 5.99 | 6.16 | 6.31 | 6.44 | 6.55 | 6.66 | 6.76 | 6.84 | 6.93 |
| 16 | 4.13 | 4.79 | 5.19 | 5.49 | 5.72 | 5.92 | 6.08 | 6.22 | 6.35 | 6.46 | 6.56 | 6.66 | 6.74 | 6.82 |
| 17 | 4.10 | 4.74 | 5.14 | 5.43 | 5.66 | 5.85 | 6.01 | 6.15 | 6.27 | 6.38 | 6.48 | 6.57 | 6.66 | 6.73 |
| 18 | 4.07 | 4.70 | 5.09 | 5.38 | 5.60 | 5.79 | 5.94 | 6.08 | 6.20 | 6.31 | 6.41 | 6.50 | 6.58 | 6.65 |
| 19 | 4.05 | 4.67 | 5.05 | 5.33 | 5.55 | 5.73 | 5.89 | 6.02 | 6.14 | 6.25 | 6.34 | 6.43 | 6.51 | 6.58 |
| 20 | 4.02 | 4.64 | 5.02 | 5.29 | 5.51 | 5.69 | 5.84 | 5.97 | 6.09 | 6.19 | 6.28 | 6.37 | 6.45 | 6.52 |
| 21 | 4.00 | 4.61 | 4.99 | 5.26 | 5.47 | 5.65 | 5.79 | 5.92 | 6.04 | 6.14 | 6.23 | 6.32 | 6.39 | 6.47 |
| 22 | 3.99 | 4.59 | 4.96 | 5.22 | 5.43 | 5.61 | 5.75 | 5.88 | 5.99 | 6.09 | 6.19 | 6.27 | 6.35 | 6.42 |
| 23 | 3.97 | 4.57 | 4.93 | 5.19 | 5.40 | 5.57 | 5.72 | 5.84 | 5.95 | 6.05 | 6.14 | 6.23 | 6.30 | 6.37 |
| 24 | 3.96 | 4.55 | 4.91 | 5.17 | 5.37 | 5.54 | 5.68 | 5.81 | 5.92 | 6.02 | 6.11 | 6.19 | 6.26 | 6.33 |
| 25 | 3.94 | 4.53 | 4.88 | 5.14 | 5.35 | 5.51 | 5.65 | 5.78 | 5.89 | 5.98 | 6.07 | 6.15 | 6.22 | 6.29 |
| 26 | 3.93 | 4.51 | 4.87 | 5.12 | 5.32 | 5.49 | 5.63 | 5.75 | 5.89 | 5.95 | 6.04 | 6.12 | 6.19 | 6.26 |
| 27 | 3.92 | 4.49 | 4.85 | 5.10 | 5.30 | 5.46 | 5.60 | 5.72 | 5.83 | 5.92 | 6.01 | 6.09 | 6.16 | 6.22 |
| 28 | 3.91 | 4.48 | 4.83 | 5.08 | 5.28 | 5.44 | 5.58 | 5.70 | 5.80 | 5.90 | 5.98 | 6.06 | 6.13 | 6.19 |
| 29 | 3.90 | 4.47 | 4.81 | 5.06 | 5.26 | 5.42 | 5.56 | 5.67 | 5.78 | 5.87 | 5.96 | 6.03 | 6.10 | 6.17 |
| 30 | 3.89 | 4.45 | 4.80 | 5.05 | 5.24 | 5.40 | 5.54 | 5.65 | 5.76 | 5.85 | 5.93 | 6.01 | 6.08 | 6.14 |
| 35 | 3.85 | 4.40 | 4.74 | 4.98 | 5.17 | 5.32 | 5.45 | 5.57 | 5.67 | 5.75 | 5.84 | 5.91 | 5.98 | 6.04 |
| 40 | 3.82 | 4.37 | 4.70 | 4.93 | 5.11 | 5.26 | 5.39 | 5.50 | 5.60 | 5.69 | 5.76 | 5.83 | 5.90 | 5.96 |
| 45 | 3.80 | 4.34 | 4.66 | 4.89 | 5.07 | 5.22 | 5.34 | 5.45 | 5.55 | 5.63 | 5.71 | 5.78 | 5.84 | 5.90 |
| 50 | 3.79 | 4.32 | 4.63 | 4.86 | 5.04 | 5.19 | 5.31 | 5.41 | 5.51 | 5.59 | 5.67 | 5.73 | 5.80 | 5.85 |
| 100 | 3.71 | 4.22 | 4.52 | 4.73 | 4.90 | 5.03 | 5.14 | 5.24 | 5.33 | 5.40 | 5.47 | 5.54 | 5.59 | 5.65 |
| $\infty$ | 3.64 | 4.12 | 4.40 | 4.60 | 4.76 | 4.88 | 4.99 | 5.08 | 5.16 | 5.23 | 5.29 | 5.35 | 5.40 | 5.45 |

Table A.6: Upper $\alpha = 0.05$ probability point for the Studentized Range Distribution

Table entries are $q_{0.05;\nu,t}$, where $P[q > q_{0.05;\nu,t}] = 0.05$

| $\nu\backslash t$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 6.08 | 8.33 | 9.80 | 10.9 | 11.7 | 12.4 | 13.0 | 13.5 | 14.0 | 14.4 | 14.8 | 15.1 | 15.4 | 15.6 |
| 3 | 4.50 | 5.91 | 6.82 | 7.50 | 8.04 | 8.48 | 8.85 | 9.18 | 9.46 | 9.72 | 9.95 | 10.2 | 10.4 | 10.5 |
| 4 | 3.93 | 5.04 | 5.76 | 6.29 | 6.71 | 7.05 | 7.35 | 7.60 | 7.83 | 8.03 | 8.21 | 8.37 | 8.52 | 8.66 |
| 5 | 3.64 | 4.60 | 5.22 | 5.67 | 6.03 | 6.33 | 6.58 | 6.80 | 6.99 | 7.17 | 7.32 | 7.47 | 7.60 | 7.72 |
| 6 | 3.46 | 4.34 | 4.90 | 5.30 | 5.63 | 5.90 | 6.12 | 6.32 | 6.49 | 6.65 | 6.79 | 6.92 | 7.03 | 7.14 |
| 7 | 3.34 | 4.16 | 4.68 | 5.06 | 5.36 | 5.61 | 5.82 | 6.00 | 6.16 | 6.30 | 6.43 | 6.55 | 6.66 | 6.76 |
| 8 | 3.26 | 4.04 | 4.53 | 4.89 | 5.17 | 5.40 | 5.60 | 5.77 | 5.92 | 6.05 | 6.18 | 6.29 | 6.39 | 6.48 |
| 9 | 3.20 | 3.95 | 4.41 | 4.76 | 5.02 | 5.24 | 5.43 | 5.59 | 5.74 | 5.87 | 5.98 | 6.09 | 6.19 | 6.28 |
| 10 | 3.15 | 3.88 | 4.33 | 4.65 | 4.91 | 5.12 | 5.30 | 5.46 | 5.60 | 5.72 | 5.83 | 5.93 | 6.03 | 6.11 |
| 11 | 3.11 | 3.82 | 4.26 | 4.57 | 4.82 | 5.03 | 5.20 | 5.35 | 5.49 | 5.61 | 5.71 | 5.81 | 5.90 | 5.98 |
| 12 | 3.08 | 3.77 | 4.20 | 4.51 | 4.75 | 4.95 | 5.12 | 5.27 | 5.39 | 5.51 | 5.61 | 5.71 | 5.80 | 5.88 |
| 13 | 3.06 | 3.73 | 4.15 | 4.45 | 4.69 | 4.88 | 5.05 | 5.19 | 5.32 | 5.43 | 5.53 | 5.63 | 5.71 | 5.79 |
| 14 | 3.03 | 3.70 | 4.11 | 4.41 | 4.64 | 4.83 | 4.99 | 5.13 | 5.25 | 5.36 | 5.46 | 5.55 | 5.64 | 5.71 |
| 15 | 3.01 | 3.67 | 4.08 | 4.37 | 4.59 | 4.78 | 4.94 | 5.08 | 5.20 | 5.31 | 5.40 | 5.49 | 5.57 | 5.65 |
| 16 | 3.00 | 3.65 | 4.05 | 4.33 | 4.56 | 4.74 | 4.90 | 5.03 | 5.15 | 5.26 | 5.35 | 5.44 | 5.52 | 5.59 |
| 17 | 2.98 | 3.63 | 4.02 | 4.30 | 4.52 | 4.70 | 4.86 | 4.99 | 5.11 | 5.21 | 5.31 | 5.39 | 5.47 | 5.54 |
| 18 | 2.97 | 3.61 | 4.00 | 4.28 | 4.49 | 4.67 | 4.82 | 4.96 | 5.07 | 5.17 | 5.27 | 5.35 | 5.43 | 5.50 |
| 19 | 2.96 | 3.59 | 3.98 | 4.25 | 4.47 | 4.65 | 4.79 | 4.92 | 5.04 | 5.14 | 5.23 | 5.31 | 5.39 | 5.46 |
| 20 | 2.95 | 3.58 | 3.96 | 4.23 | 4.45 | 4.62 | 4.77 | 4.90 | 5.01 | 5.11 | 5.20 | 5.28 | 5.36 | 5.43 |
| 21 | 2.94 | 3.56 | 3.94 | 4.21 | 4.42 | 4.60 | 4.74 | 4.87 | 4.98 | 5.08 | 5.17 | 5.25 | 5.33 | 5.40 |
| 22 | 2.93 | 3.55 | 3.93 | 4.20 | 4.41 | 4.58 | 4.72 | 4.85 | 4.96 | 5.06 | 5.14 | 5.23 | 5.30 | 5.37 |
| 23 | 2.93 | 3.54 | 3.91 | 4.18 | 4.39 | 4.56 | 4.70 | 4.83 | 4.94 | 5.03 | 5.12 | 5.20 | 5.27 | 5.34 |
| 24 | 2.92 | 3.53 | 3.90 | 4.17 | 4.37 | 4.54 | 4.68 | 4.81 | 4.92 | 5.01 | 5.10 | 5.18 | 5.25 | 5.32 |
| 25 | 2.91 | 3.52 | 3.89 | 4.15 | 4.36 | 4.53 | 4.67 | 4.79 | 4.90 | 4.99 | 5.08 | 5.16 | 5.23 | 5.30 |
| 26 | 2.91 | 3.51 | 3.88 | 4.14 | 4.35 | 4.51 | 4.65 | 4.77 | 4.88 | 4.98 | 5.06 | 5.14 | 5.21 | 5.28 |
| 27 | 2.90 | 3.51 | 3.87 | 4.13 | 4.33 | 4.50 | 4.64 | 4.76 | 4.86 | 4.96 | 5.04 | 5.12 | 5.19 | 5.28 |
| 28 | 2.90 | 3.50 | 3.86 | 4.12 | 4.32 | 4.49 | 4.62 | 4.74 | 4.85 | 4.94 | 5.03 | 5.11 | 5.18 | 5.24 |
| 29 | 2.89 | 3.49 | 3.85 | 4.11 | 4.31 | 4.47 | 4.61 | 4.73 | 4.84 | 4.93 | 5.01 | 5.09 | 5.18 | 5.23 |
| 30 | 2.89 | 3.49 | 3.85 | 4.10 | 4.30 | 4.46 | 4.60 | 4.72 | 4.82 | 4.92 | 5.00 | 5.08 | 5.15 | 5.21 |
| 35 | 2.87 | 3.46 | 3.81 | 4.07 | 4.26 | 4.42 | 4.56 | 4.67 | 4.77 | 4.86 | 4.95 | 5.02 | 5.09 | 5.15 |
| 40 | 2.86 | 3.44 | 3.79 | 4.04 | 4.23 | 4.39 | 4.52 | 4.63 | 4.73 | 4.82 | 4.90 | 4.98 | 5.04 | 5.11 |
| 45 | 2.85 | 3.43 | 3.77 | 4.02 | 4.21 | 4.36 | 4.49 | 4.61 | 4.70 | 4.79 | 4.87 | 4.94 | 5.01 | 5.07 |
| 50 | 2.84 | 3.42 | 3.76 | 4.00 | 4.19 | 4.34 | 4.47 | 4.58 | 4.68 | 4.77 | 4.85 | 4.92 | 4.98 | 5.04 |
| 100 | 2.81 | 3.36 | 3.70 | 3.93 | 4.11 | 4.26 | 4.38 | 4.48 | 4.58 | 4.66 | 4.73 | 4.80 | 4.86 | 4.92 |
| $\infty$ | 2.77 | 3.31 | 3.63 | 3.86 | 4.03 | 4.17 | 4.29 | 4.39 | 4.47 | 4.55 | 4.62 | 4.68 | 4.74 | 4.80 |

Table A.7: Upper $\alpha$ probability point for the F distribution: $\alpha = 0.05$

Table entries are $F_{0.05;\nu_1,\nu_2}$, where $P[F > F_{0.05;\nu_1,\nu_2}] = 0.05$

| $\nu_2 \backslash \nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 25 | 30 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 244 | 246 | 248 | 249 | 250 | 251 |
| 2 | 18.6 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.5 | 19.5 | 19.5 |
| 3 | 10.1 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 | 8.63 | 8.62 | 8.59 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 | 4.52 | 4.50 | 4.46 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 | 3.83 | 3.81 | 3.77 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 | 3.40 | 3.38 | 3.34 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 | 3.11 | 3.08 | 3.04 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.07 | 3.01 | 2.94 | 2.89 | 2.86 | 2.83 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 | 2.73 | 2.70 | 2.66 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.79 | 2.72 | 2.65 | 2.60 | 2.57 | 2.53 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 | 2.50 | 2.47 | 2.43 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.60 | 2.53 | 2.46 | 2.41 | 2.38 | 2.43 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 | 2.34 | 2.31 | 2.27 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.48 | 2.40 | 2.33 | 2.28 | 2.25 | 2.20 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.42 | 2.35 | 2.28 | 2.23 | 2.19 | 2.15 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.38 | 2.31 | 2.23 | 2.18 | 2.15 | 2.10 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.34 | 2.27 | 2.19 | 2.14 | 2.11 | 2.06 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.28 | 2.20 | 2.12 | 2.07 | 2.04 | 1.99 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.23 | 2.15 | 2.07 | 2.02 | 1.98 | 1.94 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.20 | 2.13 | 2.05 | 2.00 | 1.96 | 1.19 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 | 2.18 | 2.11 | 2.03 | 1.97 | 1.94 | 1.89 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 | 2.15 | 2.07 | 1.99 | 1.94 | 1.90 | 1.85 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 | 2.13 | 2.06 | 1.97 | 1.92 | 1.88 | 1.84 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 | 2.12 | 2.04 | 1.96 | 1.91 | 1.87 | 1.82 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 | 2.10 | 2.03 | 1.94 | 1.89 | 1.85 | 1.81 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 | 2.09 | 2.01 | 1.93 | 1.88 | 1.84 | 1.79 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 | 2.00 | 1.92 | 1.84 | 1.78 | 1.74 | 1.69 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.92 | 1.84 | 1.75 | 1.69 | 1.65 | 1.59 |
| 100 | 3.94 | 3.09 | 2.70 | 2.46 | 2.31 | 2.19 | 2.10 | 2.03 | 1.97 | 1.93 | 1.85 | 1.77 | 1.68 | 1.62 | 1.57 | 1.52 |

Table A.8: Upper $\alpha$ probability point for the F distribution: $\alpha = 0.01$

Table entries are $F_{0.01;\nu_1,\nu_2}$, where $P[F > F_{0.01;\nu_1,\nu_2}] = 0.01$

| $\nu_2 \backslash \nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 25 | 30 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4052 | 4999 | 5403 | 5625 | 5764 | 5859 | 5928 | 5981 | 6022 | 6056 | 6106 | 6157 | 6209 | 6240 | 6261 | 6287 |
| 2 | 98.5 | 99.0 | 99.2 | 99.2 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.5 | 99.5 | 99.5 |
| 3 | 34.1 | 30.8 | 29.5 | 28.7 | 28.2 | 27.9 | 27.7 | 27.5 | 27.4 | 27.2 | 27.0 | 26.9 | 26.7 | 26.6 | 26.5 | 26.4 |
| 4 | 21.2 | 18.0 | 16.7 | 16.0 | 15.5 | 15.2 | 15.0 | 14.8 | 14.7 | 14.6 | 14.4 | 14.2 | 14.0 | 13.9 | 13.8 | 13.8 |
| 5 | 16.3 | 13.3 | 12.1 | 11.4 | 11.0 | 10.7 | 10.5 | 10.3 | 10.2 | 10.0 | 9.89 | 9.72 | 9.55 | 9.45 | 9.38 | 9.29 |
| 6 | 13.8 | 10.9 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.72 | 7.56 | 7.40 | 7.30 | 7.23 | 7.14 |
| 7 | 12.2 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 | 6.47 | 6.31 | 6.16 | 6.06 | 5.99 | 5.91 |
| 8 | 11.3 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.67 | 5.52 | 5.36 | 5.26 | 5.20 | 5.12 |
| 9 | 10.6 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 5.11 | 4.96 | 4.81 | 4.71 | 4.65 | 4.57 |
| 10 | 10.0 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.71 | 4.56 | 4.41 | 4.31 | 4.25 | 4.17 |
| 11 | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.79 | 4.63 | 4.54 | 4.40 | 4.25 | 4.10 | 4.01 | 3.94 | 3.86 |
| 12 | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 | 4.16 | 4.01 | 3.86 | 3.76 | 3.70 | 3.62 |
| 13 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 3.96 | 3.82 | 3.66 | 3.57 | 3.51 | 3.43 |
| 14 | 8.86 | 6.51 | 5.56 | 5.04 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.80 | 3.66 | 3.51 | 3.41 | 3.35 | 3.27 |
| 15 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.67 | 3.52 | 3.37 | 3.28 | 3.21 | 3.13 |
| 16 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 | 3.55 | 3.41 | 3.26 | 3.16 | 3.10 | 3.02 |
| 17 | 8.40 | 6.11 | 5.18 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 | 3.46 | 3.31 | 3.16 | 3.07 | 3.00 | 2.92 |
| 18 | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 | 3.37 | 3.23 | 3.08 | 2.98 | 2.92 | 2.84 |
| 19 | 8.18 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 | 3.30 | 3.15 | 3.00 | 2.91 | 2.84 | 2.76 |
| 20 | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 | 3.23 | 3.09 | 2.94 | 2.84 | 2.78 | 2.69 |
| 21 | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 | 3.17 | 3.03 | 2.88 | 2.79 | 2.72 | 2.64 |
| 22 | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.35 | 3.26 | 3.12 | 2.98 | 2.83 | 2.73 | 2.58 |
| 23 | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 3.07 | 2.93 | 2.78 | 2.69 | 2.62 | 2.54 |
| 24 | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 | 3.03 | 2.89 | 2.74 | 2.64 | 2.58 | 2.49 |
| 25 | 7.77 | 5.57 | 4.68 | 4.18 | 3.85 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 | 2.99 | 2.85 | 2.70 | 2.60 | 2.54 | 2.45 |
| 26 | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.18 | 3.09 | 2.96 | 2.81 | 2.66 | 2.57 | 2.50 | 2.42 |
| 27 | 7.68 | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.15 | 3.06 | 2.93 | 2.78 | 2.63 | 2.54 | 2.47 | 2.38 |
| 28 | 7.64 | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.23 | 3.12 | 3.03 | 2.90 | 2.75 | 2.60 | 2.51 | 2.44 | 2.35 |
| 29 | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.09 | 3.00 | 2.87 | 2.73 | 2.57 | 2.48 | 2.41 | 2.33 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 | 2.84 | 2.70 | 2.55 | 2.45 | 2.39 | 2.30 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 | 2.80 | 2.66 | 2.52 | 2.37 | 2.27 | 2.20 | 2.11 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 | 2.63 | 2.50 | 2.35 | 2.20 | 2.10 | 2.03 | 1.94 |
| 100 | 6.90 | 4.82 | 3.98 | 3.51 | 3.21 | 2.99 | 2.82 | 2.69 | 2.59 | 2.50 | 2.37 | 2.22 | 2.07 | 1.97 | 1.89 | 1.80 |

# Appendix B

# Solutions to Exercises

## B.1 Chapter 1

1.1 a. experiment

b. treatment factor is "arthroscopic surgery or sham surgery" and response variable is speed of walking after surgery

c. control for the placebo effect, the effect of responding positively just because a patient receives any kind of treatment

1.2 a. conditions are "recirculated air" and "fresh air"-observed.

b. response variable is categorical - having cold or not a week after flight

c. Possible reasons are:

i. traveling is stressful which may increase the chances of catching a cold.

ii. close contact with individuals in aircraft

1.3 a. The group of children that does not receive the massage should also receive some kind of attention from their parents. In this way both groups are getting attention

1.4 a. factor of "interest" is "body piercing or not"; response variable is categorical: smokes or not

b. observational study – the conditions "have body piercings", "not have body piercings" are observed, not assigned

c. No, we cannot conclude this, because the two groups of females, one with body piercings and the other without body piercings, may differ in other ways that may be conducive to sexual activity

1.5 a. treatments are "injection of bone marrow cells" and "injection of regular blood"

b. response variables are:

i. results of test comparing blood pressure in ankle and arm

ii. differences in oxygen inside and outside tissue

c. while the two treatments are being compared on the two legs of the same person, thus controlling for extraneous variables associated with different persons, the randomization within a person is to balance out the effects of extraneous variables associated with the two legs, such as prior differences in circulation between the two legs.

d. Yes, a block is a pair of legs on a subject. The two treatments can be compared within the same person and the results for the different persons pooled to form a conclusion

1.6 a. factor of interest is "derived strength or comfort from religion or "not derived strength or comfort from religion"; response variable is length of life

b. observational study, since the conditons are observed, not assigned to subjects

c. i. diet, perhaps people who derive strength from religion eat healthier

ii. life style, perhaps people who derive strength from religion do not smoke as much or drink alcohol as much

1.7 Since there are two different methods of memorizing difficult material, the subjects could be blocked into pairs so that within each pair the two subjects are similar with regard to characteristics that might be related to memorization ability such academic ability

1.8 a. there are 4 treatments: the blowing up of the balloons of the four colors

b. the treatments will be assigned to different time slots–thus the experimental units are time slots

c. randomization would be used by randomly assigning the treatments to the time slots. The purpose would be to balance out any effects due to time on the amount of time to blow up the balloons

d. there is direct control of peoples' different abilities to blow up balloons by having the balloons all blown up by the same person

1.9 The experimental units are pens of 3 pigs. There are 6 experimental units. Treatments are different concentrations of garlic powder in the feed of a pen of 3 animals and these are randomly assigned to pens. In order for pig to be the experimental unit the pigs would have to be fed individually (not group fed) and the different concentrations assigned to the individual pigs.

1.10 a. Experiment because the treatments: supervised aerobic exercise therapy, exercise-placebo, and usual care are (randomly) are assigned to the 108 women for the purpose of determining whether there are effects of aerobic exercise therapy on quality of life.

b. Experimental units are the 108 women who had been treated for breast cancer

c. No. In an experiment in general subjects are informed of the different kinds of treatments and sometimes they can be blinded as to what particular treatment they are getting such as in a medical study, when everyone is getting the same looking pill. In this study subjects would know what exercise they are doing – you can't disguise the exercises to look the same. Thus subjects would not be blinded.

1.11 Observational study. Condition "long term meditators" not assigned to subjects.

1.12
a. Treatments are "use dryer balls in dryer" and "do not use dryer balls in dryer."
b. Experimental units are the next 40 loads of laundry at the different time periods. The treatments are (randomly) assigned to these loads/time periods to determine which load gets the dryer balls and which loads do not.
c. machine effects - directly controlled by using only one machine. size of load, type of clothing, changes in one machine over time, are controlled by randomization.

1.13
a. Factor A: Composition of pasta (white, whole grain) Factor B: Length of Pasta (1.25, 6, 12 inches)
b. Treatments are the 6 combinations of composition and length of pasta. The treatments are (white,1.25), (white,6), (white,12), (whole grain,1.25), (whole grain, 6), (whole grain, 12).
c. Experimental units are the time slots on each day corresponding to the boilings of pasta
d. Yes, each day is a block/grouping of 6 time slots when the 6 boilings occur.
e. On each day, write the 6 combinations or treatments on six slips of paper. Mix the slips thoroughly and pull out one slip. This determines the particular combination of composition and length to boil.

1.14
a. Factor A: Type of cookie (Chips Ahoy Chocolate Chip, Oreo's,Nutter Butter) Factor B: Type of Liquid (milk, orange juice, water) in which the cookie was dipped
b. Treatments are the 9 combinations of type of cookie and type of liquid
c. Experimental units are 45 cups
d. No blocking, the cups were not grouped in any way
e. Write down the 9 combinations on slips of paper, 5 per combination, for a total of 45 slips of paper and put into a bowl. Label the 45 cups with the whole numbers 1 through 45 and put the 45 labels in a bowl. Pull out a slip for a combination and a label for a cup. The selected combination slip is the combination that is assigned the selected label for the cup. Do this 44 other times.

1.15
a. Factor A: Type of liquid (water, water+lemon, water+salt) Factor B: Type of Cup (styrofoam, paper)
b. Whole units = time slots when twenty ounces of liquid heated to 160 degrees Split units = two cups each 10 ounces of heated liquid

c. Yes, each pot of liquid is a grouping of two 10 ounce cups of liquid

d. On twelve slips of paper write the levels of liquid type, 4 per liquid type. Pull one slip out a random. This determines the liquid to be heated at a time slot. Put the words styrofoam and paper on two slips of paper. Pull one out. This determines which type of cup gets the first half of the poured liquid.

1.16 a. This is a re-using type of block design. The 3 blocks are the 3 blinds, each reused at different time slots for the three different decoy types.

b. Decoy type (taxidermy mounted decoys, plastic shell decoys, and full-bodied decoys)

c. Each blind uses all three types of decoys so the randomization would determine which decoy type is used first, which second, and which third.

1.17 Determining when water is "boiling" was subjective. The researcher was aware from the other study that adding salt to the water increased the boiling temperature. So in theory if the student is not blinded regarding the treatment and observation of the thermometer, then the student may consciously or unconsciously wait longer to determine "boiling" depending on the treatment.

1.18 a. The experimental units are the 9 heating runs.

b. The measurement units are the 5 pieces of meat in a pan for a total of 5x9 = 45 measurement units

c. There is only one replicate per treatment combination corresponding to the one heating run per treatment combination.

d. Not a valid design in the sense that the 5 pieces per treatment combination should not be treated as independent replicates. The statistical analysis would need to take into account the fact that 5 pieces were heated at a time.

1.19 a. Experimental units are the pots. There are 15 pots.

b. Variation in the nutrients in the Miracle Gro soil from pot to pot within a group. Variation in the genetics of the 3 plants per pot from pot to pot.

c. There are 3 measurement units (plants) per pot, for a total of 45 measurement units.

d. Yes. The plants are not true replicates, so there are not 15 replicates per watering regimen. There are 5 replicates per watering regimen.

1.20 a. Treatments are 1)no music and 2) music

b. Distance run on the treadmill during 10 minute period

c. Age (all 16), and all physically fit with no health problems

d. Each person is a block, here a grouping of two time slots at which the running on the treadmill was conducted for the two treatments.

e. Students might do better on the 2nd day when listening to music, not because of the music but because of self-competition, that is they are trying to do better than their first day time. Or the students might do better on the 1st day, not because of no music but because on the first day they are perhaps not as tired compared to the 2nd day.

f. In a completely randomized design each student would only do one run on the treadmill with the treatment (no music, music) randomly determined.

## B.2 Chapter 2

2.1 $\overline{y} = 73.38$, $s = 4.66$,

The 5 stretched lengths deviate above or below the mean of 73.38 cm by on average 4.66 cm.

2.2 $\overline{y} = 1.25$, $s = 0.2$, $s_{\overline{y}} = \frac{s}{\sqrt{n}} = \frac{0.2}{\sqrt{48}} = 0.03$

Sample standard deviation $s = 0.2$ measures variation of weight gains of individual pigs in the sample around the sample mean $\overline{y} = 1.25$. The standard error $s_{\overline{y}} = 0.03$ gives a crude measure of the error associated with $\overline{y} = 1.25$, treating $\overline{y}$ as an estimate of the population mean weight gain, $\mu$.

2.3
- a. The sample mean $\overline{y}$ has a normal distribution with mean $\mu_{\overline{y}} = \mu = 50$ and standard deviation $\sigma_{\overline{y}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{16}} = 1.25$
- b. Standard normal distribution
- c. $t$ distribution with $\nu = 15$ degrees of freedom

2.4
- a. 1.711
- b. 0.80

2.5
- a. Sample mean $\overline{y} = 32.6$, midpoint of interval
- b. 95% error margin = 1/2 width of interval = 7.8/2 = 3.9
- c. standard error $= \frac{s}{\sqrt{121}} = 3.9/2.045 = 1.91$
- d. We are 95% confident that the population mean number of hours studied per week is between 28.7 and 36.5 hours.
- e. Yes, different set of $n = 30$ students would result in a different mean and standard deviation and thus a different interval

2.6 $\overline{y} = 244.3$, $s = 12.4$, $t = 1.21$, $P - value = 0.2508 > 0.05$, No reason to believe true mean differs from stated. There is about a 25% chance of obtaining a sample mean as far from the hypothesize null population mean of 240 as the observed value of 244.3 due to sampling.

2.7
- a. Relevant population mean is $\mu$ = mean mental health score for the population of all long-term meditators in Australia. $H_o : \mu = 75.75$ versus $\mu \neq 75.75$.
- b. $t = 14.3$
- c. Reject $H_o$, accept $H_a$. There is evidence that the mean mental health score for the population of all long-term meditators is different than the norm score of 75.75.
- d. $(83.99, 86.31)$. Use $\infty$ row in Table A.2. We are 95% confident that the mean mental health score for the population of long-term meditators is somewhere between 83.99 and 86.31.
- e. Yes, possible Type I error.

2.8
- a. Relevant population mean of interest is mean perceived midpoint of curved glass by a population of participants. $H_o : \mu = 30.00$ versus $\mu \neq 30.00$.

b. $P < 0.001 < 0.05$ and thus reject $H_o$, accept $H_a$. There is evidence that peoples' perceived midpoint is different than the true midpoint of 30. The data points to peoples' perceived midpoint being lower.

2.9 $E = 2/3 = 0.67$ Use approximation of $E = 0.7$ in Table 2.3. Power = 0.915. There is a 91.5 chance that the one sample $t$ will conclude, based on the data, that the (pop) mean perceived midpoint is lower than 30 when in fact the population mean is lower than 30.

2.10 $E = 5/8 = 0.625$. Use $E = 0.6$ in Table 2.3. $n = 25$ gives power of 0.898, approximately equal to 0.90.

## B.3 Chapter 3

3.1 a. Type of fertilizer. Total amount of tomatoes from a plant

b. Experimental units = plant/plot combination

c. Completely randomized design. Types of fertilizer assigned completely at random to plant/plot combination. Plants/plots were not grouped in any way prior to randomization

d. Fertility of soil in plots - fertilizers randomly assigned to plots Natural variability of plants - fertilizers randomly assigned to plants

3.2 a. Candles (one scented and one unscented) are paired/grouped by the day on which they are burned.

b. Type of candle to (scented or unscented); burning time of candle

c. $\overline{d} = 25$, $s_d = 58.2$, $t = 1.36$, $df = 9$, From software two sided $P-value = 0.2075 > 0.05$ No evidence of a difference in mean burning times between the two types of candles.

3.3 a. Paired design - reusing. Subjects are used before and after being put on diet.

b. Completely randomized. Treatments "told applicant attracted to interviewer" and "not told applicant attracted" were assigned completely at random to sixty male students

c. Paired design - sorting/grouping. Letters are paired according to the destination/city.

3.4 a. Time of Period (before or after) of measurement of mental ability

b. Two time periods when measurements taken for each patient

c. No. 'Before' and 'After' are inherent characteristics of time periods.

d. Since the conditions 'Before', 'After' are not assigned at random then differences in measurements taken before and after might be confounded with other time effects.

3.5 a. Route taken

b. Travel time (hrs)

c. Driving habits of drivers.

d. Independent samples t test. Variances not assumed to be equal. $n_A = 5, \overline{y_A} = 17.7, s_A = 5.9$; $n_B = 5, \overline{y_B} = 22.0, s_B = 5.6$; $df = 7.98, t = -1.10, P = 0.3047 > 0.05$. Not enough evidence of a difference in driving times between two routes.

e. Have each driver use both routes A and B in a random order.

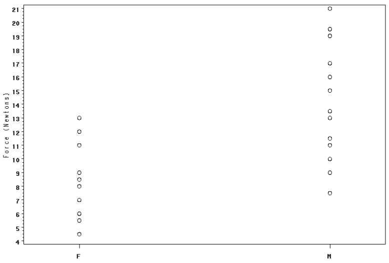
3.6 a. Paired samples design - reusing. Corresponding to each squirrel are two time periods, one when FT twigs are given and one when NFT twigs are given.

b. One sided test with $\overline{d} = 2.84$, $s_d = 2.34$, $t = 2.72$, $df = 4$, $P-value = 0.0266 < 0.05$. There is evidence that squirrels eat more of the FT twigs than the NFT twigs.

3.7 a. The plots of the stands of slash pines are paired according to location.

b. One sided test. $\overline{d} = 13$, $s_d = 127.2$, $t = 0.32$, $df = 9$, $P - value = 0.3770 > 0.05$. There is not enough evidence that 'improved' trees have a greater mean inner bark volume than 'unimproved' trees.

3.8 a. i. Lower lip forces for females lower on average than males; spread of lower lip forces for females smaller than spread for males



ii. Yes, t = -3.98, df = 24.9, P-value = 0.0005

iii. males and females are different groups which are not blocked in any way.

b. comparison of male upper lip forces with female upper lip forces, comparison of male lower lip forces with male upper lip forces

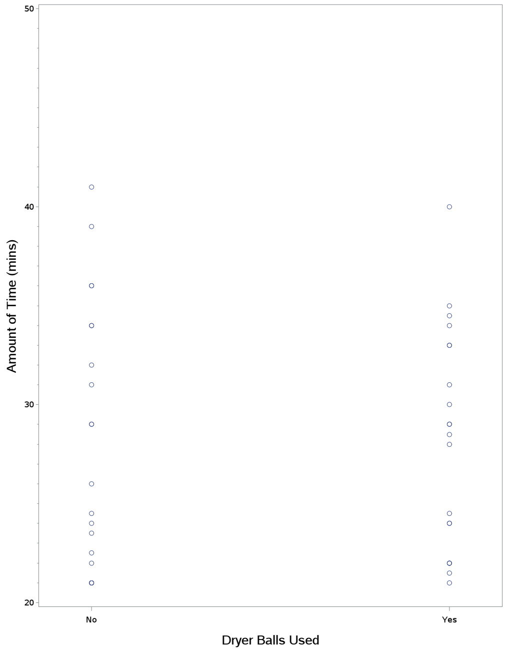
3.9 a. $H_o : \mu_d = 0$ where $\mu_d$ is the true mean of differences in perceived midpoint (straight glass - curved glass). $H_a : \mu_d \neq 0$

b. No, do not have the standard deviation of the differences

c. $P < 0.001 < 0.05$ and thus there is evidence of a difference in mean perceived midpoints for the two glass types.

3.10 $E = 50/100 = 0.5$. From Table 3.4, power is 0.799, approximately equal to 0.80, for 50 chicks per group

3.11 a. Average level of drying times and variation in drying times appear to be similar for the two groups: dryer balls used and dryer balls not used



b. Dryer balls used: mean = 28.38 min, stdev = 6.61 min. Dryer balls not used: mean = 28.30 min., stdev = 5.52 min

c. independent samples t test (unpooled variances). $H_o : \mu_1 \geq \mu_2$ versus $H_a : \mu_1 < \mu_2$ where $\mu_1$ = true mean drying time with dryer balls, and $\mu_2$ = true mean drying time without dryer balls. $t = 0.04$, $P = 0.4846 > 0.05$. There is no evidence in this study that dryer balls reduce mean drying time

## B.4 Chapter 4

4.1 a.

| | $y_{ij}$ | $=$ | $\overline{y}_{..}$ | $+$ | $A_i$ | $+$ | $e_{ij}$ |
|---|---|---|---|---|---|---|---|
| Drug A | 20 | = | 24.67 | + | -2.67 | + | -2 |
| | 22 | = | 24.67 | + | -2.67 | + | 0 |
| | 25 | = | 24.67 | + | -2.67 | + | 3 |
| | 24 | = | 24.67 | + | -2.67 | + | 2 |
| | 19 | = | 24.67 | + | -2.67 | + | -3 |
| Drug B | 21 | = | 24.67 | + | 0.33 | + | -4 |
| | 26 | = | 24.67 | + | 0.33 | + | 1 |
| | 26 | = | 24.67 | + | 0.33 | + | 1 |
| | 27 | = | 24.67 | + | 0.33 | + | 2 |
| | 25 | = | 24.67 | + | 0.33 | + | 0 |
| Drug C | 30 | = | 24.67 | + | 2.33 | + | 3 |
| | 24 | = | 24.67 | + | 2.33 | + | -3 |
| | 26 | = | 24.67 | + | 2.33 | + | -1 |
| | 25 | = | 24.67 | + | 2.33 | + | -2 |
| | 30 | = | 24.67 | + | 2.33 | + | 3 |

| | $y_{ij} - \overline{y}_{..}$ | $=$ | $A_i$ | $+$ | $e_{ij}$ |
|---|---|---|---|---|---|
| Drug A | -4.67 | = | -2.67 | + | -2 |
| | -2.67 | = | -2.67 | + | 0 |
| | 0.33 | = | -2.67 | + | 3 |
| | -0.67 | = | -2.67 | + | 2 |
| | -5.67 | = | -2.67 | + | -3 |
| Drug B | -3.67 | = | 0.33 | + | -4 |
| | 1.33 | = | 0.33 | + | 1 |
| | 1.33 | = | 0.33 | + | 1 |
| | 2.33 | = | 0.33 | + | 2 |
| | 0.33 | = | 0.33 | + | 0 |
| Drug C | 5.33 | = | 2.33 | + | 3 |
| | -0.67 | = | 2.33 | + | -3 |
| | 1.33 | = | 2.33 | + | -1 |
| | 0.33 | = | 2.33 | + | -2 |
| | 5.33 | = | 2.33 | + | 3 |

b.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Grand Mean | 1 | 9126.67 | | | |
| Drug | 2 | 63.33 | 31.67 | 4.75 | 0.0302 |
| Error | 12 | 80.00 | 6.67 | | |
| Total | 15 | 9270 | | | |

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Drug | 2 | 63.33 | 31.67 | 4.75 | 0.0302 |
| Error | 12 | 80.00 | 6.67 | | |
| Total (Corrected) | 14 | 143.33 | | | |

c. Yes, F = 4.75 > 3.89

4.2 a.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Grand Mean | 1 | 1728 | | | |
| Treatments | 4 | 85 | 21.25 | 5.06 | $< 0.01$ |
| Error | 25 | 105 | 4.2 | | |
| Total | 30 | 1918 | | | |

b. 5 treatments

c. 6 replications per treatment

d. Yes, since $P < 0.01$.

4.3 $F = \frac{0.157}{0.136} = 1.15$ with numerator $df = 2$ and denominator $df = 9$. $P = 0.3578$, not significant at the 0.05 level of significance.

4.4 a. Type of drink

b. Treatments are Cocal cola, Orange Juice, Water

c. Experimental units are cup/ice combination.

d. Ice cube size, amount of liquid, rate of pouring.

e. $y_{ij} = \overline{\mu}_{\cdot} + \alpha_i + \epsilon_{ij}$ where

- $i$ is an index on type of drink with $i = 1$(Coca cola), $i = 2(OJ)$, $i = 3(Water)$
- $y_{ij}$ is the $j^{th}$ observation on amount of time for the $i^{th}$ type of drink
- $\overline{\mu}_{\cdot}$ = true grand mean amount of time averaged over all types of drink
- $\alpha_i$ is the true effect of the $i^{th}$ type of drink on melting time $y_{ij}$
- $\epsilon_{ij}$ is the effect of extraneous variables on melting time $y_{ij}$.

f. i. $H_o : \alpha_1 = \alpha_2 = \alpha_3 = 0$
$H_a$ : not all $\alpha_i^{\prime} s = 0$

ii. $F = \frac{790.5}{3.87} = 102.2$, $P < 0.0001$. There is evidence of a difference in melting times among the three types of beverages.

iii. The true errors $\epsilon_{ij}$ are independent, normally distributed each with mean of 0 and common standard deviation $\sigma$.

4.5 a. Means are 43.1, 89.4, 68.0, and 40.5, respectively for Brands A,B,C,and D. Standard deviations are 3.0, 2.2, 2.2, and 2.4, respectively, for brands A,B,C,D. Yes.

b.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Brand | 3 | 15953.47 | 5317.82 | 866.12 | P < 0.0001 |
| Error | 36 | 221.03 | 6.14 | | |
| Total (Corrected) | 39 | 16174.5 | | | |

Estimate of common variance is MSE = 6.14

c. Yes, P < 0.0001 < 0.05.

4.6 7988.38, 84.09

4.7 a. $\overline{\mu}_. = (4+4+7)/3 = 5$. True effects are $\alpha_1 = 4-5 = -1$, $\alpha_2 = 4-5 = -1$, $\alpha_3 = 7-5 = 2$

b. By population effects model,
$Y = \mu_2 + \epsilon$ or
$Y = \overline{\mu}_. + \alpha_2 + \epsilon$
Thus decomposition of 6 is $6 = 5 + (-1) + 2$

c. True average of MSE is $E[MSE] = \sigma^2 = 2$

d. True average of MSTR is according to Section 4.4
$E[MSTR] = \sigma^2 + \frac{1}{3-1}\sum_{i=1}^{3} n_i\alpha_i^2$ or
$E[MSTR] = 2 + \frac{1}{3-1}[10(-1)^2 + 10(-1)^2 + 10(2)^2]$ or $E[MSTR] = 32$.
MSTR for a particular run of the experiment will differ from 32.

e. By the model and Section 4.4,
$\overline{y}_{1.} - \overline{y}_{2.} = (\mu_1 - \mu_2) + (\overline{\epsilon}_{1.} - \overline{\epsilon}_{2.})$
Thus even if $\mu_1 - \mu_2 = 0$, $\overline{y}_{1.} - \overline{y}_{2.}$ can be different than 0 due to $\overline{\epsilon}_{1.} - \overline{\epsilon}_{2.}$ being different than 0, that is different due to errors.

4.8 a. $y_{ij} = \overline{\mu}_. + \alpha_i + \epsilon_{ij}$
where
$i = 1(Men), 2(NCWomen), 3(OCWomen)$ indexes the different groups
$j = 1, ..., 15$ for $i = 1, 2$ and $j = 1, ..., 14$ for $i = 3$, indexes subject $j$ within a particular group $i$
$y_{ij}$ represents the $j^{th}$ observation on frequency of sexual thoughts/desire to engage in sexual activity within group $i$
$\overline{\mu}_.$ = the true/population grand mean of frequency averaged over all groups
$\alpha_i$ = the true/population effect of the $i^{th}$ group on frequency
$\epsilon_{ij}$ = the effect of extraneous variables on the $j^{th}$ frequency in the $i^{th}$ group.

b. $H_o : \alpha_1 = \alpha_2 = \alpha_3 = 0$
$H_a$ : not all $\alpha$'s are equal to 0

c. i. numerator degrees of freedom = 2 and denominator degrees of freedom equal to 41

ii. There is only a 0.001 probability that the F ratio (MSTR/MSE) will take on a value like 8.72 or higher if in fact the null hypothesis is true.

iii. P-value = 0.001 which is less than 0.05, and thus null hypothesis is rejected

4.9 (pooled samples) t test: t = 0.04, df = 38, P = 0.9691

ANOVA: F = 0.00, num df = 1, den df = 38, P = 0.9691

square of t stat = F,. df for t test equal to den df for F test, P values same for both tests.

## B.5 Chapter 5

5.1 a. m = 10

b. $t_{0.01/2} = 2.787$ for $\nu = 25$, ME = 9.7, $CL_e \geq 0.9$

c. 3.73, 12.9

d. 5.14, 12.6

e. Bonferroni, Unadjusted t procedure.

5.2 a. $6.6 < \mu_2 - \mu_1 < 15.4$, significant, $13.2 < \mu_2 - \mu_3 < 22.0$, significant, $2.2 < \mu_1 - \mu_3 < 11.0$, significant

b. 99% percent confident that all 3 intervals from part(a) are simultaneously correct.

c. narrower

5.3 $t_{0.05/2;\nu=20} = 2.086$, $q^*/\sqrt{2} = 2.95/\sqrt{2} = 2.086$

5.4 a. MSE = 0.817

b. $\nu = 40$, $t = 3$, $q_{0.05;40,3} = 3.44$ Men vs NC Women: $(0.598, 2.202)$ Men vs OC Women: $(-0.258, 1.378)$ NC Women vs OC Women: $(-1.658, -0.022)$

c. Means for Men and NC Women are significantly different (interval does not include 0), Means for NC Women and OC Women are significantly different (interval does not include 0), Means for Men and OC Women are not significantly different (interval includes 0)

5.5 a. $C_1 = (\frac{1}{4})\mu_1 + (-1)\mu_2 + (\frac{1}{4})\mu_3 + (\frac{1}{4})\mu_4 + (\frac{1}{4})\mu_5$
$C_2 = (\frac{1}{2})\mu_1 + (0)\mu_2 + (-\frac{1}{2})\mu_3 + (-\frac{1}{2})\mu_4 + (\frac{1}{2})\mu_5$
$C_3 = (1)\mu_1 + (0)\mu_2 + (0)\mu_3 + (0)\mu_4 + (-1)\mu_5$
$C_4 = (0)\mu_1 + (0)\mu_2 + (1)\mu_3 + (0)\mu_4 + (-1)\mu_5$

b. $\hat{C}_1 = 0.9$, $s\{\hat{C}_1\} = 0.52$
$\hat{C}_2 = -5.4$, $s\{\hat{C}_2\} = 0.46$
$\hat{C}_3 = 2.8$, $s\{\hat{C}_3\} = 0.66$
$\hat{C}_4 = 6.2$, $s\{\hat{C}_4\} = 0.66$

c. $H_o : C_1 = 0$, $H_a : C_1 \neq 0$, $t = 1.73$, Bonferroni t percentile = $t_{0.05/2(4);15} = 2.84$, $|1.73| < 2.84$, do not reject $H_o$, There is not enough evidence of a difference in mean abrasiveness averaged across all additives and mean abrasiveness with no additive.

$H_o : C_2 = 0$, $H_a : C_2 \neq 0$, $t = -11.74$, $|-11.74| \geq 2.84$, reject $H_o$, There is evidence of a difference in mean abrasiveness with whitener and mean abrasiveness with fluoride.

$H_o : C_3 = 0$, $H_a : C_3 \neq 0$, $t = 4.24$, $|4.24| \geq 2.84$, reject $H_o$, There is evidence of a difference in mean abrasiveness with Whitener only and mean abrasiveness with Whitener plus freshener

$H_o : C_4 = 0$, $H_a : C_4 \neq 0$, $t = 9.39$, $|9.39| \geq 2.84$, reject $H_o$, There is evidence of a difference in mean abrasiveness with Fluoride only and mean abrasiveness with Fluoride plus freshener

5.6 a. $(12.27 - 12.07) \pm \frac{4.11}{\sqrt{2}}\sqrt{5.74}\sqrt{\frac{1}{20} + \frac{1}{8}} = (-2.71, 3.11)$

b. B vs A, significant difference. Estimate that with B:Upper canines, mean bond strength is greater than mean bond strength for A: upper incisors by anywhere from 3.126 to 7.529 MPa.
B vs C, no significant difference
B vs D, significant difference. Estimate that with B:Upper canines, mean bond strength is greater than mean bond strength for D: Lower incisors by anywhere from 1.063 to 5.586 MPa.
B vs E, no significant difference.
B vs F, no significant difference.

c. We are (at least) 95% confident that all conclusions in part b. are correct.

## B.6 Chapter 6

6.1 a. $\hat{\alpha}_1 = -1$, $\hat{\alpha}_2 = 1$

b. $\hat{\beta}_1 = -2.175$, $\hat{\beta}_2 = -0.025$, $\hat{\beta}_3 = 3.025$, $\hat{\beta}_4 = -.825$

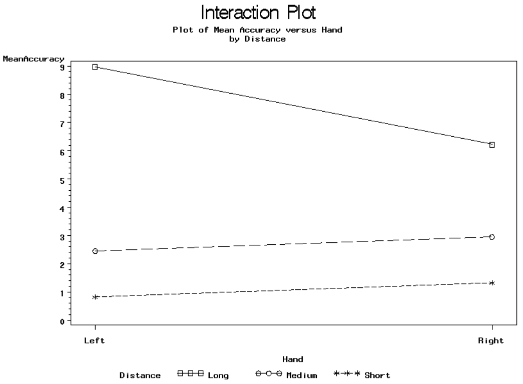
c. $\hat{\alpha\beta}_{11} = 1.85$, $\hat{\alpha\beta}_{21} = -1.85$ $\hat{\alpha\beta}_{12} = 0$, $\hat{\alpha\beta}_{22} = 0$ $\hat{\alpha\beta}_{13} = -0.65$, $\hat{\alpha\beta}_{23} = 0.65$ $\hat{\alpha\beta}_{14} = -1.2$, $\hat{\alpha\beta}_{24} = 1.2$

d. MSAB = $\frac{31.71}{3} = 3.725$, F = $\frac{10.58}{3.625} = 2.92$, $2.92 < F_{0.05;3,16} = 3.24$, No evidence of interaction

e. $SS_{Potash} = 24$, $MS_{Potash} = \frac{24}{1} = 24$, $F = \frac{24}{3.625} = 6.62$, $6.62 > F_{.05;1,16} = 4.49$, Evidence of Potash effects

f. $SS_{Nitrogen} = 87.375$, $MS_{Nitrogen} = \frac{87.375}{3} = 29.125$, $F = \frac{29.125}{3.625} = 8.03$, $8.03 > F_{.05,3,16} = 3.24$, Evidence of Nitrogen Effects

6.2 a. Differences in accuracy between levels of distance don't depend much on hand.

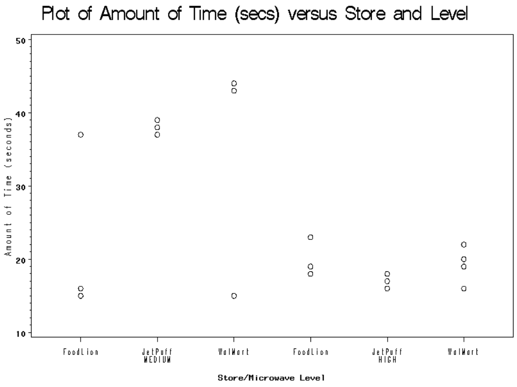


b. Test of interaction between hand and distance not significant at the 0.10 level ($F = 1.20, P - value = 0.3184, \nu_1 = 2, \nu_2 = 24$
Test of Distance Effects significant at 0.05 level $F = 15.74, P-value < 0.0001, \nu_1 = 2, \nu_2 = 24$. For i = 1 (long), i = 2 (short), and i = 3 (short), $\overline{y}_{1.} = 7.60, \overline{y}_{2.} = 2.70, \overline{y}_{3.} = 1.08$,

$$\begin{array}{rcccl} 1.88 & \leq & \mu_{1.} - \mu_{2.} & \leq & 7.92 \\ 3.50 & \leq & \mu_{1.} - \mu_{3.} & \leq & 9.55 \\ -1.40 & \leq & \mu_{2.} - \mu_{3.} & \leq & 4.65 \end{array}$$

Test of Hand effects not significant at 0.05 level $F = 0.35, P - value = 0.5607, \nu_1 = 1, \nu_2 = 24$.

6.3 a. There is more variation in times when setting is medium as compared to when setting is high. There appears to be no differences in brands when setting is high - perhaps there are differences in brands when setting is medium but this may depend on outliers. Caution should be exercised in drawing conclusions because of possible outliers and variation not being same across treatments.



b. $y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$ where

- $i$ is an index on type of brand with $i = 1$(Food Lion), $i = 2$(Jet Puff), $i = 3$(WalMart)
- $j$ is an index on microwave level with $j = 1$(Medium), $j = 2$(High)
- $k$ is an index on the amount of time for a particular brand and microwave combination
- $y_{ijk}$ is the $k^{th}$ observation on amount of time for the $i^{th}$ brand and $j^{th}$ level
- $\mu_{\cdot\cdot}$ = true grand mean amount of time averaged over all brands and levels
- $\alpha_i$ is the true effect of the $i^{th}$ brand on amount of time $y_{ijk}$
- $\beta_j$ is the true effect of the microwave level on amount of time $y_{ijk}$
- $\alpha\beta_{ij}$ is the true interaction effect between brand $i$ and level $j$ on amount of time $y_{ijk}$
- $\epsilon_{ijk}$ is the effect of extraneous variables on amount of time $y_{ijk}$.
- $\epsilon_{ijk}$ are independent normal random variables, each with mean 0 and variance $\sigma^2$

c.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Store | 2 | 300.250 | 150.125 | 2.70 | 0.0940 |
| Level | 1 | 1066.667 | 1066.667 | 19.21 | 0.0004 |
| Store*Level | 2 | 436.083 | 218.042 | 3.93 | 0.0384 |
| Error | 18 | 999.500 | 55.528 | | |
| Total (Corrected) | 23 | 2802.500 | | | |

i. Estimate of variance is 55.528

ii. Interaction is significant ($F = 3.93, P = 0.0384 < 0.10$)
Comparison of Brands when Setting = Medium (j=1):

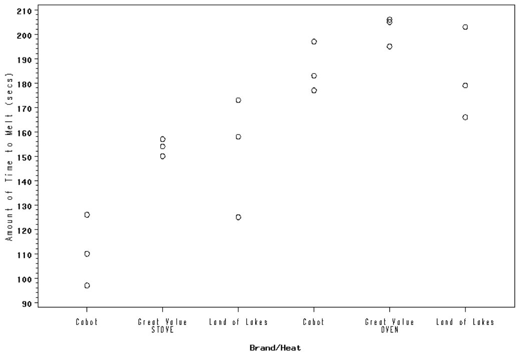
$$\begin{array}{rcccl} 3.8 & \leq & \mu_{21}-\mu_{11} & \leq & 30.7 \\ 2.1 & \leq & \mu_{31}-\mu_{11} & \leq & 29.0 \\ -15.2 & \leq & \mu_{31}-\mu_{21} & \leq & 11.7 \end{array}$$

Comparisons of Brands when Setting = High (j=2):

$$\begin{array}{rcccl} -16.0 & \leq & \mu_{22}-\mu_{12} & \leq & 11.0 \\ -13.7 & \leq & \mu_{32}-\mu_{12} & \leq & 13.2 \\ -11.2 & \leq & \mu_{32}-\mu_{22} & \leq & 11.2 \end{array}$$

6.4 a. There appears to be a heat source effect with amount of time larger for the oven. There appears to be a brand effect with Cabot and Land of Lake resulting in smaller times to melt but this comparison may depend on heat source.

Plot of Amount of Time (seconds) versus Brand and Heat Source



b. $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$ where

- $i$ is an index on brand with $i = 1$(Cabot), $i = 2$(GreatValue), $i = 3$(LandOLake)
- $j$ is an index on method with $j = 1$(Oven), $j = 2$(Stove)
- $k$ is an index on the amount of time for a particular brand and method of melting combination
- $y_{ijk}$ is the $k^{th}$ observation on amount of time for the $i^{th}$ brand and $j^{th}$ method
- $\mu_{..}$ = true grand mean amount of time averaged over all brands and methods
- $\alpha_i$ is the true effect of the $i^{th}$ brand on amount of time $y_{ijk}$
- $\beta_j$ is the true effect of the method on amount of time $y_{ijk}$
- $\alpha\beta_{ij}$ is the true interaction effect between brand $i$ and method $j$ on amount of time $y_{ijk}$
- $\epsilon_{ijk}$ is the effect of extraneous variables on amount of time $y_{ijk}$.
- $\epsilon_{ijk}$ are independent normal random variables, each with mean 0 and variance $\sigma^2$

c.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Brand | 2 | 2683.0 | 1341.5 | 6.09 | 0.0149 |
| Heat | 1 | 11806.7 | 11806.7 | 53.63 | < .0001 |
| Brand*Heat | 2 | 1470.8 | 735.4 | 3.34 | 0.0703 |
| Error | 12 | 2642.0 | 220.2 | | |
| Total (Corrected) | 17 | 18602.5 | | | |

i. Interaction is significant ($F = 3.34, P = 0.0703 < 0.10$)
Comparisons of Brands when Heat = Oven (j = 1):

$$\begin{array}{rcccl} -16.0 & \leq & \mu_{21} - \mu_{11} & \leq & 48.6 \\ -35.3 & \leq & \mu_{31} - \mu_{11} & \leq & 29.3 \\ -19.3 & \leq & \mu_{31} - \mu_{21} & \leq & 13.0 \end{array}$$

Comparisons of Brands when Heat = Stove (j=2):

$$\begin{array}{rcccl} 10.4 & \leq & \mu_{22} - \mu_{12} & \leq & 75.0 \\ 8.7 & \leq & \mu_{32} - \mu_{12} & \leq & 73.3 \\ -34.0 & \leq & \mu_{32} - \mu_{22} & \leq & 30.6 \end{array}$$

6.5 a. 4 levels of A; 3 levels of B

b. 35 + 1 = 36

c. Degrees of freedom for interaction = 6; MSAB = 80/6 = 13.3; MSE = 400/24 = 16.7; F = 13.3/16.7 = 0.80 < 2.51, not significant.

d. MSA = 310/3 = 103.3, MSE = 400/24 = 16.7, F = 103.3/16.7 = 6.19 > 3.01, significant

e. SSB = 100, MSB = 100/2 = 50.0, MSE = 400/24 = 16.7, F = 50.0/16.7 = 2.99 < 3.40, not significant

6.6 All diets contain a basal diet. What changes is whether or not cholesterol and thiouracil is added. Let factor A: cholesterol or not and factor B: thiouracil or not. Then the four combinations of A and B determine the diets in the table.

## B.7 Chapter 7

7.1 a. Time periods at which the two treatments (waterbed, regular) were assigned for each baby. Total of 18 EUs, 2 for each baby

b. An example of a completely randomized design, say 18 babies, are randomly assigned to the two treatments with 9 babies sleeping on the waterbed and 9 babies sleeping on a regular mattress.

7.2 a. Type C blocking; a block is a sample of coal

b. Experimental units are halves of the sample assigned at random for each sample to the two labs.

c. In a completely randomized design the 10 samples could have been assigned completely at random to the two labs, with Lab1 receiving 5 samples and Lab2 receiving a different 5 samples.

7.3 a. Time to exhaustion; Diets 1, 2, 3; Time slots (3 day periods) assigned to 3 diets for each person.

b. Subject. Variation in subjects which might affect time to exhaustion such as general health, weight.

c. Have 18 subjects, say, assigned completely at random to the 3 diets, with 6 persons per diet. Different groups of subjects for the 3 diets.

7.4 a. Replication/Day

b. Time slots of the burning of a candle

c. Location effect on table, changes in micro-environment from one candle lighting to another.

d. In a completely randomized design the 28 time slots at which the candles are to be burned would be randomly assigned to the 28 candles. With this design in theory 8 tan candles might be lit first, etc.

e. $y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$ where

- $i = 1, 2, ..., 7$ is an index on the replication/day $j = 1, 2, 3, 4$ is an index on the color of the candle $j = 1(Tan)$, $j = 2(Blue)$, $j = 3(Purple)$, $j = 4(White)$.
- $y_{ij}$ is the observation on burning time for the $i^{th}$ block and $j^{th}$ color.
- $\mu_{..}$ is the grand mean of burning time
- $\rho_i$ is the true effect of the $i^{th}$ block on the burning time $y_{ij}$
- $\tau_j$ is the true effect of the $j^{th}$ color on the burning time $y_{ij}$
- $\epsilon_{ij}$ is the effect of extraneous variable on the burning time $y_{ij}$

f.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Color | 3 | 12398.4 | 4132.8 | 2.77 | 0.0713 |
| Day | 6 | 17795.7 | 2966.0 | 1.99 | 0.1204 |
| Error | 18 | 26820.9 | 1490.0 | | |
| Total (Corrected) | 27 | 57015.0 | | | |

$F = 2.77, P = 0.0713$, not enough evidence at $\alpha = 0.05$ of differences in mean burn time across colors.

7.5 a. Groups of 48 hens (4 levels) and periods (4 levels)

b. EUs are time periods at which mean eggs are calculated, total of 16 EUs, 4 for each group of 48 hens

c. $y_{ijk} = \mu_{...} + \rho_i + \kappa_j + \tau_k + \epsilon_{ijk}$
$i = 1, 2, 3, 4$ representing the 4 groups of 48 hens
$j = 1, 2, 3, 4$ representing the 4 periods
$k = 1, 2, 3, 4$ representing the 4 concentrations of molasses with $k = 1(0g/kg)$, $k = 2(70g/kg)$, $k = 3(140g/kg)$, $k = 4(210g/kg)$
and where

- $y_{ijk}$ represents the mean egg weight of the $i^{th}$ group of hens at the $j^{th}$ period, getting the $k^{th}$ concentration of molasses
- $\mu_{...}$ represents the true grand mean of mean egg weight
- $\rho_i$ $(i = 1, ..., r)$ represents the true main effect of the $i^{th}$ group of chickens on mean egg weight
- $\kappa_j$ $(j = 1, ..., r)$ represents the true main effect of the $j^{th}$ period on mean egg weight
- $\tau_k$ $(k = 1, ..., r)$ represents the true main effect of the $k^{th}$ concentration of molasses on mean egg weight
- $\epsilon_{ijk}$ represents the effects of extraneous variables on mean egg weight

Assume that the $\epsilon_{ijk}$'s are independent normal random variables,each with mean of 0 and common variance $\sigma^2$
Assume that there is no interaction between each of the blocking factors, group, period and molasses concentration

d.

| Source of Variation | df | SS | MS | F | Pvalue |
|---|---|---|---|---|---|
| Group | 3 | 9.03 | 3.01 | 8.26 | 0.0150 |
| Period | 3 | 5.07 | 1.69 | 4.64 | 0.0525 |
| Molasses | 3 | 6.37 | 6.37 | 5.83 | 0.0327 |
| Error | 6 | 2.19 | 0.36 | | |
| Total | 15 | 22.66 | | | |

Significant difference in mean egg weights among the diets/molasses concentrations ($F = 5.83, P = 0.0327$

e. 0 - 70: (-1.427152,1.527152)
0 - 140: (-0.602152,2.352152)
0 - 210: (0.047848,3.002152)
70 - 140: (-0.652152, 2.302152)
70 - 210: (-0.002152,2.952152)
140 - 210: (-0.827152, 2.127152)
There was a significant difference in mean egg weight only with the 0 and 210 mg/kg diets.

7.6 a.

| Source of Variation | df | SS | MS | F | Pvalue |
|---|---|---|---|---|---|
| Car | 3 | 149.2 | 49.7 | 11.5 | 0.0067 |
| Position | 3 | 573.2 | 191.1 | 44.3 | 0.0002 |
| Brand | 3 | 185.2 | 61.7 | 14.3 | 0.0038 |
| Error | 6 | 25.9 | 4.3 | | |
| Total | 15 | 933.4 | | | |

Significant differences in treadwear among the brands ($F = 14.3, P = 0.0038$

b. A-B: (-0.333230, 9.833230)
A-C: (-9.083230, 1.083230)
A-D: (-1.583230, 8.583230)
B-C: (-13.833230, -3.666770)
B-D: (-6.333230, 3.833230)
C-D: (2.416770, 12.583230)
Significant differences in treadwear between B and C, C and D.

7.7 a. Blocks are the subjects, 12 altogether. Treatments are the three levels of caffeine (0, 1, 3 mg per kg body weight). Experirmental units are the time slots/days since treatments are assigned (at random) to days for each subject.

b. The type of blocking is Type B[i], reusing subjects at different time slots/days in order to receive different treatments. To balance out any order effects of the treatments on the responses.

c. GM = 59. SSTR = 96, MSTR = 96/(3-1) = 48. Not possible to determine MSE, would need individual heart rates in order to calculate block effects.

7.9 a. Factors are type of container and type of liquid. Response variable is number of minutes for ice cube to melt.

b. Treatments are the combinations (styrofoam, cola), (styrofoam, water), (styrofoam, juice), (glass,cola), (glass, water), (glass,juice), (plastic, cola), (plastic, water), (plastic,juice). There are 9 treatments.

c. Since melting is done one cup at a time then experimental units are cups at time slots. 30 experimental units.

d. One extraneous variable would be the micro-environment at the time when a cube is being melted. Another extraneous variable would be the size of the cube which may vary from melting to melting. These two variables are controlled by randomization of treatments to cubes and time slots.

e. Blocks are the days on which complete replications of the 9 treatments is conducted. There are 5 blocks.

7.11 a. Factor of interest = air conditioning system with 4 levels

b. Twenty homes, 4 per block. Within each block, four air conditioning systems were randomly assigned to the homes.

c. Response variable = electricity usage (KWh) in 1-month period

d. Purpose of blocking is to reduce the size of experimental error by removing the blocking effects (floor space, type of insulation, etc.), likely resulting in a more precise comparison of the four air conditioning system since the systems are now compared within each block where floor space, type of insulation, etc. are similar. The effects of floor space, insulation system, etc. are taken out of experimental error and now represented as a non-error term in the model.

e. Assuming 20 homes, in a completely randomized design, the 20 homes would be assigned completely at random to the 4 air-conditioning systems, 5 per system. The homes would NOT be blocked/grouped before randomization. The effects of floor space, type of insulation, etc. would be part of experimental error.

f.

$$y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij} \tag{B.1}$$

where

- $y_{ij}$ = electricity usage for home in the $i^{th}$ block (i=1,2,3,4,5) and $j^{th}$ system (j=1,2,3,4)
- $\mu_{..}$ represents the true grand mean of electricity usage
- $\rho_i$ represents the true effect of $i^{th}$ level of the blocking variable (floor space, type of insulation, etc.) on electricity usage
- $\tau_j$ represents the true effect of the $j^{th}$ level of system on electricity usage
- $\epsilon_{ij}$ represents as usual the effects of extraneous variables on the observation of $y_{ij}$, electricity usage at the $ij$ combination of the blocking variable (floor space, type of insulation) and system type. This would include the effects of extraneous variables other than those used to block the homes.
  Assume that the 20 errors are values of independent normal random variables, each with mean of 0 and same variance, $\sigma^2$. Assume that there is no interaction between the blocking factor and air-conditioning system.

g.

| Source of Variation | Df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| System | 3 | 930.15 | 310.05 | 2.18 | 0.1431 |
| Block | 4 | 4959.70 | 1239.93 | 8.73 | 0.0015 |
| Error | 12 | 1705.10 | 142.09 | | |
| Total (Corrected) | 19 | 7594.95 | | | |

h. If the analysis had been analyzed incorrectly as a completely randomized design (CRD) then the error would be 116 minus the mean of usage for all homes with system 1, 109.4, and thus 116 - 109.4 = 6.6
Using the correct block design analysis the error would be
116-(grand mean + block 1 effect + system 1 effect)

Grand mean = 118.45
Block 1 effect = block 1 mean - grand mean = 142.25 - 118.45 = 23.8
System 1 effect = system 1 mean - grand mean = 109.40 - 118.45 = -9.05
Correct error using block design model is thus:
116 - (118.45 + 23.8 - 9.05) = 116 - 133.2 = -17.2
Note that block design error, -17.2, can also be calculated as
-17.2 = incorrect error - block effect = 6.6 - 23.8 = -17.2.

## B.8 Chapter 8

8.1 No, the errors appear to be dependent. After the residual at time 1, there appears to be an upward trend, implying that time to melt was lower than expected in the early trials and then higher than expected in the later trials.

8.2 a.

| Fan Status | Flavor | Burning Time | TimeOrder | Predicted | Residual |
|---|---|---|---|---|---|
| On2 | Vanilla | 15 | 5 | 14.7 | 0.3 |
| On2 | Vanilla | 16 | 11 | 14.7 | 1.3 |
| On2 | Vanilla | 13 | 18 | 14.7 | -1.7 |
| On2 | Cinnamon | 14 | 2 | 15.7 | -1.7 |
| On2 | Cinnamon | 17 | 8 | 15.7 | 1.3 |
| On2 | Cinnamon | 16 | 14 | 15.7 | 0.3 |
| On4 | Vanilla | 19 | 6 | 19.3 | -0.3 |
| On4 | Vanilla | 21 | 15 | 19.3 | 1.7 |
| On4 | Vanilla | 18 | 17 | 19.3 | -1.3 |
| On4 | Cinnamon | 21 | 1 | 20.3 | 0.7 |
| On4 | Cinnamon | 20 | 3 | 20.3 | -0.3 |
| On4 | Cinnamon | 20 | 12 | 20.3 | -0.3 |
| Off | Vanilla | 27 | 4 | 27.0 | 0.0 |
| Off | Vanilla | 29 | 7 | 27.0 | 2.0 |
| Off | Vanilla | 25 | 10 | 27.0 | -2.0 |
| Off | Cinnamon | 26 | 9 | 27.3 | -1.3 |
| Off | Cinnamon | 28 | 13 | 27.3 | 0.7 |
| Off | Cinnamon | 28 | 16 | 27.3 | 0.7 |

b. Check for assumption of constant variance of error terms. Plot does not indicate extreme violation of assumption. Spread of points (burn times) roughly same across treatments.

c. Check for assumption of independence of errors. No pattern of residuals versus time and thus no evidence assumption violated.

d. Check for assumption of constant variance of error terms. Plot does not indicate any gross violation of assumption - spread of points (residuals) roughly same across all treatments

e. Check for assumption of constant variance of error terms. Plot does not indicate any gross violations of assumption. No widening or narrowing of plot as predicted burn time increases.

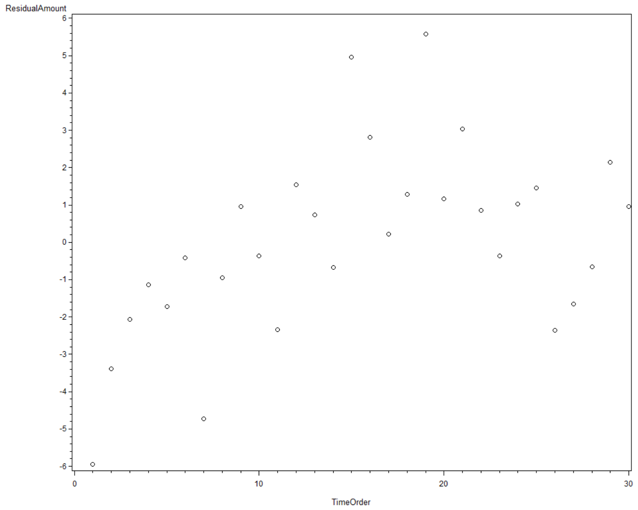
f. Histogram of residuals - used to check normality of errors. No evidence that assumption is grossly violated. Histogram of residuals approximately symmetric bell-shaped.

g. Quantile-quantile plot of residuals - used to check normality of errors. No evidence that assumption is grossly violated - plot is roughly linear.

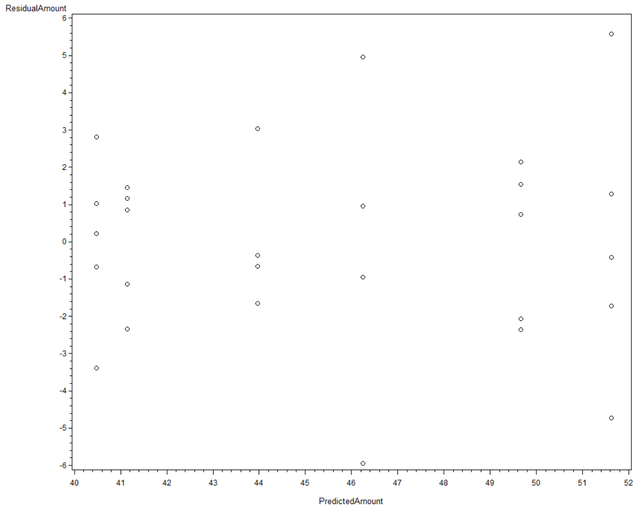
8.3 Yes, homogeneity of variance assumption. Sample standard deviations increase with increasing mean. Largest standard deviation is approximately 10 times the smallest standard deviation.

8.4 Residuals are given in the table below.

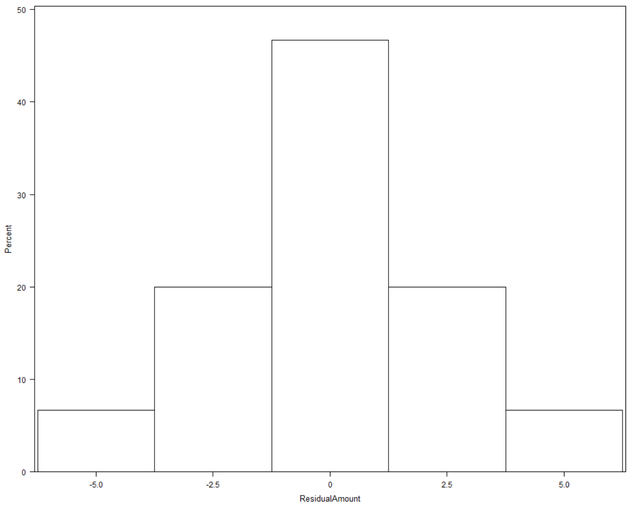
| Time Order | Cup | Liquid | Amount of Time(mins) | Predicted | Residual |
|---|---|---|---|---|---|
| 1 | Paper | Coffee | 40.3 | 46.24 | -5.94 |
| 2 | Plastic | Coffee | 37.1 | 40.48 | -3.38 |
| 3 | Styrofoam | Water | 47.6 | 49.66 | -2.06 |
| 4 | Plastic | Water | 40.0 | 41.14 | -1.14 |
| 5 | Styrofoam | Coffee | 49.9 | 51.62 | -1.72 |
| 6 | Styrofoam | Coffee | 51.2 | 51.62 | -0.42 |
| 7 | Styrofoam | Coffee | 46.9 | 51.62 | -4.72 |
| 8 | Paper | Coffee | 45.3 | 46.24 | -0.94 |
| 9 | Paper | Coffee | 47.2 | 46.24 | 0.96 |
| 10 | Paper | Water | 43.6 | 43.96 | -0.36 |
| 11 | Plastic | Water | 38.8 | 41.14 | -2.34 |
| 12 | Styrofoam | Water | 51.2 | 49.66 | 1.54 |
| 13 | Styrofoam | Water | 50.4 | 49.66 | 0.74 |
| 14 | Plastic | Coffee | 39.8 | 40.48 | -0.68 |
| 15 | Paper | Coffee | 51.2 | 46.24 | 4.96 |
| 16 | Plastic | Coffee | 43.3 | 40.48 | 2.82 |
| 17 | Plastic | Coffee | 40.7 | 40.48 | 0.22 |
| 18 | Styrofoam | Coffee | 52.9 | 51.62 | 1.28 |
| 19 | Styrofoam | Coffee | 57.2 | 51.62 | 5.58 |
| 20 | Plastic | Water | 42.3 | 41.14 | 1.16 |
| 21 | Paper | Water | 47.0 | 43.96 | 3.04 |
| 22 | Plastic | Water | 42.0 | 41.14 | 0.86 |
| 23 | Paper | Water | 43.6 | 43.96 | -0.36 |
| 24 | Plastic | Coffee | 41.5 | 40.48 | 1.02 |
| 25 | Plastic | Water | 42.6 | 41.14 | 1.46 |
| 26 | Styrofoam | Water | 47.30 | 49.66 | -2.36 |
| 27 | Paper | Water | 42.3 | 43.96 | -1.66 |
| 28 | Paper | Water | 43.3 | 43.96 | -0.66 |
| 29 | Styrofoam | Water | 51.8 | 49.66 | 2.14 |
| 30 | Paper | Coffee | 47.2 | 46.24 | 0.96 |

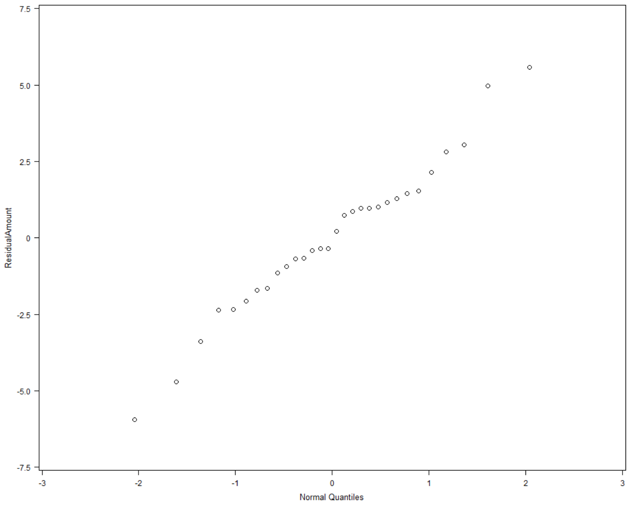


The assumption of independence of errors appears to be in question. The early trials appear to result in more negative residuals implying amounts of time smaller than predicted.



The assumption of homogeneity of error variances appears to be approximately satisfied as there appears to be no relationship between spread of the residuals and predicted amount of time.





The assumption of normality of the errors holds approximately as the residuals have a bell shaped histogram and the qqplot shows a linear relationship.

## B.9 Chapter 9

9.1 a. Whole plot factor is temperature. Whole plot experimental unit is growth chamber. Variations in treatment (temperature) within a chamber; environmental location of chamber

b. Completely randomized design. Chambers are not blocked. Temperatures assigned completely at random to chambers.

c. Split plot factor is strain of petunia (A,B,C). Split plot experimental unit is pot/location in chamber. Some experimental error factors are variation in pot soil, locations of pots within chambers.

d. The whole units, here growth chambers, serve as blocks or groups of petunias.

e.

$$y_{ijk} = \mu + \alpha_i + \epsilon^{w}_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^{s}_{ijk} \tag{B.2}$$

where $i = 1, 2, 3$ indexes temperature, $j = 1, 2, 3$ strain of petunia, and $k = 1, 2, 3$ indexes chamber associated with a particular temperature, and

- $Y_{ijk}$ is the growth of the petunia, at the $i^{th}$ level of temperature, $k^{th}$ chamber nested within the $i^{th}$ level of temperature, and $j^{th}$ level of petunia strain.
- $\mu$ is the grand mean of growth averaged over a population of chambers, all levels of temperature, and all levels of strain of petunia.
- $\alpha_i$ is the true effect of the $i^{th}$ level of the temperature on growth of petunia.
- $\epsilon^{w}_{k(i)}$ is the error term for the $k^{th}$ chamber nested within the $i^{th}$ level of the temperature, representing the effect of extraneous variables associated with the chamber.
- $\beta_j$ is the true effect of the $j^{th}$ level of strain on growth
- $\alpha\beta_{ij}$ is the true interaction effect on growth of the $i^{th}$ level of temperature and the $j^{th}$ strain.
- $\epsilon^{s}_{ijk}$ is the error term for the split unit, here pot/location, associated with the $i^{th}$ level of temperature, $k^{th}$ chamber nested under the $i^{th}$ level of temperature,and the $j^{th}$ strain, representing the effect of extraneous variables for this unit.

  It is assumed that the 9 whole unit(chamber) errors are normally distributed each with mean of 0 and constant variance. It is assumed that the 27 split unit (pot) errors are normally distributed each with mean of 0 and constant variance. It is assumed that the whole unit and split unit errors are independent. It is assumed that the growths within each chamber are equally correlated.

9.2 a. Whole plot factor is music type. Whole plot experimental unit is session or time period of session. Variation in environmental conditions associated with different sessions such as other noise, etc.

b. Completely randomized design. Music types assigned completely at random to 9 sessions (sessions are not grouped in any way).

c. Split plot factor is font of list of words. Split plot experimental unit is subject. Extraneous variables associated with subject are memorizing ability, health of person, etc.

d. The whole units, here sessions, serve as blocks of three subjects.

e.

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{B.3}$$

where $i = 1, 2, 3$ indexes music type, $j = 1, 2, 3$ font color, and $k = 1, 2, 3$ indexes session associated with a particular music type, and

- $Y_{ijk}$ is the proportion of correctly memorized words, at the $i^{th}$ level of music, $k^{th}$ session nested within the $i^{th}$ level of music, and $j^{th}$ level of font color.
- $\mu$ is the grand mean of proportion of correctly memorized words averaged over a population of sessions, all levels of music, and all levels of font color.
- $\alpha_i$ is the true effect of the $i^{th}$ level of the music type on proportion of correctly memorized words.
- $\epsilon^w_{k(i)}$ is the error term for the $k^{th}$ session nested within the $i^{th}$ level of music type, representing the effect of extraneous variables associated with the session.
- $\beta_j$ is the true effect of the $j^{th}$ level of font color on proportion of correctly memorized words
- $\alpha\beta_{ij}$ is the true interaction effect on proportion of correctly memorized words of the $i^{th}$ level of music type and the $j^{th}$ font color.
- $\epsilon^s_{ijk}$ is the error term for the split unit, here subject, associated with the $i^{th}$ level of music type, $k^{th}$ session nested under the $i^{th}$ level of music type,and the $j^{th}$ font color, repesenting the effect of extraneous variables for the subject.
  It is assumed that the 9 whole unit(session) errors are normally distributed each with mean of 0 and constant variance. It is assumed that the 27 split unit (subject) errors are normally distributed each with mean of 0 and constant variance. It is assumed that the whole unit and split unit errors are independent. It is assumed that the fractions of words correct within each session are equally correlated.

9.3 a. Whole unit factor is oven temperature. Whole unit is oven run/session. Characteristics of oven runs/sessions such as slight variations in oven temperature at different runs with same temperature setting

b. Completely randomized design. Oven temperatures are assigned completely at random to the runs. Runs are not grouped in any way and then temperatures assigned at random within groups.

c. Split unit factor is Type of Ice Cube (bottle, tap, and salt). Split unit factor is type of cube. Extraneous variables include size of cube, temperature variability within parts of oven, etc.

d.

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{B.4}$$

where $i = 1, 2, 3$ indexes oven temperature, $j = 1, 2, 3$ ice type, and $k = 1, 2, 3$ indexes oven run/session associated with a particular oven temperature

- $Y_{ijk}$ is the amount of time for cube to melt, at the $i^{th}$ level of temperature, $k^{th}$ oven run nested within the $i^{th}$ level of temperature, and $j^{th}$ level of ice cube type.
- $\mu$ is the grand mean of amount of time averaged over a population of oven runs/sessions, all levels of temperature, and all levels of ice cube type.
- $\alpha_i$ is the true effect of the $i^{th}$ level of temperature on amount of time for ice cube to melt.
- $\epsilon^w_{k(i)}$ is the error term for the $k^{th}$ oven run nested within the $i^{th}$ level of temperature, representing the effect of extraneous variables associated with the run
- $\beta_j$ is the true effect of the $j^{th}$ level of ice type on amount of time to melt
- $\alpha\beta_{ij}$ is the true interaction effect on amount of time to melt for the $i^{th}$ level of temperature and the $j^{th}$ ice type.
- $\epsilon^s_{ijk}$ is the error term for the split unit, here ice cube, associated with the $i^{th}$ level of temperature, $k^{th}$ oven run nested under the $i^{th}$ level of temperature,and the $j^{th}$ type, repesenting the effect of extraneous variables for the cube.
  It is assumed that the 9 whole unit (oven runs) errors are normally distributed each with mean of 0 and constant variance. It is assumed that the 27 split unit (ice cube) errors are normally distributed each with mean of 0 and constant variance. It is assumed that the whole unit and split unit errors are independent. It is assumed that the amounts of time of within each run are equally correlated.

e.

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Temperature | 2 | 89149.4 | 44574.7 | 10.32 | 0.0114 |
| Error (Run(Temperature)) | 6 | 25906.4 | 4317.7 | | |
| IceType | 2 | 59849.2 | 29924.6 | 13.53 | 0.0008 |
| Temperature*IceType | 4 | 10491.3 | 2622.8 | 1.19 | 0.3659 |
| Error (Cube) | 12 | 26544.2 | 2212.0 | | |
| Total (corrected) | 26 | 211940.5 | | | |

f. No evidence of interaction (F = 1.19, P-value = 0.3659)

g. Evidence of temp main effects (F = 10.32, P-value = 0.0114)
Tukey-Kramer pairwise comparisons of temperatures with $i = 1(250), i = 2(300), i = 3(350)$

$$\begin{array}{rcccl} -6.9 & \leq & \mu_{1\cdot} - \mu_{2\cdot} & \leq & 183.1 \\ 44.1 & \leq & \mu_{1\cdot} - \mu_{2\cdot} & \leq & 234.1 \\ -44.0 & \leq & \mu_{2\cdot} - \mu_{3\cdot} & \leq & 146.0 \end{array}$$

Evidence of Ice Type main effects (F = 13.53, P-value = 0.0008)
Tukey-Kramer pairwise comparisons of ice cube type with $j = 1(tap), j = 2(bottle), j = 3(salt)$

$$\begin{array}{rcccl} -15.8 & \leq & \mu_{\cdot 1} - \mu_{\cdot 2} & \leq & 102.5 \\ 55.1 & \leq & \mu_{\cdot 1} - \mu_{\cdot 3} & \leq & 173.4 \\ 11.7 & \leq & \mu_{\cdot 2} - \mu_{\cdot 3} & \leq & 130.0 \end{array}$$

h. Normality and homogeneity of split plot errors satisfied approximately.

9.4 a. Whole plot factor is fertilizer. Whole plot experimental unit is plot.

b. Block design. Plots grouped by blocks and then fertilizer assigned at random to plots within a block.

c. Split plot factor is variety of wheat. Split plot experimental unit is smaller plot.

d. The model for the split plot design in this example is:

$$y_{ijk} = \mu + \alpha_i + \rho_k + \epsilon_{ik}^{w} + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}^{s} \tag{B.5}$$

where

- $Y_{ijk}$ is the observation on yield at the $i^{th}$ fertilizer, $k^{th}$ block, and $j^{th}$ wheat variety
- $\mu$ is the grand mean of yields averaged over a population of blocks, all levels of fertilizer, and both wheat varieties.
- $\alpha_i$ is the true effect of the $i^{th}$ level of fertilizer on yield
- $\rho_k$ is the true effect of the $k^{th}$ level of block
- $\epsilon_{ik}^{w}$ is the error term for the whole plot assigned to fertilizer $i$ in block $k$ representing the effect of extraneous variables associated with the whole plot.
- $\beta_j$ is the true effect of the $j^{th}$ level of variety on yield
- $\alpha\beta_{ij}$ is the true interaction effect on the yield of the $i^{th}$ level of fertilizer and the $j^{th}$ level of wheat variety
- $\epsilon_{ijk}^{s}$ is the error term for the smaller plot receiving the $j^{th}$ level of wheat variety in the $k^{th}$ block for fertilizer $i$, representing the effects of extraneous variables associated with the smaller plot.
  It is assumed that the 8 whole unit (plot) errors are normally distributed each with mean of 0 and constant variance. It is assumed that the 16 split unit (smaller plots) errors are normally distributed

each with mean of 0 and constant variance. It is assumed that the 2 block effects are normally distributed each with mean of 0 and constant variance. It is assumed that the whole unit, split unit, and block effects are independent. It is assumed that the yields within each yield are equally correlated.

e.

| Source of Variation | df | SS | MS | F | P-value |
|---|---|---|---|---|---|
| Block | 1 | 131.1 | 131.1 | 56.77 | 0.0048 |
| Fertilizer | 3 | 40.2 | 13.4 | 5.80 | 0.0914 |
| Error (Block*Fertilizer)) | 3 | 6.93 | 2.30 | | |
| Wheat | 1 | 2.25 | 2.25 | 1.07 | 0.3599 |
| Fertilizer*Wheat | 3 | 1.55 | 0.52 | 0.25 | 0.8612 |
| Error | 4 | 8.43 | 2.12 | | |
| Total (corrected) | 15 | 190.4 | | | |

f. No evidence of interaction (F = 0.25, P-value = 0.8612)

g. No evidence of fertilizer main effects (F = 5.80, P-value = 0.0914)
No evidence of wheat variety main effects (F = 1.07, P-value = 0.3599)

## B.10 Chapter 10

10.1 a. 9 planes. Variation from plane to plane for given design such as quality, weight of printer paper, inability to construct plane exactly same way from plane to plane

b. 3 different throws with same plane. Inability to throw the same plane with the same force, direction, from throw to throw, drafts in dormitory hallway.

c. The model for the data is:

$$y_{ijk} = \mu + \alpha_i + \epsilon_{ij} + \eta_{ijk} \tag{B.6}$$

where $i = 1, ..., t = 3$, with $i$ being an index on the plane design, $j = 1, ..., r = 3$, with $j$ being an index on the replicate plane at each level of design, and $k = 1, ..., n = 3$, with $k$ being an index on the measurement unit (throw) for the $j^{th}$ replicate plane for the $i^{th}$ design.

- $y_{ijk}$ is the value of flight distance for the $k^{th}$ flight for the $j^{th}$ replicate plane with the $i^{th}$ design.
- $\mu$ is the grand mean of flight distances averaged over a population of flights, all design levels and a population of planes.
- $\alpha_i$ is the true effect of the $i^{th}$ design on flight distance
- $\epsilon_{ij}$ is the error term associated with the $j^{th}$ replicate plane of the $i^{th}$ design, representing the effect of extraneous variables associated with the plane.
- $\eta_{ijk}$ is the error for the $k^{th}$ flight of the $j^{th}$ replicate plane associated with the $i^{th}$ design, representing the effect of extraneous variables associated with the flight

The model assumes that the experimental or plane errors are independent normal random variables each with mean 0 and common variance $\sigma^2_\epsilon$ and that the flight errors are independent normal random variables each with mean 0 and common variance $\sigma^2_\eta$. It is also assumed that the plane errors are independent of flight errors.

d. Mean distances for Dart, Lightning, and Thunder are 575.2, 474.2, and 171.9, respectively. Mean Square for Design = 396426.3, Mean Square Error for Planes is 230.9, Mean Square Error for Flights is 230.9. F = 601.8, $P < 0.0001$ with 2 and 6 numerator, denominator degrees of freedom.

e. Calculate mean distance for each plane and analyze means using the analysis for one factor completely randomized design (Ch. 4)

10.2 a. Experimental units are pens. Extraneous variables include microclimate/location effects of pen, variation in application of test diet to pen, social interaction of 8 hens in pen

b. Measurement units are individual hens within a pen. Extraneous variables include variation in starting weight of hen, genetics affecting weight gain of individual hens.

c. Yes, since we have standard deviations, MSE could be calculated by Equation 4.9, Chapter 4. Also, since we have means treatment effects could be calculated as well and thus a test for diet effects could be conducted.

10.3 a. Whole units are large time slots for a particular balloon being blown up. Split units are smaller time slots at which particular balloon particular repeated blown up.

b. Whole unit factor is color of balloon blown up at a large time slot. Split unit factor is time scale of the smaller time slots at which the balloon is blown up (0, 5, 10 minutes).

c. Completely randomized design. Colors of balloons assigned completely at random to the larger time slots.

d. Each balloon serves as a block of 3 times slots when inflation time is measured.

10.4 a. Brand of Gum and Chewing Time

b. One type, time slot when a stick of gum of a particular brand and chewing time is chewed.

c. No. While there is an amount of time scale, each stick of gum is not repeated measured through time. Each stick of gum is only chewed once for a certain amount of time. The design is a two factor completely randomized design (Chapter 6).

10.5 Two types of experimental units are individual planes and time slots when individual plane is flown. Whole unit factor is plane design. Split unit factor is time scale (time = 1, 2, 3) corresponding to the three flights. Blocks are individual planes or set of 3 time slots corresponding to individual planes.

10.6 a. Whole units are branches. Split units are time points when repeated measures on number of mealybugs is observed.

b. Whole unit factor is treatment for mealybugs. Split unit factor is time scale for day on which repeated observations on number of mealybug taken (baseline before treatment, Day 1, 2, 3,4,5,6,7)

c. Whole units are branches which are blocked by plant.

d. Two types of blocking. Blocking of branches by plant. Each branch is a block of time points for the repeated measures.

10.7 a. Whole units are the beakers with their liquids. Position of beakers on the burner, amount of liquid in a beaker.

b. Split units are the time slots or occasions for each beaker/liquid when the temperatures are measured. Extraneous variables include variables in effect at the particular time slots, such as position of thermometer in the beaker, environmental conditions at that occasion.

c. Whole unit factor is type of liquid. Split unit factor is time variable, here number of minutes (1,2,3,4,5) after liquid removed from the burner.

d. Whole units are blocked by Day. Randomization occurs within each day. Within each day the three liquids are assigned at random to the three beakers.

e. There are two types of blocking. One blocking variable is Day, a grouping of beakers by Day (Type A). The other blocking variable is the beaker/liquid since each beaker with its liquid is reused (Type B[ii]).

10.8 a. Type of Box Packaging. Treatments are control, wax paper, metal foil, plastic, metal foil/plastic.

b. Experimental units are boxes. 5 boxes per treatment for a total of 25 boxes.

c. Measurement units are crackers, 3 per box, for a total of 75 crackers.

d. Randomized complete block design. Experimental units are blocked by chamber.

10.9 a. Whole units are ice masses or ice masses at time slots when an ice mass was tested. Extraneous variables include size of the ice mass, amount of salt put on the ice mass, environmental conditions when ice mass tested.

b. Split units are the time slots or occasions when the ice masses are measured. Extraneous variables include micro environment, measurement error, when amount of melted water measured.

c. Whole unit factor is type of salt since types of salt are are assigned to the ice masses at time slots. Split unit factor is time variable, number of minutes after ice mass treated with type of salt.

d. The whole units are blocked by Day, with randomization of types of salt to ice masses occurring independently from day to day.

e. Days are blocks, groupings (Type A) of ice masses. Ice masses are blocks since each ice mass is reused over time (Type B[ii]).

# Bibliography

[1] Agresti, A., C. Franklin. 2007 *Statistics: The Art and Science of Learning from Data* Pearson.

[2] Casella, George. 2008 *Statistical Design* Springer.

[3] Chance Magazine, 2000.

[4] G. Cobb. 1997. *Introduction to Design and Analysis of Experiments.* Springer

[5] Cochran, W.G., G.M. Cox. 1950 *Experimental Designs* Wiley.

[6] Dean, A., D. Voss. 1999. *Design and Analysis of Experiments* Springer

[7] DeVeaux, R., P. Velleman, D. Bock 2005. *Stats Data and Models* Pearson

[8] Peck, R., C. Olsen, J. Devore. 2005. *Introduction to Statistics and Data Analysis* Second Edition.

[9] Dowdy, S., S. Wearden. 1991. *Statistics for Research* Second Edition.

[10] Hicks, C.R. 1993. *Fundamental Concepts in the Design of Experiments* Fourth Edition

[11] Devore, J., Peck, R. 2001. *Statistics The Exploration and Analysis of Data* Fourth Edition

[12] Peck, R., J. Devore. 2008. *Statistics The Exploration and Analysis of Data* Sixth Edition.

[13] Johnson, R., Sui

[14] Kuehl, R.O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*, Second Edition

[15] Kutner, M.H., C.J. Nachtsheim, J. Neter, W. Li. 2005. *Applied Linear Statistical Models*, Fifth Edition.

[16] Lawson, J. 2010. *Design and Analysis of Experiments with SAS*

[17] Littel, R.C., W.W. Stroup, R.J. Freund. 2002. *SAS for Linear Models*, Fourth Edition.

[18] McClave, J.T., T. Sincich. 2003. *A First Course in Statistics*, Eighth Edition.

[19] McClave, J.T., T. Sincich. 2003. *Statistics*, Ninth Edition.

[20] Milliken, G.A., D.E. Johnson. 1992. *Analysis of Messy Data, Volume I: Designed Experiments*

[21] Milliken, G.A., D.E. Johnson. 2009. *Analysis of Messy Data, Volume I: Designed Experiments, 2nd Edition.*

[22] Moore, D., G. McCabe. 2003. *Introduction to the Practice of Statistics.* Freeman.

[23] Morris, T.R. 1999. *Experimental Design and Analysis in Animal Sciences.* CABI Publishing.

[24] Oehlert, G.W. 2000. *A First Course in Design and Analysis of Experiments.* Freeman

[25] Quinn, G.P., Keough, M.J. 2002. *Experimental Design and Data Analysis for Biologists.* Cambridge University Press.

[26] Saliva, A.A. 1990. *Introduction to Statistics.* Saunders College Publishing.

[27] Schabenberger, O., Pierce, F. 2002. *Contemporary Statistical Models for the Plant and Soil Sciences.* Taylor and Francis.

[28] Walpole, R.E., R.H. Myers. 1985. *Probability and Statistics for Engineers and Scientists.* MacMillan.

[29] Weber, D., J. Skillings. *A First Course in the Design of Experiments. A Linear Models Approach.* CRC Press

# Index of Subjects

# Index of Studies