The Basics of Statistical Design and Analysis of Experiments

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

April 18, 2010

Contents

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 conditions of some kind are compared for groups of individuals or objects 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. 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 periods. Interest still lies in comparing the groups defined by the conditions in terms of the response variable.

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/assigned by the researcher. The different diets, low, medium, and high calorie are called 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 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 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 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 suffers 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 above 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.

1.3 Examples of Observational Studies

In this section examples of observational studies and surveys 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. I record 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 they 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.

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. In this book we will be mainly concerned with studies where the response variable is quantitative, such as longevity of a mouse or number of fatty deposits in the arteries.

The treatments in an experiment are values of a variable called the **factor** of the study. In the longevity study the factor of interest is diet. There are three values or levels of diet: low, medium, and high. 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.

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.

1.5 What's affecting the response variable?

Extraneous variables in an experiment are important to recognize and control since differences in 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 on the amount of tomatoes grown. Suppose that ten plants of about the same size and variety are used. Five plants received fertilizer A and five receive fertilizer B. The plants are assigned at random to ten plots in a garden. The resulting yields in pounds are given below:

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

Note that the fertilizer B yields appear to be slightly larger than those for A? The mean yields for fertilizer A and B, respectively, are 48.2 and 51.4. Can we say conclusively that fertilizer B is better? Note, however, that WITHIN each treatment or fertilizer group 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. How can we tell?

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** 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 association between the extraneous variable and the response variable. Then it's impossible to tell whether the 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 variables** 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 got 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

The following are 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.

- Randomization

Randomization should be used to assign treatments to experimental units. This can be done in various ways to be described in subsequent chapters. 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. For example, 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, randomization may not produce balanced groups and blocking should be used to achieve better balance.

Sometimes an experiment has to be performed over time. 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 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.

- Blocking

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

  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. In theory, randomization will, on average over many replications of the experiment, balance out the effect of extraneous factors. 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 factors not fully balanced by the randomization, and not necessarily due to microwaving.

  An alternative way of designing the experiment is as follows. Before any kind of randomization, sort the twenty oranges by size 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 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 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 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.

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 in a completely randomized fashion. 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 in each row and compare the treatments 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.

Sometimes blocks are natural sortings or groupings of the experimental units. For example, several sets of twins may be used in a study to compare the effects of two drugs. A block is a set of twins. The individuals of the twin set 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 sets of twins.

- Direct Control

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

- Replication

  Replication of a treatment refers to a series of independent assignments of experimental units to that treatment. There would not be replication in the orange example if only one orange is used for each of the microwaving/ no microwaving treatments.

  Because of extraneous variables and their effects on the response variable, replication is obviously important. A difference in orange juice with only one replication could be the result of a difference in size of the two oranges. Only with replication is it possible to conclude that there are "true" differences in amount of juice between microwaving and not microwaving.

  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 study, that is sampled many males and many females, the "true" pattern would be concluded.

  Recall from an earlier discussion that the differences in the average value of a response variable among 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 is necessary in order to measure the extent of the differences that could be due to the effects of extraneous variables.

1.9 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 a 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 observation 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. [2].

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 have 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. [2]

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 ([5], 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.

Chapter 2

Basic Concepts and The One Sample Problem

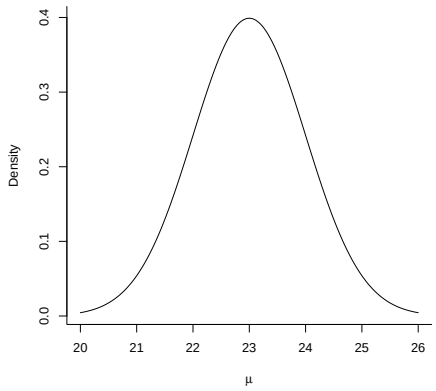
This chapter reviews the basic statistical concepts associated with inferences about a single population based on a random sample selected from that population. The notions of estimation of population parameters, standard errors of estimators, and hypothesis testing are discussed.

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 $\mu = E[y]$ and variance σ^2. The mean is also referred to as the expected value of y, denoted by $E[y]$. The variance is defined as $\sigma^2 = E[(y-\mu)^2]$, the expected value of the square of the difference between y and μ. The standard deviation of y is defined to be $\sigma = \sqrt{\sigma^2}$. The mean μ, variance σ^2, and standard deviation σ 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 μ. Figure 2.1 provides a histogram of a basic normal population with mean of y equal to $\mu = 23$ 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 μ and standard deviation σ, 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 μ and standard deviation σ 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 μ and standard deviation σ. Statistically $y_1, y_2, ..., y_n$ represent random

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

values which are independent, identically distributed as the population, each with mean μ and standard deviation σ.

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

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 standard deviation* $\sigma = 15$ *pounds. In practice neither the population mean nor the population standard deviation would typically be known. Suppose that a random sample of* $n = 10$ *students is selected. Table 2.1 gives the 10 weights and their deviations from the mean, 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 standard deviation is $s = \sqrt{3279.1/9} = 19.1$. Notice that the sample mean and standard deviation for this sample are not the same as the mean and 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 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 $\overline{y}$, denoted by $\mu_{\overline{y}} = E[\overline{y}] = \mu$. Note this says that the average of sample means under repeated sampling is equal to μ. 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 μ.

2. The standard deviation of the random variable $\overline{y}$ is $\sigma_{\overline{y}} = \sigma/\sqrt{n}$. The quantity $\sigma_{\overline{y}}$ gives a crude measure of how far "off" an observed sample mean is away from the unknown population mean μ. Note that this number is not very useful in practice, however, since σ is unknown. However we could estimate σ with s, the sample standard deviation. See below.

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

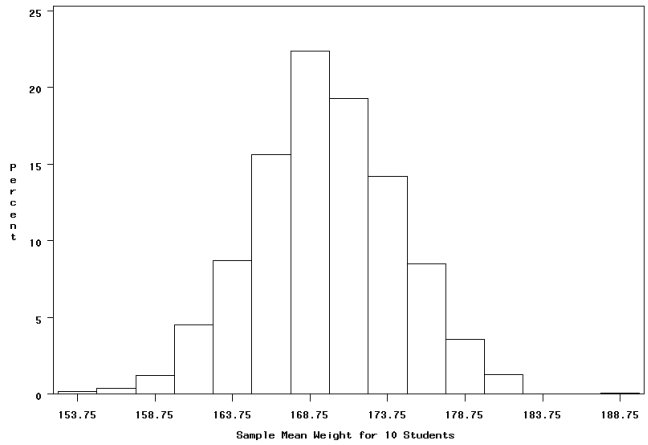
Example 2.2 *A SAS program was written to simulate the random sampling of 1000 samples of male university students, each of size* $n = 10$ *from the same population used in Example 2.1. For each sample of 10 weights obtained the sample mean was calculated. Figure 2.2 gives a histogram for the 1000 sample 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 μ. The sample will also provide an observed sample standard deviation, s. This observed sample standard deviation is used to estimate σ, 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.

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,

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

Figure 2.2: Histogram of 1000 Sample Mean Weights



has a standard normal 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

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

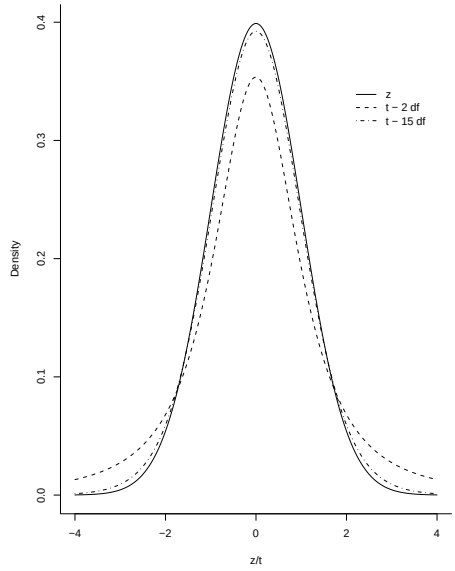
has a standard normal distribution.

If in the standardized sample mean we replace σ 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 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 μ is unknown and must be estimated based on a sample. A confidence interval estimate for a population mean is an interval of possible values for μ. Associated with the interval is a "confidence level" that indicates how confident we are that the interval actually contains μ. The one sample "t" interval for a population mean is based on the Student's t distribution or simply the t distribution.

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



Fact 2.2 *Suppose a random sample $y_1, y_2, ..., y_n$ of size n is selected from a normal population with mean μ and standard deviation σ. 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** $\nu = n - 1$.

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 ν. They 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 depends upon the parameter ν which will generally depend on sample size. Figure 2.3 gives a picture of two t distribution compared to the standard normal distribution.

Table A.2 gives the upper α probability points, denoted by $t_{\alpha;\nu}$, for certain values of α and ν. The area under the t distribution to the right of $t_{\alpha;\nu}$ is α.

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 μ and population standard deviation σ. 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 μ. A 95% *confidence interval* for μ is an interval of possible values for μ. The 95% "confidence level" refers to how confident we are that the value of μ takes on one of the values in the interval. A brief derivation of such an interval follows.

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 μ. 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 μ.

Example 2.3 *Suppose that amount of money spent by students on textbooks in a given semester are normally distributed with some (unknown) mean μ and standard deviation σ. 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 μ 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 $2.145(\$7.02) = \$15.06$. 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 confidence interval 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 $n-1$ 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 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) is 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 average survival time?

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 even if the true population mean is no different, that is 60 months, simply due to sampling variability and the fact that survival times will vary naturally. That is, is the result due solely to chance (sampling) or is the new treatment really better?

Let μ 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:

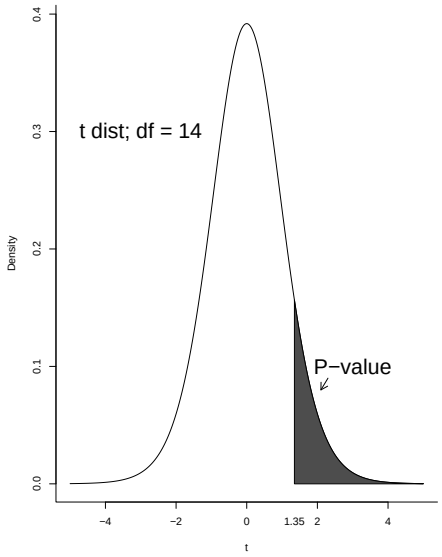
- Assume initially that H_o is true
- Assuming H_o is true, calculate a summary of the data called the *test statistic*. The probability distribution of the test statistic is known.
- Calculate the probability of obtaining a value of the test statistic like the observed value or more extreme in the direction of the alternative hypothesis. 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 α. A common value used for α 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 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 $\nu = 15 - 1 = 14$. Now the *observed value* of the test statistic in this example is

Figure 2.4: P-value for Example 2.4



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

That is, the observed sample mean of 62.7, is 1.36 standard errors above the null hypothesized value of 60. Thus the P-value is

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

which is the area of the shaded region 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 $P-value = 0.1010$. Using a significance level of $\alpha = 0.05$, since $P-value > 0.05$ 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.

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

Table 2.2: General Forms of Null and Alternative Hypothesis for Single Sample t test

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

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.

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" 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 or 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.

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 is 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 α level. Typical values used here are $\alpha = 0.01$ or $\alpha = 0.05$. Symbolically the null hypothesis is rejected is $P - value \leq \alpha$. 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 α 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 α, 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 α level be perhaps set at 0.01 rather than 0.05. Note that regardless of sample size setting the α 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, β, 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 bf 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 β 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 β or power depending upon various characteristics of the study. For example in the one sample t test β, or power depends upon:

1. Sample size. For a particular population standard deviation and particular α level, increasing sample size will decrease β.

2. The α level. For a particular population standard deviation and fixed sample size, increasing α decreases β and decreasing α increases β.

3. Population standard deviation. For fixed sample size and α level, β will be larger for populations with larger standard deviations.

4. The difference between the null hypothesis value of μ and the true value of μ under the alternative hypothesis. In our example the null hypothesis value of μ was 60 months. If the new drug is more effective and the true mean is 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 70 months then the difference or effect of the drug is an increase of 10 months. The β level will depend upon the difference or "effect" in this example. The greater the effect the smaller the level of β 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

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

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 μ_o and μ_a are null and alternative values of μ, respectively.

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 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?

2.2 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 sampling distribution of the sample mean $\overline{y}$?

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

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

2.3 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 $95th$ percentile) of the sampling distribution of $\frac{\overline{y}-100}{s/\sqrt{25}}$

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

2.4 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.

2.5 This example is taken from Devore and Peck ([7], page 555). Calorie contents for each 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? 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.

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 two treatment groups in an experiment or the two observed groups in an observational study. We will examine the case where the two samples are independent and where they are dependent. Section 3.1 will examine hypothesis testing and confidence interval estimation in the independent two 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 data have been gathered by one of two ways:

- 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 study 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 μ_1 and variance σ_1^2. Let $\overline{y}_{1\cdot}$ and s_1 represent the sample mean and standard deviation of the sample. Then from Chapter 2, $E[\overline{y}_{1\cdot}]$ $= \mu_1$ and standard error of $\overline{y}_{1\cdot}$ is $\sigma_{\overline{y}_{1\cdot}} = \sigma_1/\sqrt{n_1}$. Similarly, let $y_{21}, y_{22}, ..., y_{2n_1}$

represent the values of a random sample of size n_2 from another normal population with mean μ_2 and variance σ_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 standard error 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, μ_1 and μ_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 error of this estimator. We also need to know the 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 possible 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 error" 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 error gives the average distance differences in sample means are from the difference in population means $\mu_1 - \mu_2$. In practice after sampling the difference in sample means is known. The standard error then gives a rough idea of how far off that observed difference is from the unknown difference in population means. However the population variances σ_1^2 and σ_2^2 are unknown. So in order to be of practical value these two population variances need to be estimated to derive an estimated standard error. We will first consider hypothesis testing and confidence interval estimation that estimates the two population variances 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 error in the denominator to obtain an estimated standard error, $s_{\overline{y}_{1\cdot}-\overline{y}_{2\cdot}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. 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, ν, used in this approximate 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

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2;\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}} \tag{3.2}$$

where $t_{\alpha/2;\nu}$ is the upper $\alpha/2$ probability point from a t distribution with degrees of freedom, ν, 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 $\overline{y}_1 - \overline{y}_2$ 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 estimated standard error of $\overline{y}_1 - \overline{y}_2$. 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 μ_1 and μ_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 null hypothesized value of 0 for the

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

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

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 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.

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

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 ν 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: Plot of Weight Gain (lbs) versus Treatment

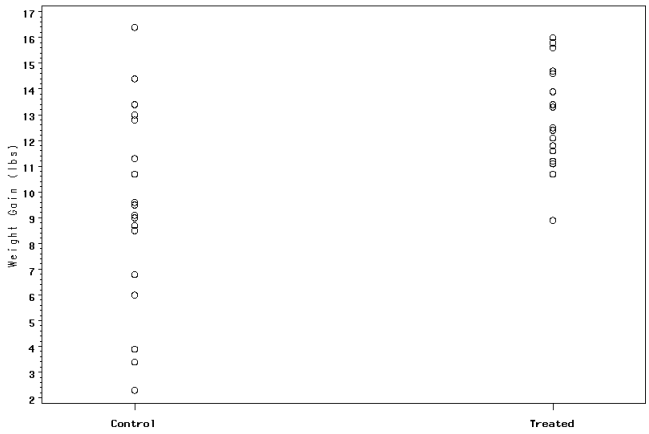


Figure 3.1 gives a plot of the weight gains versus treatment. Weight gain does appear to be improved in the treated group. There also appears to be less variability in weight gains in the treated group.

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 $\overline{y}_1 = 12.83$ and $\overline{y}_2 = 9.38$. The standard deviation of weight gains for the treated and control groups are, respectively, $s_1 = 1.87$ and $s_2 = 3.88$. Let μ_1 and μ_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

$$t = \frac{\overline{y}_{1\cdot} - \overline{y}_{2\cdot}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{12.83 - 9.38}{\sqrt{\frac{1.82^2}{22} + \frac{3.88^2}{18}}} = \frac{3.48}{0.99} = 3.46$$

Degrees of freedom would be

$$\nu = \frac{\left(\frac{1.87^2}{23} + \frac{3.88^2}{18}\right)^2}{\frac{1}{22-1}\left(\frac{1.82^2}{22}\right)^2 + \frac{1}{18-1}\left(\frac{3.88^2}{18}\right)^2} = 23.4$$

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 anywhere from 1.4 to 5.5 pounds.

3.1.3 Two Sample Pooled t test and Confidence Interval

In some experimental situations 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 true standard error of $\overline{y}_1 - \overline{y}_2$ is then $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$.

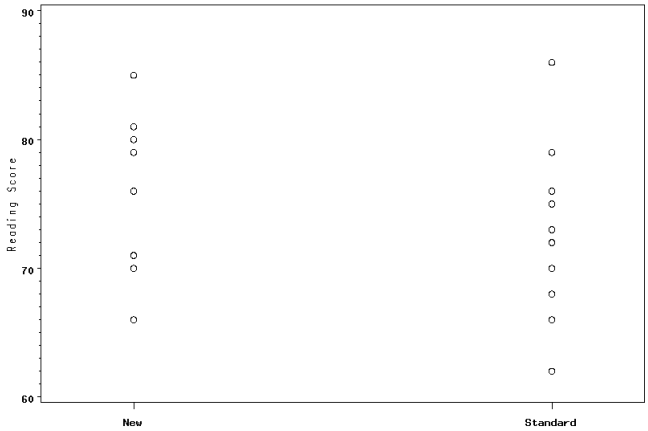
Example 3.2 *An example from McClave and Sincich [13], page 329, will illustrate. A new method of teaching reading to children who are slow learners is compared to a current standard method. The comparison is based on a reading test score given at the end of the learning period. Ten subjects are taught 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 and n_2 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 μ_1 and μ_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.

Figure 3.2: Plot of Reading Score versus Method



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

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

Note that the numerator is just the sum of squared deviations of the observations 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, σ^2 with s_p^2 we have the estimated standard error of $\overline{y}_{1\cdot} - \overline{y}_{2\cdot}$, $s_{\overline{y}_{1\cdot} - \overline{y}_{2\cdot}}$ to be

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

or

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

This is the standard error used in practice. In order to construct confidence intervals and perform hypothesis testing we need to have a sampling distribution to calculate p-values and to obtain percentiles 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 true standard error $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$, assuming equal population variances. Thus $\frac{(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ has the standard normal distribution and P-values and error

margins could be based on this distribution. However, again, the true standard error is unknown. If we replace the true standard error with the estimated 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 $\nu = n_1 + n_2 - 2$.

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

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

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 (estimated) standard error of $\overline{y}_1 - \overline{y}_2$ 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 value of the test statistic under the null hypothesis is thus

$$\frac{(\overline{y}_1 - \overline{y}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(72.23 - 68.30) - (0)}{2.62} = 1.55$$

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 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 [3] 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 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 [6]; Peck and Devore [8]).

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

An alternative approach to handling the unequal variance 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 greater discussion on transformation will be given in Chapter 8 on the checking of ANOVA assumptions.

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 [3]) 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. There may be two treatments of interest. Each individual receives one treatment on one occasion and the other treatment at 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. The individual is the block or the pair of time slots/occasions corresponding to each individual.
- **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 liquid or piece of cloth, is physically split. 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.

Let the n random pairs of observations 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 μ_1 and variance σ_1^2. The $y_2's$ are a random sample from a population with mean μ_2 and variance σ_2^2. The degree of relationship between any two y's in a pair is quantified by the **covariance**, σ_{12} and **correlation**, ρ. Readers may remember covariance and correlation from their elementary 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 analysis for paired data is based on the differences

$$d_1 = y_{11} - y_{21}$$

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

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 and standard deviation of the $d's$, denoted by $\overline{d}$ and s_d, respectively. 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 μ_d and standard deviation σ_d then we can use the one sample t confidence interval and t test from Chapter 2 to draw inferences about μ_d and thus conclusions about the differences in treatments or conditions.

The general form of the confidence interval for μ_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 μ_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 δ 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 μ_d is

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

where δ is a null hypothesized value for μ_d. P-values are determined by the t distribution with $\nu = n - 1$, 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, σ_1^2 and σ_2^2, of y in the populations are equal.

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 μ_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 based on $60 - 1 = 59$ 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 α 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 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 Pooled Two Sample t test

In this section we will give some power calculations for the one-sided two independent 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 σ, i.e.

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

Table 3.2: Words Lists for Student Experiment

List A	List B
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

Table 3.4 gives power for various values of E and common sample size n.

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 two sample 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 the study 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 with whom 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, pheronomnal 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.

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 scatterplot 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 $t = 3$, $n_1 = 4$, $n_2 = 4$, 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 $N = 12$. For each treatment let y_{ij} denote the j^{th} observation on the 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 factors such as genetics, stress factors, etc. The hypothesis test will be based on a decomposition of the data into parts.

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 $\overline{y}_{..} = 21.50$ of all 12 lifespans is called the *grand mean.*

Figure 4.1: Plot of Lifelength versus Diet

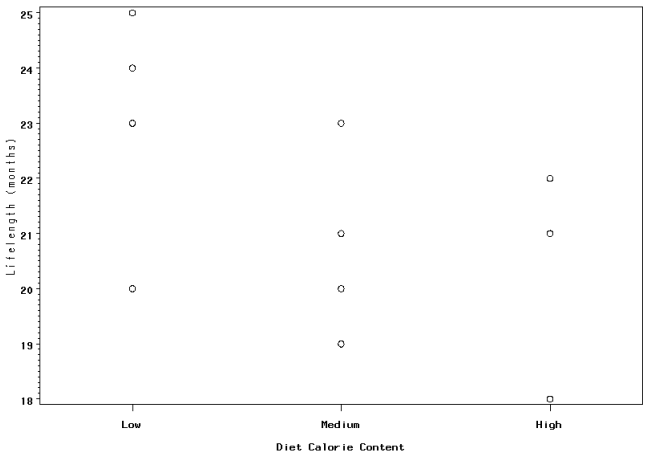


Table 4.1: Descriptives for Mice Data

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

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

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 is measured by the difference between lifelength and the mean for Diet 3, 23.00. The errors for all 4 mice receiving Diet 3 are given in Table 4.2.

In your first course in statistics these were probably called deviations from the mean. 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 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 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 lifelength of an animal the mean of the group to which the animal belongs. The errors for groups 1 and 2 are given below.

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.

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

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

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.

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:

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

Thus diet 3 has the "effect" of raising lifelength by 1.50 months compared to the grand mean of 21.50 months. Diet 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:

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

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

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

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

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

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

given in Section 4.1 is referred to as the **sample means model**. The set of equations that uses the effects,

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

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

The **population means model** for the lifelength data is:

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

and the **population effects model** is

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

In the above equations,

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

An assumption of the population model is that the $\epsilon_{ij}{}'s$ are independent normal random variables each with mean 0 and unknown variance σ^2.

It is important to realize that $\overline{\mu}_{\cdot}, \mu_i, \alpha_i, \epsilon_{ij}$ are NOT the same as $\overline{y}_{\cdot\cdot}, \overline{y}_{i\cdot}, A_i$ and e_{ij}. 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,

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

Substituting actual values we have

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

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 α_3; $e_{31} = 0$ is an estimate of ϵ_{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 $\overline{y}_{i\cdot}$ 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 $\overline{y}_{i\cdot}$ even if the μ_i are the same or obtain sample effects A_i different from 0 even if the true effects α_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 ϵ'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

$$\overline{y}_{1\cdot} = \mu_1 + \overline{\epsilon}_{1\cdot} \tag{4.4}$$
$$\overline{y}_{2\cdot} = \mu_2 + \overline{\epsilon}_{2\cdot} \tag{4.5}$$

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 observation in the decomposition table, and since there are repeats,

$$\begin{aligned} SSTR &= n_1 A_1^2 + n_2 A_2^2 + n_3 A_3^2 \\ &= 4(-0.75)^2 + 4(-0.75)^2 + 4(1.50)^2 \\ &= 2.25 + 2.25 + 9.00 \\ &= 13.50 \end{aligned} \tag{4.6}$$

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

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

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

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

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 [10], 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 σ^2 is the common variance of the errors $\epsilon_{ij}{}'s$.

Thus on average MSE is equal to the error variance σ^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 α_i are zero) then the expected value or average of $MSTR$ is greater than the expected value or average of MSE, equal to σ^2. If diet does not have an effect (α_i are all zero), then the expected values of $MSTR$ and MSE are both equal to σ^2, in which case the observed values should be similar.

In the lifelengths example, the estimate of the error variance σ^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 sampling distribution of the ratio 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

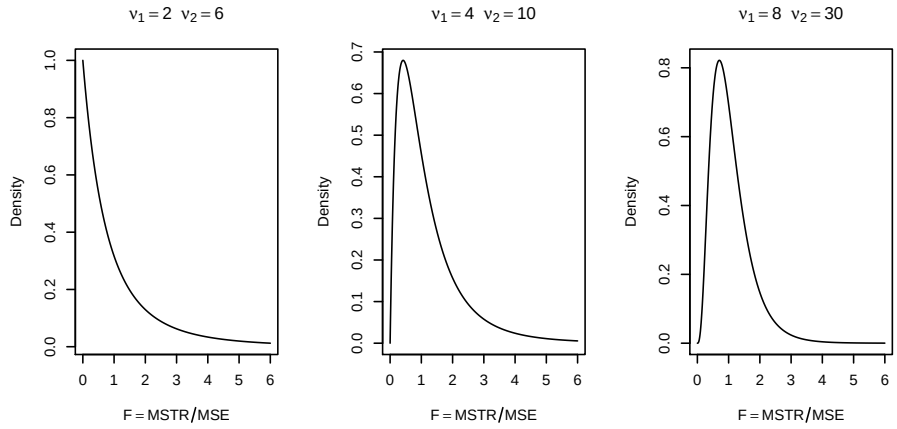
In this section we describe the sampling distribution of the ratio $\frac{MSTR}{MSE}$.

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, ν_1 and denominator degrees of freedom, ν_2.
- The F distribution represents the probability distribution of a statistic which is non-negative, such as $MSTR/MSE$.

Figure 4.2: Examples of F distributions



- 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 α probability points, denoted 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 ν_1 and ν_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.

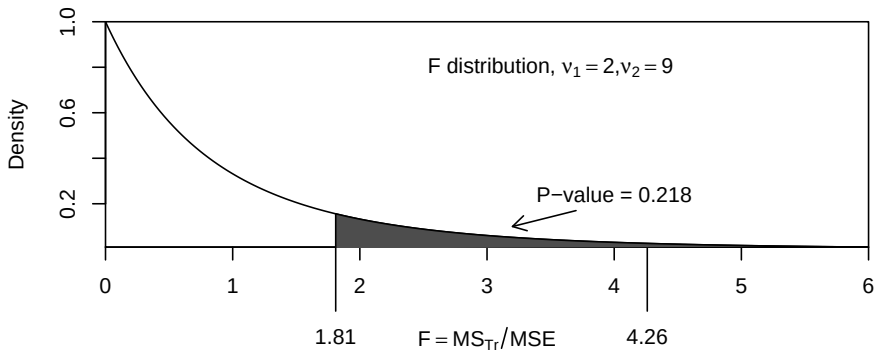
4.4.3 Summary of the F test for Treatment Effects

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,

Figure 4.3: P-value for Diet Example



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

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

$$SSE = \sum_{i=1}^{t} \sum_{j=1}^{n_i} [e_{ij}]^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} [y_{ij} - \overline{y}_{i\cdot}]^2$$

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 $\nu_1 = (t-1)$ numerator degrees of freedom and $\nu_2 = (N-t)$ denominator degrees of freedom.

At a significance level of α 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

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

α 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 α 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 denoted by $SSTOT$:

$$\begin{aligned} SSTOT &= (22)^2 + (18)^2 + (21)^2 + (22)^2 \qquad Diet1 \\ &\quad +(20)^2 + (19)^2 + (23)^2 + (21)^2 \qquad Diet2 \\ &\quad +(23)^2 + (24)^2 + (20)^2 + (25)^2 \qquad Diet3 \\ &= 1733 + 1731 + 2130 \\ &= 5594 \end{aligned} \tag{4.9}$$

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

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			

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	$SSTOT$			

$$SSTOT = SSGM + SSTR + SSE$$

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. Or we are analyzing the variation in lifelengths and 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 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

Table 4.8: Decomposition of Lifelength Corrected for 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

The grand mean of 21.50 months was subtracted from each lifelength (this is called lifelength corrected for the 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 is part diet effect of -0.75 and part error of 1.25.

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.

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

4.6 Independent Samples t Test Revisited

Consider the two sided independent samples t test introduced in Chapter 3 that assumed equal population variances. This special testing situation can be shown to be equivalent to the 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

Table 4.9: ANOVA Table for Lifelength Data: Correction for 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 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)	$N-1$	$SSTOT_C$			

μ_1. Let $y_{21}, y_{22}, ..., y_{2,n_2}$ be independent measuresment from a population with mean μ_2. Let σ^2 be the common population variance.

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

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

where the ϵ_{ij} are independent, normally distributed random variables with mean of 0 and variance of σ^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 population variance, $s_p^2 = 33.25$. 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 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 mean and one with the correction for the mean).

b. Construct the ANOVA table (not corrected for mean and corrected for 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 [11]. 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. Obatin 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.

Chapter 5

Multiple Comparisons

5.1 Introduction

If it has been concluded from the F test 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 compares 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 method 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

μ_i and μ_j are to be compared with i and j referring to two of the possible t means. Then for an α 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 $\frac{\alpha}{2}$ percentile 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 σ. 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, ν, 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 μ_i and μ_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 data 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)	$\overline{y}_{1\cdot} = 51.00$	$s_1 = 1.81$
Cover 2 (Child)	$\overline{y}_{2\cdot} = 46.90$	$s_2 = 1.72$
Cover 3 (Cereal Bowl)	$\overline{y}_{3\cdot} = 48.42$	$s_3 = 0.71$

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 ANOVA 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

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

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:

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

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 that the mean sales with the sports hero as the box cover is between 2.25 (x 10,000) and 5.95 (x 10,000) boxes higher than when the cover has a child. It is estimated that the mean sales with the sports hero as the box cover is between 0.73 (x 10,000) and 4.43 (x 10,000) boxes higher than when the cover has a bowl of cereal.

5.2.2 Contrasts - generalization of pairwise difference

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 differences but in certain pre-planned comparisons of subgroups of the means.

Let $\mu_1, ..., \mu_t$ be the 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 difference 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 $c's$ is 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 μ_1 and μ_2 with the average of the μ_3 and μ_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 in Equation 5.4 is

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

This estimate is normally distributed with mean

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

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

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

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:

$$H_0 : C = 0 \tag{5.13}$$
$$H_a : C \neq 0 \tag{5.14}$$

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 α level of significance if $|t| \geq t_{\alpha/2;N-t}$. 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 five 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 5.3: 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 5.4 provides the means and standard deviations of the satisfaction data for the five methods of instruction.

The ANOVA is given in Table 5.5. The differences in the sample mean class satisfaction scores are significantly different 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:

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

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			

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:

$$\begin{aligned}
s^2\{\hat{C}_1\} &= (1.44)(\frac{(1/2)^2}{6} + \frac{(1/2)^2}{6} + \frac{(-1/3)^2}{6} + \frac{(-1/3)^2}{6} + \frac{(-1/3)^2}{6}) \\
&= 0.20 \\
s^2\{\hat{C}_2\} &= (1.44)(\frac{(1)^2}{6} + \frac{(-1)^2}{6} + \frac{(0)^2}{6} + \frac{(0)^2}{6} + \frac{(0)^2}{6}) \\
&= 0.48 \\
s^2\{\hat{C}_3\} &= (1.44)(\frac{(0)^2}{6} + \frac{(0)^2}{6} + \frac{(1/2)^2}{6} + \frac{(1/2)^2}{6} + \frac{(-1)^2}{6}) \\
&= 0.36 \\
s^2\{\hat{C}_4\} &= (1.44)(\frac{(0)^2}{6} + \frac{(0)^2}{6} + \frac{(1)^2}{6} + \frac{(-1)^2}{6} + \frac{(0)^2}{6}) \\
&= 0.48
\end{aligned}$$

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 ± 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 comparisons 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 ± 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 ± 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 ± 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 α represents the pre-assigned probability of making a Type I error for a particular test, called the significance level of the test. $1 - \alpha$ is the preassigned confidence level associated with each confidence interval. Of interest in multiple testing is the overall or **experimentwise significance level**, denoted by α_e and the overall or **experimentwise confidence level** denoted by CL_e.

The experimentwise wise significance level α_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 α level, then

$$\alpha_e \leq m\alpha$$

In this context α is called the **comparison wise Type I error rate**. Thus if each of $m = 6$ tests is conducted at the $\alpha = 0.05$ 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 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, α_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 α is the comparison-wise significance level used for each test. So suppose we want the experimentwise error rate to be at most α, rather than $m\alpha$. Then clearly if we carry out each test at comparison level of $\frac{\alpha}{m}$ rather than α 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 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 α level for m tests use $\frac{\alpha}{m}$ for the comparison wise error rate.

5.4.1 Set of m Contrasts

For a set of 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

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

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 overall 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 percentile 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 percentiles 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, α_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, α_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 percentile 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 overall confidence level of at least 95% are

$$\begin{array}{rcl} (-1.33) & \pm & 2.69(\sqrt{0.20}) \\ (0.03) & \pm & 2.69(\sqrt{0.48}) \\ (0.72) & \pm & 2.69(\sqrt{0.36}) \\ (-1.35) & \pm & 2.69(\sqrt{0.48}) \end{array}$$

or

$$\begin{array}{rcl} -1.33 & \pm & 1.20 \\ 0.03 & \pm & 1.86 \\ 0.72 & \pm & 1.61 \\ -1.35 & \pm & 1.86 \end{array}$$

Thus the four Bonferroni intervals with overall confidence level of at least 95% are

$$\begin{array}{rcccl} -2.53 & \leq & C_1 & \leq & -0.13 \\ -1.83 & \leq & C_2 & \leq & 1.89 \\ -0.89 & \leq & C_3 & \leq & 2.33 \\ -3.21 & \leq & C_4 & \leq & 0.51 \end{array}$$

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 α 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

$$\overline{y}_i - \overline{y}_j \pm (t_{\frac{\alpha}{2m};\nu})\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Determination of the appropriate t percentile 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 has 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:*

$$\begin{aligned}
(51.00 - 46.90) &\pm 2.69\sqrt{2.26}\sqrt{\tfrac{1}{6} + \tfrac{1}{6}} \\
(51.00 - 48.42) &\pm 2.69\sqrt{2.26}\sqrt{\tfrac{1}{6} + \tfrac{1}{6}} \\
(46.90 - 48.42) &\pm 2.69\sqrt{2.26}\sqrt{\tfrac{1}{6} + \tfrac{1}{6}}
\end{aligned}$$

or

$$\begin{aligned}
4.10 &\pm 2.33 \\
2.58 &\pm 2.33 \\
-1.52 &\pm 2.33
\end{aligned}$$

Thus the three intervals are:

$$\begin{aligned}
1.77 &\leq \mu_1 - \mu_2 \leq 6.43 \\
0.25 &\leq \mu_1 - \mu_3 \leq 4.91 \\
-3.85 &\leq \mu_2 - \mu_3 \leq 0.81
\end{aligned}$$

Note that the t percentile used for the Bonferroni intervals, 2.69, is larger than the t percentile 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 α probability point from the Studentized Range Distribution, tabulated in Tables A.5 ($\alpha = 0.05$) and A.6 ($\alpha = 0.01$). The tables depend upon a degrees of freedom parameter, ν, 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 α, 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 overall 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*

$$\begin{aligned}(51.00 - 46.90) &\pm \tfrac{3.67}{\sqrt{2}}\sqrt{2.26}\sqrt{\tfrac{1}{6}+\tfrac{1}{6}}\\ (51.00 - 48.42) &\pm \tfrac{3.67}{\sqrt{2}}\sqrt{2.26}\sqrt{\tfrac{1}{6}+\tfrac{1}{6}}\\ (46.90 - 48.42) &\pm \tfrac{3.67}{\sqrt{2}}\sqrt{2.26}\sqrt{\tfrac{1}{6}+\tfrac{1}{6}}\end{aligned}$$

or

$$\begin{aligned}4.10 &\pm 2.25\\ 2.58 &\pm 2.25\\ -1.52 &\pm 2.25\end{aligned}$$

Thus the three intervals are:

$$\begin{aligned}1.85 &\leq \mu_1 - \mu_2 \leq 6.35\\ 0.33 &\leq \mu_1 - \mu_3 \leq 4.83\\ -3.77 &\leq \mu_2 - \mu_3 \leq 0.73\end{aligned}$$

Notice that the multiplier, $\frac{3.67}{\sqrt{2}} = 2.60$, on 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. Multiple 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 the family wise significance level. The Bonferroni procedure controls the family wise 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

Table 5.6: Multipliers for Multiple Comparison Procedures

			Multiplier		
t	m	ν	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

do not have to be of equal size. The Tukey-Kramer procedure is used to control the family wise 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 usually called the Tukey procedure.

3 Suppose that the family 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

$$estimate \pm (multiplier) \times SE(estimate)$$

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 family of all pairwise comparisons when there are $t = 3$ treatments ($m = 3$ comparisons) comparisons and when there are $t = 4$ treatments ($m = 6$ comparisons) for certain degrees of freedom ν.

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 5.7: 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

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 1 error rate and statistical power. The Tukey-Kramer procedure is often used because it offers better protection against the type 1 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 along with P-values are given in Table 5.7 for the cereal data.

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 lower bound on the the family wise confidence level?

c. Suppose that the Bonferroni procedure is to be used with the family wise confidence level required to be at least than 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 family wise confidence level specified to be exactly 0.99. What is the appropriate percentile 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 a family wise 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 Multiple t procedure and the Tukey-Kramer procedure give the same multiplier on the standard error and thus the same confidence 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 are combinations of the levels of A and B. 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 μ_{ij} is the mean production for treatment A_{ij}.

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 ϵ_{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 water regimens and the true main effects of watering regimens are defined similarly.

Note that each treatment mean can be written as the sum of the grand mean + effect of fertilizer + 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

$$y_{11k} = \mu_{11} + \epsilon_{11k} = \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + \epsilon_{11k}$$

$$y_{12k} = \mu_{12} + \epsilon_{12k} = \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + \epsilon_{12k}$$

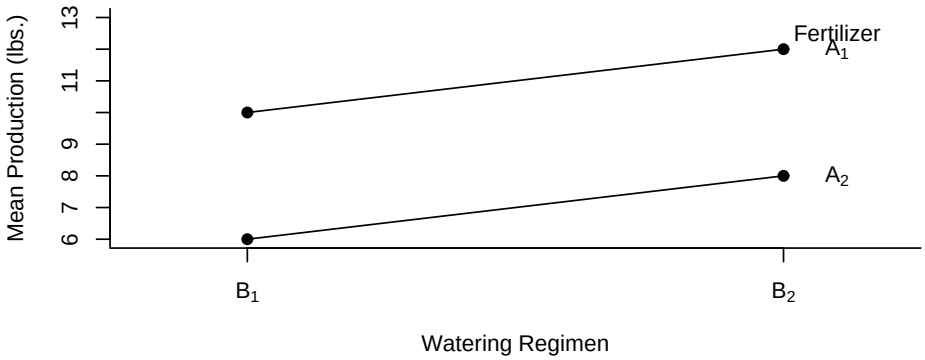
$$y_{21k} = \mu_{21} + \epsilon_{21k} = \mu_{\cdot\cdot} + \alpha_2 + \beta_1 + \epsilon_{21k}$$

$$y_{22k} = \mu_{22} + \epsilon_{22k} = \mu_{\cdot\cdot} + \alpha_2 + \beta_2 + \epsilon_{22k}$$

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 treatment effect, and factor B treatment 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 effect of water regimen does not depend on fertilizer or the effect of fertilizer 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 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 where interaction exist shortly.

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.

6.3 Interaction Model

Suppose that in the tomato production example the (true) treatment means were as in Table 6.3.

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	$\mu_{21} = 6$	$\mu_{22} = 8$	$\mu_{2\cdot} = 7$	$\alpha_2 = -3$
	$\mu_{\cdot 1} = 8$	$\mu_{\cdot 2} = 12$	$\mu_{\cdot\cdot} = 10$	
	$\beta_1 = -2$	$\beta_2 = 2$		

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 effect of fertilizer DOES DEPEND upon watering regimen or the effect of watering regimen on production 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 means CANNOT expressed as the sum of the grand mean, fertilizer effect, and watering regimen effect:

$$\begin{array}{lll} \mu_{11} \neq \mu_{\cdot\cdot} + \alpha_1 + \beta_1 & \text{or} & 10 \neq 10 + 3 + (-2) = 11 \\ \mu_{12} \neq \mu_{\cdot\cdot} + \alpha_1 + \beta_2 & \text{or} & 16 \neq 10 + 3 + 2 = 15 \\ \mu_{21} \neq \mu_{\cdot\cdot} + \alpha_2 + \beta_1 & \text{or} & 6 \neq 10 + (-3) + (-2) = 5 \\ \mu_{22} \neq \mu_{\cdot\cdot} + \alpha_2 + \beta_2 & \text{or} & 8 \neq 10 + (-3) + 2 = 9 \end{array}$$

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

$$10 = 10 + 3 + (-2) + (-1)$$

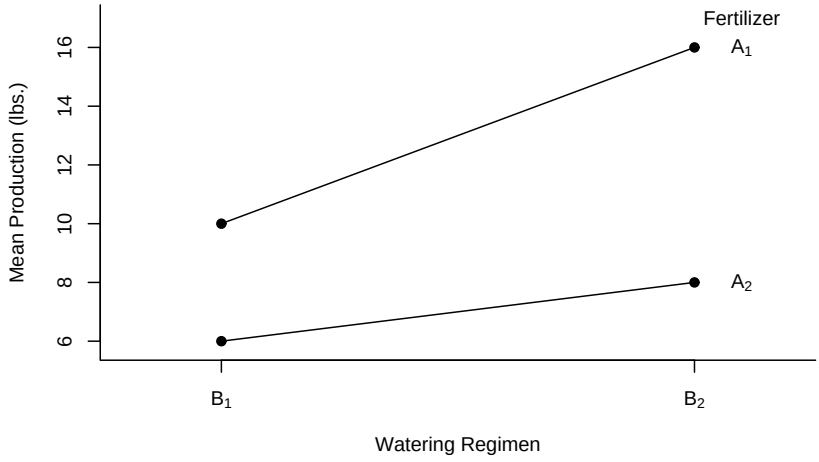
To get (-1), $\mu_{\cdot\cdot} + \alpha_1 + \beta_1$ was subtracted from μ_{11}. Thus the new equation looks like

$$\mu_{11} = \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + [\mu_{11} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_1)] \quad \text{or} \quad 10 = 10 + 3 + (-2) + [-1]$$

The necessary adjustments are illustrated for all equations below:

$$\begin{array}{lll} \mu_{11} = \mu_{\cdot\cdot} + \alpha_1 + \beta_1 + [\mu_{11} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_1)] & \text{or} & 10 = 10 + 3 + (-2) + [-1] \\ \mu_{12} = \mu_{\cdot\cdot} + \alpha_1 + \beta_2 + [\mu_{12} - (\mu_{\cdot\cdot} + \alpha_1 + \beta_2)] & \text{or} & 16 = 10 + 3 + 2 + [1] \\ \mu_{21} = \mu_{\cdot\cdot} + \alpha_2 + \beta_1 + [\mu_{21} - (\mu_{\cdot\cdot} + \alpha_2 + \beta_1)] & \text{or} & 6 = 10 + (-3) + (-2) + [1] \\ \mu_{22} = \mu_{\cdot\cdot} + \alpha_2 + \beta_2 + [\mu_{22} - (\mu_{\cdot\cdot} + \alpha_2 + \beta_2)] & \text{or} & 8 = 10 + (-3) + 2 + [-1] \end{array}$$

Figure 6.2: Tomato Production - Interaction between Watering Regimen and Fertilizer



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.

$$\begin{aligned}
\mu_{11} &= \mu_{..} + \alpha_1 + \beta_1 + [\mu_{11} - (\mu_{..} + \alpha_1 + \beta_1)] \\
\mu_{11} &= \mu_{..} + \alpha_1 + \beta_1 + \alpha\beta_{11}
\end{aligned} \tag{6.5}$$

$$\begin{aligned}
\mu_{12} &= \mu_{..} + \alpha_1 + \beta_2 + [\mu_{12} - (\mu_{..} + \alpha_1 + \beta_2)] \\
\mu_{12} &= \mu_{..} + \alpha_1 + \beta_2 + \alpha\beta_{12}
\end{aligned}$$

$$\begin{aligned}
\mu_{21} &= \mu_{..} + \alpha_2 + \beta_1 + [\mu_{21} - (\mu_{..} + \alpha_2 + \beta_1)] \\
\mu_{21} &= \mu_{..} + \alpha_2 + \beta_1 + \alpha\beta_{21}
\end{aligned}$$

$$\begin{aligned}
\mu_{22} &= \mu_{..} + \alpha_2 + \beta_2 + [\mu_{22} - (\mu_{..} + \alpha_2 + \beta_2)] \\
\mu_{22} &= \mu_{..} + \alpha_2 + \beta_2 + \alpha\beta_{22}
\end{aligned}$$

Thus the "full" model, means and effects, for each of the treatments is:

$$\begin{aligned}
y_{11k} &= \mu_{11} + \epsilon_{11k} \\
&= \mu_{..} + \alpha_1 + \beta_1 + \alpha\beta_{11} + \epsilon_{11k} \\
\\
y_{12k} &= \mu_{12} + \epsilon_{12k} \\
&= \mu_{..} + \alpha_1 + \beta_2 + \alpha\beta_{12} + \epsilon_{12k} \\
\\
y_{21k} &= \mu_{21} + \epsilon_{21k} \\
&= \mu_{..} + \alpha_2 + \beta_1 + \alpha\beta_{21} + \epsilon_{21k} \\
\\
y_{22k} &= \mu_{22} + \epsilon_{22k} \\
&= \mu_{..} + \alpha_2 + \beta_2 + \alpha\beta_{22} + \epsilon_{22k}
\end{aligned} \tag{6.6}$$

The model with interaction written as one equation 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.

6.4 Data Decomposition

When a two factor experiment is carried out the data that results 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.

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 μ_{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 means $\mu_{12}, \mu_{21}, \mu_{22}$, respectively. The values $\overline{y}_{1\cdot\cdot} = 10.5, \overline{y}_{2\cdot\cdot} = 6.88, \overline{y}_{\cdot 1\cdot} = 7.50$, and $\overline{y}_{\cdot 2\cdot} = 9.88$ are sample estimates of $\mu_{1\cdot}, \mu_{2\cdot}, \mu_{\cdot 1}, \mu_{\cdot 2}$, respectively. The values $\hat{\alpha}_1 = 1.81$ and $\hat{\alpha}_2 = -1.81$ are estimates of the true effects α_1 and α_2. The values $\hat{\beta}_1 = -1.19$ and $\hat{\beta}_2 = 1.19$ are estimates of the true effects β_1 and β_2. Finally $\overline{y}_{\cdots} = 8.69$ is an estimate of the true grand mean $\mu_{\cdots}$.

We can write each observed tomato yield y in terms of the estimated parameters. As an example,

Table 6.4: Tomato Production with Means

		B_1		B_2		
A_1	$\overline{y}_{11\cdot} = 9.25$	8 8 9 12	$\overline{y}_{12\cdot} = 11.75$	11 11 12 13	$\overline{y}_{1\cdot\cdot} = 10.5$	$\hat{\alpha}_1 = 1.81$
A_2	$\overline{y}_{21\cdot} = 5.75$	5 6 6 6	$\overline{y}_{22\cdot} = 8.00$	7 8 8 9	$\overline{y}_{2\cdot\cdot} = 6.88$	$\hat{\alpha}_2 = -1.81$
		$\overline{y}_{\cdot 1\cdot} = 7.50$ $\hat{\beta}_1 = -1.19$		$\overline{y}_{\cdot 2\cdot} = 9.88$ $\hat{\beta}_2 = 1.19$		$\overline{y}_{\cdots} = 8.69$

$$
\begin{aligned}
y_{111} = 8 &= \overline{y}_{11\cdot} + e_{111} \\
&= 9.25 + (8 - 9.25) \\
&= 9.25 + (-1.25) \\
&= 8.69 + 1.81 + (-1.19) + [9.25 - (8.69 + 1.81 - 1.19)] + (-1.25) \\
&= 8.69 + 1.81 + (-1.19) + (-0.06) + (-1.25) \\
&= \overline{y}_{\cdots} + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\alpha\beta}_{11} + e_{111}
\end{aligned}
$$

$$
\begin{aligned}
y_{123} = 12 &= \overline{y}_{12\cdot} + e_{123} \\
&= 11.75 + (12 - 11.75) \\
&= 11.75 + 0.25 \\
&= 8.69 + 1.81 + (1.19) + [11.75 - (8.69 + 1.81 + 1.19)] + (0.25) \\
&= 8.69 + 1.81 + 1.19 + 0.06 + 0.25 \\
&= \overline{y}_{\cdots} + \hat{\alpha}_1 + \hat{\beta}_2 + \hat{\alpha\beta}_{12} + e_{123}
\end{aligned}
$$

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.

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

$$SSTOT = SSGM + SSA + SSB + SSAB + SSE$$

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 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

$$SSTOT_C = SST - SSGM$$

Degrees of freedom associated with $SST(corrected) = N - 1$. In this example, $SST(corrected) = 1299 - 1208.26 = 90.74$ and $df = N - 1 = 16 - 1 = 15$. When correcting for the 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, for A main effects, and 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 expected values of the various mean squares can be shown to be

$$E[MSE] = \sigma^2 \tag{6.8}$$

$$E[MSA] = \sigma^2 + \frac{nb\sum_{i=1}^{a}\alpha_i^2}{a-1} \tag{6.9}$$

$$E[MSB] = \sigma^2 + \frac{na\sum_{j=1}^{b}\beta_j^2}{b-1} \tag{6.10}$$

$$E[MSAB] = \sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2}{(a-1)(b-1)} \tag{6.11}$$

Thus if there are no main effects for A, that is all α_i are 0, then $E[MSA]$ and $E[MSE]$ are both equal to σ^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 α_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 α 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 α 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 $SSA = nb\sum_{i=1}^{a}\hat{\alpha}_i^2 = SSA = nb\sum_{i=1}^{a}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot\cdot})^2$ 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.

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" distribution with $\nu_1 = (a-1)$ numerator degrees of freedom and $\nu_2 = (N - ab)$ denominator degrees of freedom.

At a significance level of α 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 α 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 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,

$$H_a : \text{ not all } \mu_{\cdot j}' s \text{ are equal}$$

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" distribution with $\nu_1 = (b-1)$ numerator degrees of freedom and $\nu_2 = (N-ab)$ denominator degrees of freedom.

At a significance level of α 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 α 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 α 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 α 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, μ_{ij}, rather than main effect means are more appropriate. For example treatment combination means involving levels of A are compared and this is

Table 6.7: 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	33.50	16.25	1.35	
Total (Corrected)	15	91.44			

done at each level of factor B. Or treatment combination means involving levels of B are compared and this is done at each level of A. Examples are provided in the following 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 a factor when these comparisons are not representative across all 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 means. The 0.10 level will normally be used for testing interaction in this text unless otherwise stated.

6.6 Examples

6.6.1 Tomato Weight Example

Table 6.7 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 $\alpha = 0.05$ level with $F = 0.05, P = 0.8335$, providing no formal 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 ($F = 38.82, P < 0.0001$) and water ($F = 16.66, P = 0.0015$) affect tomato production.

Table 6.8: Paper Towel Example Data

Paper Towel	Water	Liquid Dishwashing Detergent	Vegetable Oil
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

6.6.2 Paper Towel Example - No Interaction

A former student conducted an experiment to compare the amount 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: 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 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 6.8.

Figure 6.3: Plot of Amount of Liquid Absorbed versus Towel/Liquid

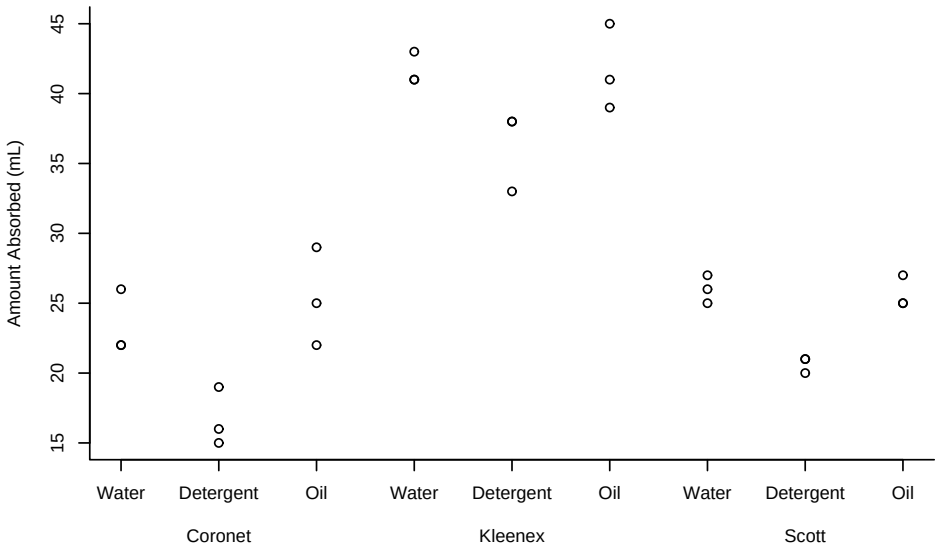


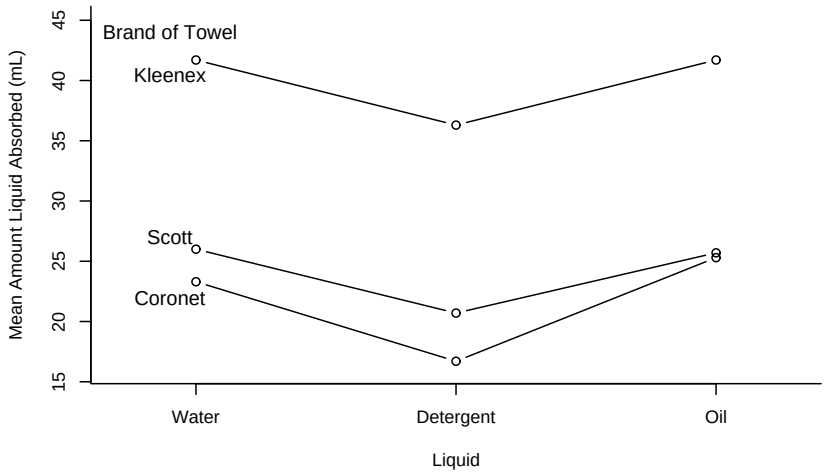
Figure 6.3 is a plot of the amounts of liquid absorbed versus the treatment combination of brand of towel and liquid. A few observations can be made based on the plot. Kleenex appears to have been most absorbent regardless of type of liquid used. The comparison of liquid types is similar across the brands, 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 6.9 and an interaction plot is given in Figure 6.4.

Table 6.9: Means of Amount Absorbed (mL): Paper Towel Example

	Liquid		
	Water	Dishwashing Detergent	Vegetable Oil
Paper Towel			
Coronet	23.3	16.7	25.3
Kleenex	41.7	36.3	41.7
Scott	26.0	20.7	25.7

Figure 6.4: Plot of Amount of Liquid Absorbed versus Towel/Liquid



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 linear model for the data is given by

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \tag{6.12}$$

with

$i = 1, 2, 3$ representing i^{th} level (Coronet, Kleenex, Scott) of Paper Towel

$j = 1, 2, 3$ representing j^{th} level (water, detergent, oil) of Liquid

$k = 1, 2, 3$ is an index on the response amount of liquid absorbed with a treatment combination

y_{ijk} represents the k^{th} observation on amount absorbed for towel i and liquid j

$\mu_{..}$ = the true grand mean of amount absorbed

α_i = the true main effect of the i^{th} level of paper towel on amount absorbed

β_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

Table 6.10: 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			

ϵ_{ijk} = the effects of extraneous variables on the k^{th} amount at the i^{th} paper towel and j^{th} liquid

The ANOVA table for the Paper Towel example is given in Table 6.10. There is no evidence of interaction between Towel Brand and Liquid at the 0.10 level of significance with $F = 0.65, P = 0.6350$. There is evidence at the 0.05 level of both brand effects ($F = 180.05, P < .0001$) and Liquid effects ($F = 22.82, P < .0001$).

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 for the Coronet, Kleenex, and Scott brands, are respectively, $\overline{y}_{1\cdot\cdot} = 21.8$, $\overline{y}_{2\cdot\cdot} = 39.9$, and $\overline{y}_{3\cdot\cdot} = 24.1$

For two levels i and i' of Towel Brand, the general form of the interval for $\mu_{i\cdot} - \mu_{i'\cdot\cdot}$ and simultaneous confidence level of 95% is

$$\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot} \pm \frac{q_{0.05;\nu,a}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{bn} + \frac{1}{bn}}$$

where $\overline{y}_{i\cdot}$ and $\overline{y}_{i\prime\cdot\cdot}$ 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

$$\begin{aligned} 21.8 - 39.9 &\pm \tfrac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\tfrac{1}{9} + \tfrac{1}{9}} \\ 21.8 - 24.1 &\pm \tfrac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\tfrac{1}{9} + \tfrac{1}{9}} \\ 39.9 - 24.1 &\pm \tfrac{3.61}{\sqrt{2}}\sqrt{4.85}\sqrt{\tfrac{1}{9} + \tfrac{1}{9}} \end{aligned}$$

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:

$$\begin{array}{rcccl} -20.8 & \leq & \mu_{1.} - \mu_{2.} & \leq & -15.4 \\ -5.0 & \leq & \mu_{1.} - \mu_{3.} & \leq & 0.4 \\ 13.1 & \leq & \mu_{2.} - \mu_{3.} & \leq & 18.5 \end{array}$$

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 overall 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 for Water, Detergent, and Oil, are, respectively, $\overline{y}_{.1.} = 30.3$, $\overline{y}_{.2.} = 24.6$, and $\overline{y}_{.3.} = 30.9$

The general form of the interval for $\mu_{.j} - \mu_{.j'}$ is

$$\overline{y}_{.j.} - \overline{y}_{.j'.} \pm \frac{q_{0.05;\nu,b}}{\sqrt{2}} \sqrt{MSE} \sqrt{\frac{1}{an} + \frac{1}{an}}$$

where $\overline{y}_{.j.}$ and $\overline{y}_{.j'.}$ refer, respectively, to the marginal means of amount absorbed for levels j and j' of the factor liquid. The denominator $an = (3)(3)$ 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:

$$\begin{array}{rcccl} 3.0 & \leq & \mu_{.1} - \mu_{.2} & \leq & 8.4 \\ -3.3 & \leq & \mu_{.1} - \mu_{.3} & \leq & 2.1 \\ -9.0 & \leq & \mu_{.2} - \mu_{.3} & \leq & -3.6 \end{array}$$

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 overall 95% confidence level.

6.6.3 Example with Interaction

This example is taken from Littel, Stroup, and Freund [12]. 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 6.11.

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 6.12. 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

The interaction plot more clearly shows evidence of interaction between method and variety as noted in Figure 6.5.

The linear model for the data is given by

$$y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \tag{6.13}$$

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)

Table 6.11: 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

Table 6.12: 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	
						$\overline{y}_{\cdots} = 18.4$

Figure 6.5: Plot of Yield versus Method/Variety

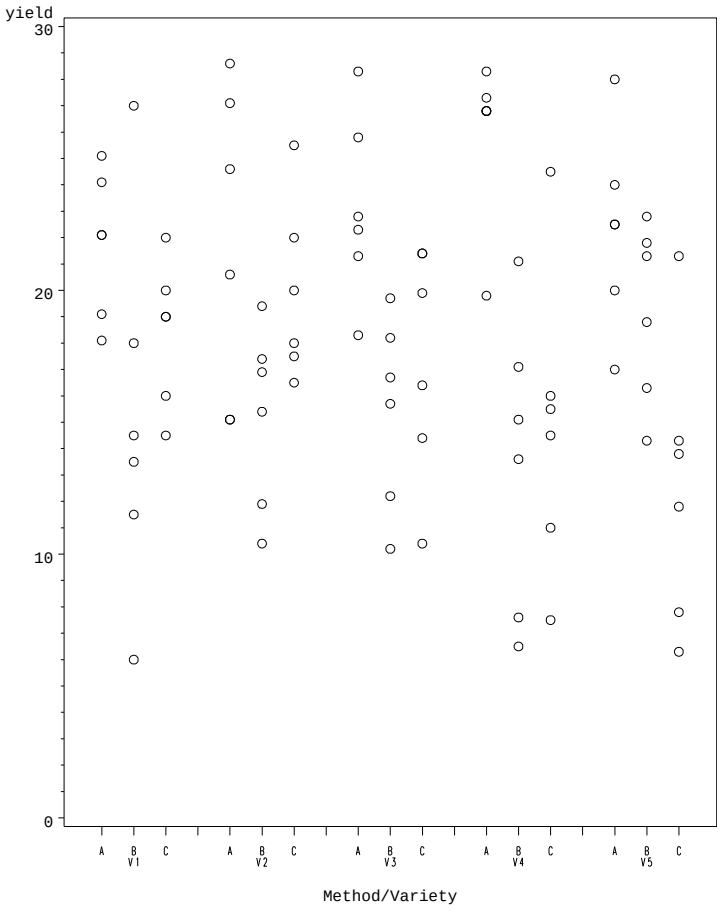


Figure 6.6: Interaction Plot Grass Data

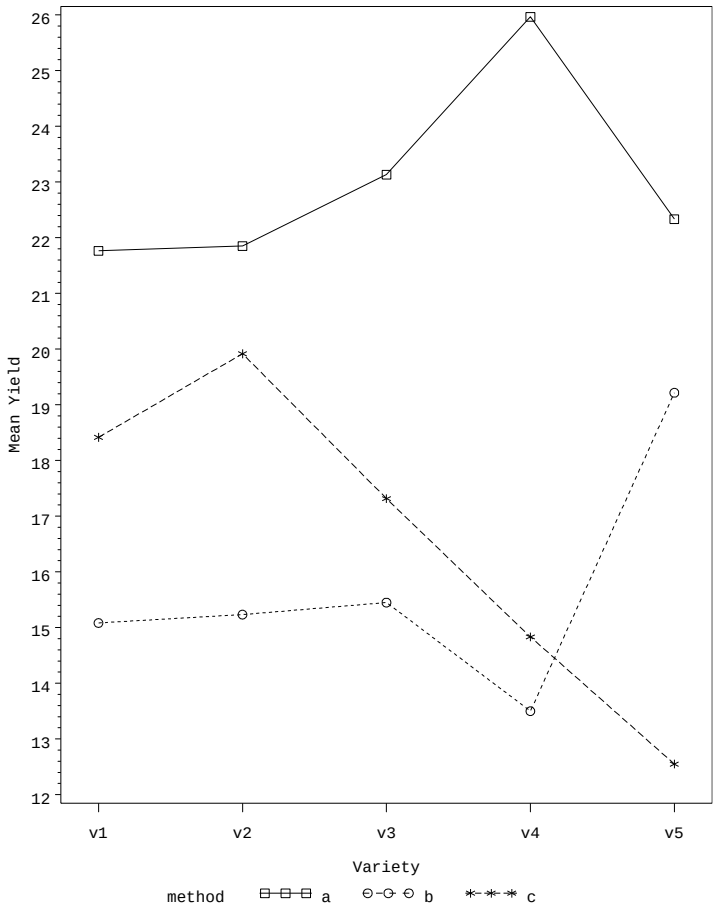


Table 6.13: ANOVA Table for Grasses Example

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			

y_{ijk} represents the k^{th} observation on yield of dry matter for method i and variety j

$\mu_{..}$ = the true grand mean of yield of dry matter

α_i = the true main effect of the i^{th} method on yield

β_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

ϵ_{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.

The ANOVA table for the Grasses example is given in Table 6.13. 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 is the same for the 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 μ_{ij} of 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

$$\begin{aligned}
\overline{y}_{11.} - \overline{y}_{21.} &\pm \frac{q_{0.05;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\tfrac{1}{n}+\tfrac{1}{n}} \\
\overline{y}_{11.} - \overline{y}_{31.} &\pm \frac{q_{0.05;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\tfrac{1}{n}+\tfrac{1}{n}} \\
\overline{y}_{21.} - \overline{y}_{31.} &\pm \frac{q_{0.05;\nu,t}}{\sqrt{2}}\sqrt{MSE}\sqrt{\tfrac{1}{n}+\tfrac{1}{n}}
\end{aligned}$$

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 6.12, 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:

$$\begin{array}{rcrcl} 0.5 & \leq & \mu_{11} - \mu_{21} & \leq & 12.9 \\ -2.8 & \leq & \mu_{11} - \mu_{31} & \leq & 9.6 \\ -9.5 & \leq & \mu_{21} - \mu_{31} & \leq & 2.9 \end{array}$$

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:

$$\begin{array}{rcrcl} 0.4 & \leq & \mu_{12} - \mu_{22} & \leq & 12.8 \\ -4.3 & \leq & \mu_{12} - \mu_{32} & \leq & 8.1 \\ -10.9 & \leq & \mu_{22} - \mu_{32} & \leq & 1.5 \end{array}$$

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:

$$\begin{array}{rcrcl} 1.5 & \leq & \mu_{13} - \mu_{23} & \leq & 13.9 \\ -0.4 & \leq & \mu_{13} - \mu_{33} & \leq & 12.0 \\ -8.1 & \leq & \mu_{23} - \mu_{33} & \leq & 4.3 \end{array}$$

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:

$$\begin{array}{rcrcl} 6.3 & \leq & \mu_{14} - \mu_{24} & \leq & 18.7 \\ 5.0 & \leq & \mu_{14} - \mu_{34} & \leq & 17.4 \\ -7.5 & \leq & \mu_{24} - \mu_{34} & \leq & 4.9 \end{array}$$

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:

$$\begin{array}{rcrcr} -3.1 & \leq & \mu_{15} - \mu_{25} & \leq & 9.3 \\ 3.5 & \leq & \mu_{15} - \mu_{35} & \leq & 15.9 \\ 0.4 & \leq & \mu_{25} - \mu_{35} & \leq & 12.8 \end{array}$$

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 The yield of tomato plants (pounds per plant) depends upon the type of fertilizer used. Two important constituents of fertilizer are (A) potash content (percent) and (B) nitrogen content (percent). At an agricultural experiment station several fertilizer combinations are used. The yield is measured for three tomato plants with each combination. The mean yield at each combination is given below. (Saliva, 1990)

	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.

e. Carry out an F test to determine if there are true Potash main effects.

f. Carry out an F test to determine if there are true Nitrogen main effects.

6.2 The author's son used his Nerf gun to shoot at a target on a glass door. The target wa 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 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 source 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(seconds)
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	28	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.4.

d. At a significance level of $\alpha = 0.05$ are the A main effects significant? Use a critical F value from Table A.4 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.4 to make a decision.

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, σ, 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. Note that **there is still randomization possible**. The brands can be randomly assigned to the 4 tire positions within each car. The purpose of the randomization within each car is to balance out the effects of other extraneous variables such as position effects.

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/objects into blocks, each block containing t subjects/objects. This would include natural groupings such as twins or litters of animals, tire positions on a car, etc.

b. Reuse each subject/object in different time slots, number of time slots equal to the number of treatments.

c. Physically split large chunks of material such as a batch of milk, plot of land, etc. into parts, the number of parts equal to the number of treatments.

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. 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 we would have an **incomplete block design**.

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.

b. (Johnson Sui) Two methods of memorizing difficult material are being tested to determine whether one produces better retention. Nine pairs of students are included in the study. The students in each pair are matched according to IQ and academic background and then assigned to the two methods at random. A memorization test is given to all the students, and the following scores are obtained.

	Student Pair								
Method A	90	86	72	65	44	52	46	38	43
Method B	85	87	70	62	44	53	53	42	46

Each pair of students forms a block.

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. Time slots/periods are the experimental units.

The blocking/grouping here is of time slots by person and then **randomization is undertaken within each grouping**.

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. **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 **after**.

An example of this comes from Moore and McCabe ([15], 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.

c. It is claimed that an industrial safety program is effective in reducing the loss of working hours due to factor 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.

d. 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 Type C or Splitting Material 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. The splitting may be splitting of batches of material, such as milk, cloth, chemical mixture, etc.

   (Johnson & Sui,[9]). 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	F
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.

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 blocking factor as one of the two factors, say A, in a two factor study. If there are a levels of the blocking factor A and b levels of the factor of interest B, then let μ_{ij} be the true mean of the response variable at the i^{ith} level of the blocking factor A and j^{th} level of factor B.
Then the model is

$$\begin{aligned} y_{ij} &= \mu_{ij} + \epsilon_{ij} \\ &= \mu_{..} + \rho_i + \tau_j + (\rho\tau)_{ij} + \epsilon_{ij} \end{aligned} \tag{7.1}$$

where

- $\mu_{..}$ represents the true grand mean
- ρ_i $(i = 1, ..., a)$ represents the true effect of i^{th} level of the blocking factor
- τ_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 factor and the j^{th} level of the factor of interest, and
- ϵ_{ij} represents as usual the effects of extraneous variables on the observation of Y at the ij combination of the blocking factor 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 factor 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 ρ is being used instead of α for the effect of the blocking factor. Also the symbol τ is being used instead of the symbol β for the factor of interest. 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 ϵ_{ijk} with $y_{ij} - \overline{y}_{ij\cdot}$ where $\overline{y}_{ij\cdot}$ is the sample mean at the i^{th} level of the blocking factor and j^{th} level of the factor of interest. However there is only one observation at the i^{th} level of the blocking factor and j^{th} level of A. Thus $\overline{y}_{ij\cdot}$ 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 ϵ_{ij} we get

$$\epsilon_{ij} = y_{ij} - (\mu_{\cdot\cdot} + \rho_i + \tau_j)$$

This suggests estimating the error term ϵ_{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/treatment combination we use a value that served as interaction effect in Chapter 6. This is legitimate assuming there is no true interaction between block and treatment.

To illustrate the ideas consider the following example taken from Kutner , Nachtsheim, Neter, and Li [11].

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

Table 7.1: Auditor Proficiency Measures

	Training Method		
Block	1	2	3
1	73	81	92
2	76	79	89
3	75	76	87
4	74	77	90
5	76	71	88
6	73	75	85
7	68	72	88
8	64	71	82
9	65	73	81
10	62	69	78

Figure 7.1: Plot of Proficiency Measure versus Block by Treatment

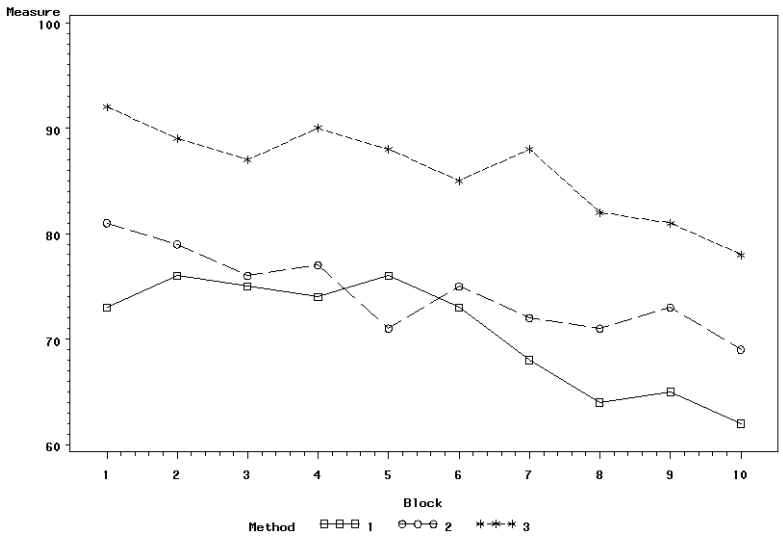


Figure 7.1 provides a plot of the proficiency measures versus Block by Method. Note that Method 3 results in the highest measures regardless of the amount of time elapsed since college graduation. Also there is no evidence of interaction between block and method.

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 mean proficiency for block 1 is $\overline{y}_{1.} = 82$ and the 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 we have that 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:

$$y_{11} = 73 = 77 + 5 + (-6.4) + (-2.6).$$

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 these sums of squares and mean squares. 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

Table 7.2: Decomposition Table - Auditor Data

i	j	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

the graphs representing the different treatments are roughly parallel. Figure 7.1 indicates that the assumption of no interaction between block and method is reasonable for the auditor example.

7.3.2 F test for Treatment Effect in a 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

$$SSTR = a\sum_{j=1}^{b} \hat{\tau}_j^2 = a\sum_{j=1}^{b} (\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})^2$$

and

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b} [e_{ij}]^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} [y_{ij} - (\overline{y}_{\cdot\cdot} + \hat{\rho}_i + \hat{\tau}_j)]^2$$

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common standard variances holds, then $MSTR/MSE$ has the "F" 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 α 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 α 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 reduce experimental error associated with the presumed relationship between the blocking factor 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,

$$H_o : \mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{a\cdot}$$

The alternative hypothesis is

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

or equivalently,

$$H_a : \text{ not all } \rho_{i\cdot}{}'s \text{ are equal}$$

The test statistic is

$$F = \frac{MSBL}{MSE} = \frac{SSBL/(a-1)}{SSE/(a-1)(b-1))}$$

where

$$SSBL = b\sum_{i=1}^{a} \hat{\rho}_i^2 = b\sum_{i=1}^{a} (\overline{y}_{\cdot i} - \overline{y}_{\cdot\cdot})^2$$

and

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b} [e_{ij}]^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} [y_{ij} - (\overline{y}_{\cdot\cdot} + \hat{\rho}_i + \hat{\tau}_j)]^2$$

If the null hypothesis is true and the assumption of independent, normally distributed errors with mean 0 and common standard variances holds, then $MSBL/MSE$ has the "F" 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 α 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 α 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

$$E[MSE] = \sigma^2 \tag{7.2}$$

$$E[MSBL] = \sigma^2 + \frac{b\sum_{i=1}^{a}\rho_i^2}{a-1} \tag{7.3}$$

$$E[MSTR] = \sigma^2 + \frac{a\sum_{j=1}^{b}\tau_j^2}{b-1} \tag{7.4}$$

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.

7.3.3 Pairwise Comparisons Using the Tukey-Kramer Procedure

The general form of endpoints for the Tukey-Kramer interval for the difference of two treatment means $\mu_{\cdot j} - \mu_{\cdot j'}$, adapted from Chapter 6, is

$$\overline{y}_{\cdot j} - \overline{y}_{\cdot j'} \pm \frac{q_{\alpha;\nu,b}}{\sqrt{2}}\sqrt{MSE}\sqrt{\frac{1}{a}+\frac{1}{a}}$$

where $\overline{y}_{\cdot j}$ and $\overline{y}_{\cdot j'}$ 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 ν 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 Blocks and Analysis of Block Design

Consider again the model for the block design given earlier:

$$y_{ij} = \mu + \rho_i + \tau_j + \epsilon_{ijk}$$

The following items are noted:

a. ρ_i measures the effect of the i^{th} level of the blocking factor of interest. In this example the block corresponding to a set of 3 individuals who finished school about the same number of years ago. So the block effect is the effect of the length of time corresponding to the particular 3 individuals. τ_j measures the j^{th} level of the treatment factor of interest. Here this is the effect of the particular method of training employed. The $\epsilon_{ij}s$ measure as in other models the effects of extraneous variables associated with the experimental units. In the auditor example the experimental unit corresponds to an individual in the group of three. If the experiment were performed again a different individual from the 3 in the block may be assigned to a method.

b. It is assumed in our model that the ϵ_{ij} are independent normal random variables with mean 0 and standard deviation σ.

c. The model assumes no interaction between the blocking factor 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.

d. 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.

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) then we can use the paired samples t test to compare the two treatments. 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 of two treatments using the paired samples t test is equivalent to the F ratio obtained from an analysis of variance of a block design. An example follows.

The equivalence will be illustrated with the word recall example from Chapter 7. 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 factor 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

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 square of the t statistic is equal to the F ratio. Note here that $0.11^2 = 0.01$. This equivalence only holds for the two sided test.

7.6 Two Blocking Factors – 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 factors: car and wheel position. 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. The letters A, B, C, ... used to denote the treatments are written in the first blocking row and then the remaining rows are obtained by shifting the letters to the left once.

Randomization of treatments within the row blocking variable and within the column blocking variables 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 factors, in general a row blocking factor and a column blocking factor.

b. The number of levels of the row blocking factor is equal to the number of levels of the column blocking factor, 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.

7.6.1 Another Example

Cobb ([3], 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 driver 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 factor being subject. Each subject receives all four treatments using different word list 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.

Problems for Chapter 7

7.1 (Cobb [3], page 300). Interruptions in breathing are particularly common among premature infants. In the past, hospitals have kept babies on ordinary bassinet mattresses, but someone thought the babies might do better on waterbeds because of their gentle rocking motion. To study the effect of waterbeds on breathing, investigators attached an alarm to each of 8 premature babies; the alarm would sound whenever the baby's breathing stopped for more than 20 seconds. Each baby was monitored for two six-hour periods; during one (randomly chosen) period, the baby slept on a waterbed; during the other, the control period, the baby slept on a regular basinet mattress. The researchers recorded the number of times the alarm went off, and they measured the length of time the baby was asleep. The following numbers give the number of interruptions per hour of sleep.

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 In order to check on possible laboratory bias in the reported ash content of coal, 10 samples of coal were split in half and then sent at random to each of 10 laboratories. The laboratories reported the following ash content data:

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 In a study conducted by the Department of Health & Physical Education at the Virginia Polytechnic Institute and State University in 1983, 3 diets

were assigned for a period of 3 days to each of 6 subjects in a randomized complete block design. The subjects were assigned the following 3 diets in a random order:

Diet 1:	mixed fat and carbohydrates
Diet 2:	high fat
Diet 3:	high carbohydrates

At the end of the 3 day period each subject was put on a treadmill and the time to exhaustion, in seconds, was measured. The following data were recorded:

	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 factor 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 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 applications; a proficiency measure based on this analysis was obtained for each auditor. The results were:

	Training Method		
Block	1	2	3
1	73	81	92
2	76	79	89
3	75	76	87
4	74	77	90
5	76	71	88
6	73	75	85
7	68	72	88
8	64	71	82
9	65	73	81
10	62	69	78

a. What is the reason for using the blocking variable "time elapsed since college graduation"? What type of blocking is this?
b. What are the experimental units?
c. What are some extraneous variables that are being controlled by the randomization in each block?

7.5 As a class project students 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 explain how this experiment would be carried out in a completely randomized design.
e. Give a model for the data and describe the terms in the model in context.

f. Construct an analysis of variable table. Use it to determine if there are significant differences in burn time. If there are significant differences use Tukey's multiple comparison procedure to rank the colors.

Chapter 8

Checking Assumptions of Error Terms

8.1 Assumptions

In the models that we have considered there has always been an error term ϵ 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 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 ϵ_{ij} is not actually observed since it depends upon the mean of y_{ij}, μ_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 the error is

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

The obvious estimate of ϵ_{ij} is obtained by substituting for μ_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 ϵ_{ij}. The estimated error e_{ij} can be calculated – the true error ϵ_{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

$$\begin{aligned} e_{ij} &= y_{ij} - \overline{y}_{i.} \\ &= observed - predicted \end{aligned} \tag{8.2}$$

8.1.2 Residuals for Two Factor Completely Randomized Design

The means model for the two factor completely randomized design is from Chapter 6:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk} \tag{8.3}$$

and thus the true error is

$$\epsilon_{ijk} = y_{ij} - \mu_{ij} \tag{8.4}$$

Estimating μ_{ij}, the population treatment mean, with the sample treatment mean or predicted Y, $\overline{y}_{ij.}$ then the estimate e_{ijk} of ϵ_{ijk} is

$$e_{ijk} = y_{ij} - \overline{y}_{ij.} \tag{8.5}$$

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 Block Design

The model for the one factor block design 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 ϵ_{ij} we get

$$\epsilon_{ij} = y_{ij} - (\mu_{..} + \rho_i + \tau_j)$$

This suggests estimating the error term ϵ_{ij} with

$$\begin{aligned} e_{ij} &= \textit{observed} - \textit{predicted} \\ e_{ij} &= y_{ij} - (\overline{y}_{..} + \hat{\rho}_i + \hat{\tau}_j) \end{aligned} \tag{8.6}$$

where $\hat{\rho}_i = \overline{y}_{i.} - \overline{y}_{..}$ and $\hat{\tau}_j = \overline{y}_{.j} - \overline{y}_{..}$.

8.2 Checking for Independence

The ϵ_{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 ([5], 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 with one battery of each kind being tested in a random order.

The data with the time order and the residuals are given in Table 8.1:

Note from the table 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 are given in Table 8.2.

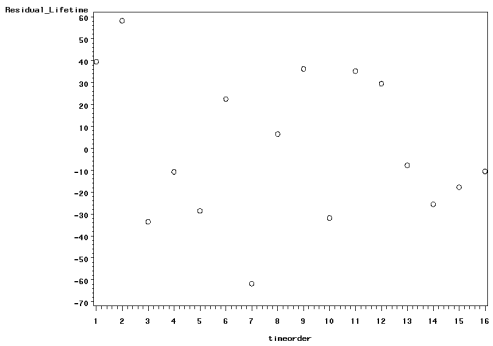
Table 8.1: Lifetime Data

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



The overall F test for the effect of battery type on lifetime 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 the Figure 8.1 Note that the residuals appear to be randomly scattered about 0 providing no evidence of dependence.

Example 8.2 *The experiment (Dean and Voss [5], 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 overall F test for treatment means is significant ($F = 3.85, P = 0.0200$) at the 0.05 level of significance indicating a difference in mean inflation time across 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 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

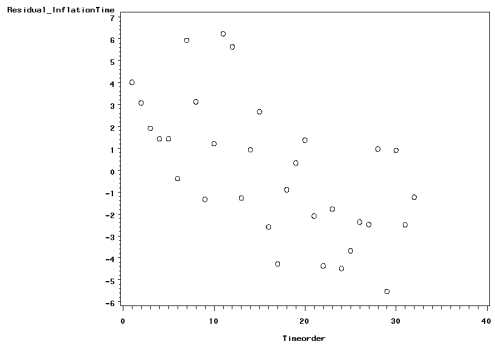
Table 8.3: Inflation Times Data

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



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 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

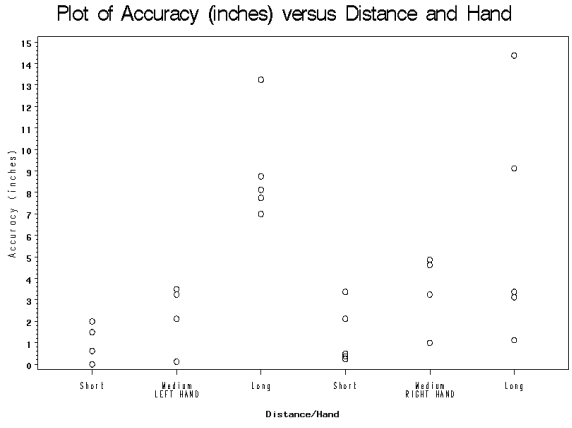
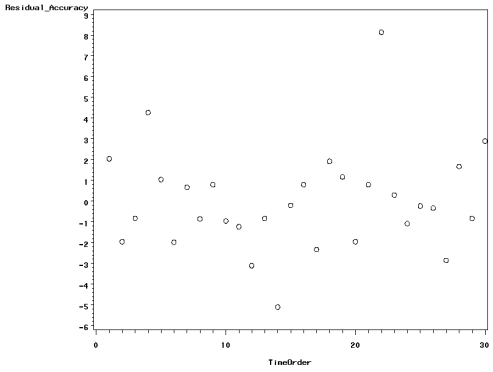


Table 8.6: Nerf Gun Data Means and Standard Deviations

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 error terms are identical, in particular, equal to some common value, σ^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.

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 56.14 and the smallest is 23.6. Thus the ratio of the largest to the smallest is $56.14/23.6 = 2.38 < 3$.

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

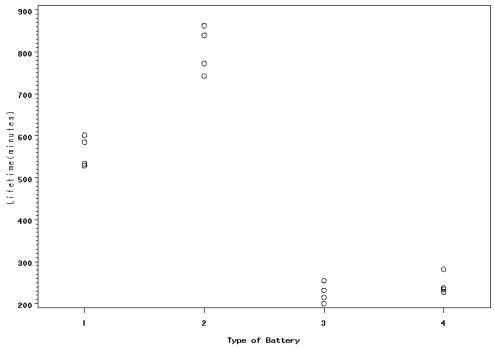


Figure 8.6: Scatterplot of Residuals versus Predicted Lifetimes

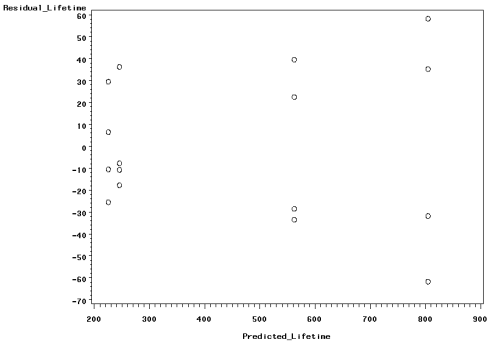
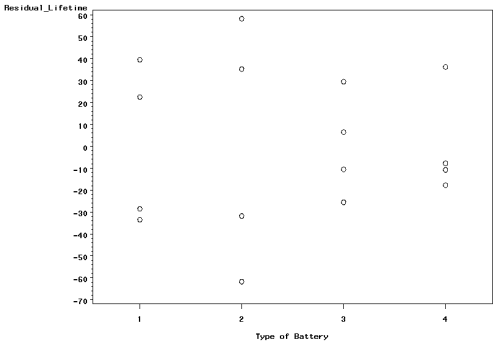


Figure 8.7: Scatterplot of Residuals versus Battery Type



we see that the residuals corresponding to battery type 2 having slightly more spread than the other batter 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 scatterplot of the response variable amount absorbed (mL) 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 $3.05/0.58 = 5.25$, slightly 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) and since the residual plot showed no patterns, we will assume that the

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

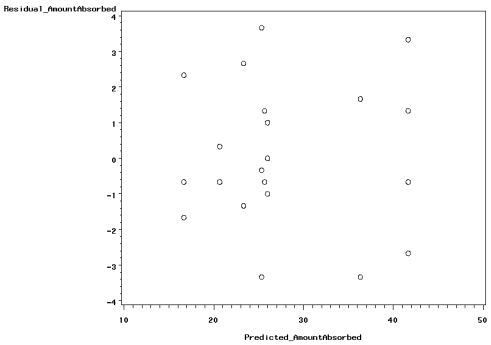
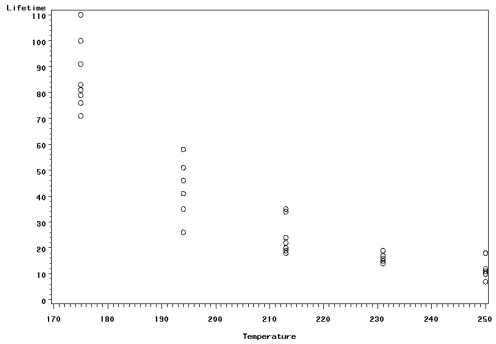


Table 8.8: Lifetimes of a Resin Under Temperature Stress

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



homogeneity assumption holds approximately.

8.3.4 Checking Homogeneity of Variance Data - Resin Example

This example is adapted from Oehlert ([16], 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.

The ratio of the largest to the smallest standard deviation is $13.1/1.8 = 7.3$ quite larger than the rule of thumb value of 3.

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			

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

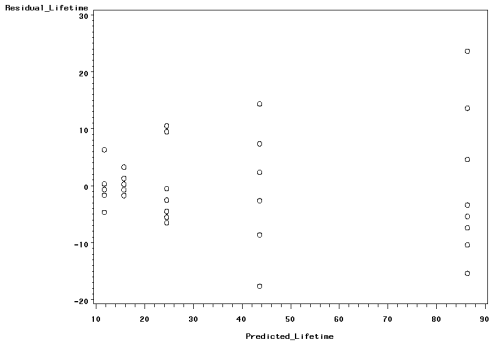
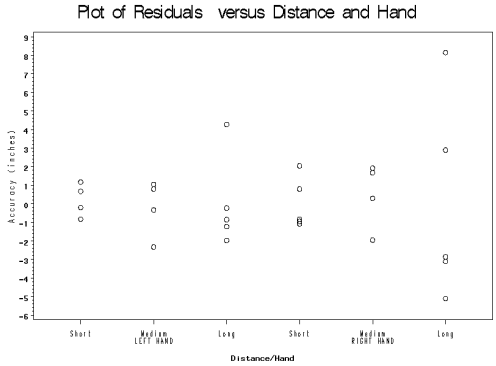


Figure 8.11: Plot of Nerf Gun Residuals versus Treatment



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. Variation appears to increase with distance regardless of the hand.

The means and standard deviations for amount absorbed are given for the treatment combinations 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 3.00. 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 the scatterplot, 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 treatments. 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

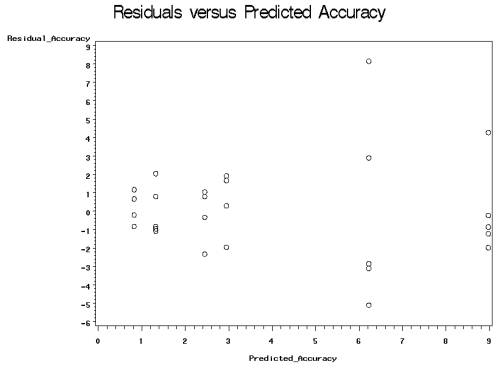


Figure 8.13: Plot of Auditor Example Residuals versus Block

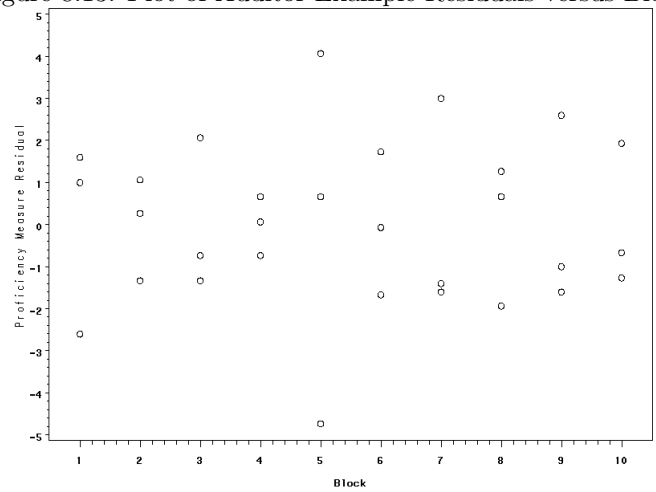


Figure 8.14: Plot of Auditor Example Residuals versus Method

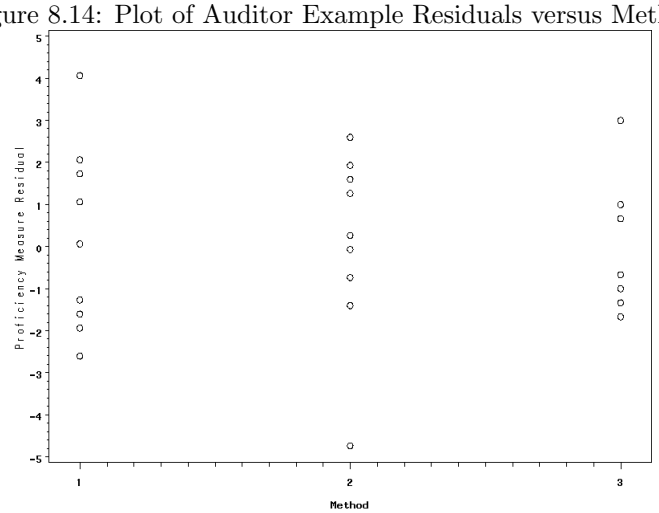
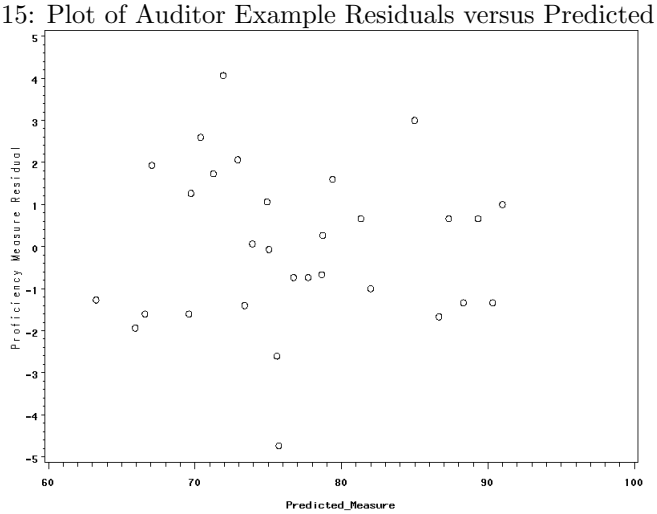


Figure 8.15: Plot of Auditor Example Residuals versus Predicted Measure



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 the 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 the z_7 would refer to the 7^{th} smallest value 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 is 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

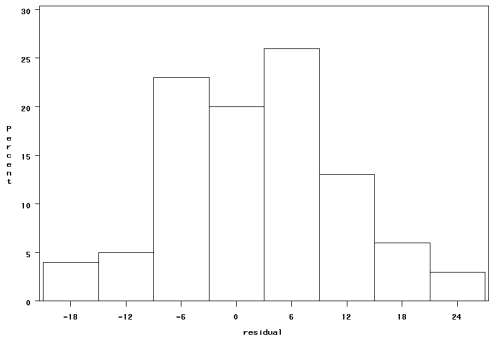


Figure 8.17: Residual QQPlot: Normal Distribution

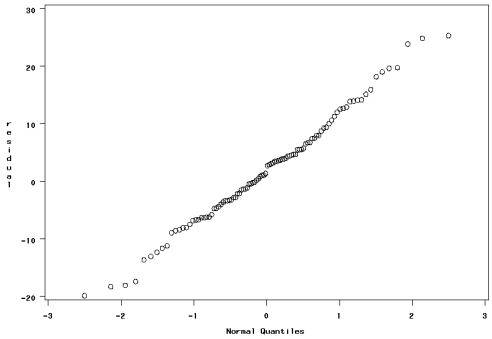


Figure 8.18: Residual Histogram: Heavy Tail Distribution

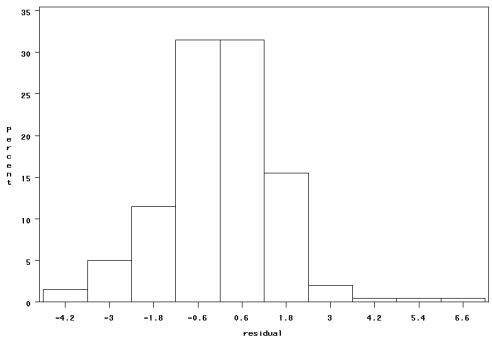


Figure 8.19: Residual QQPlot: Heavy Tail Distribution

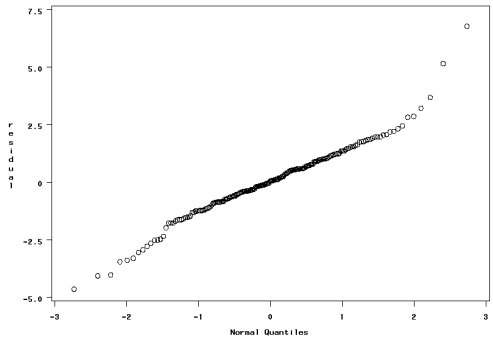


Figure 8.20: Residual Histogram: Light Tail Distribution

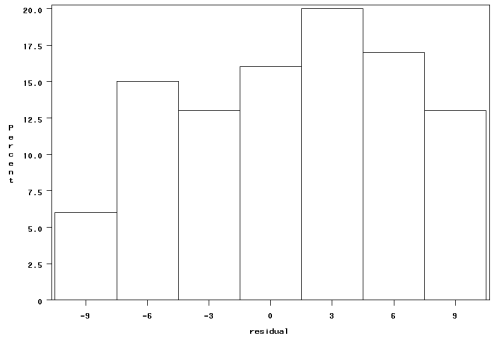


Figure 8.21: Residual Q-Q Plot: Light Tail Distribution

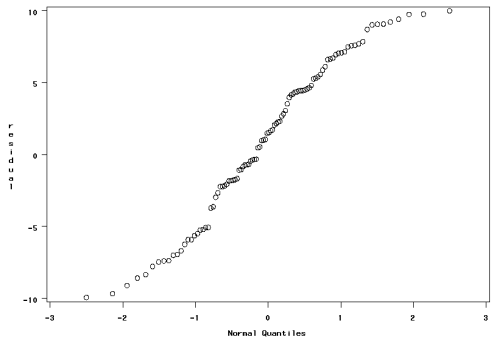


Figure 8.22: Residual Histogram: Right Tail Distribution

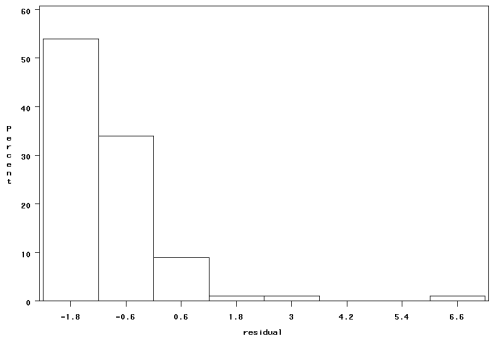


Figure 8.23: Residual Q- Q Plot: Right Tail Distribution

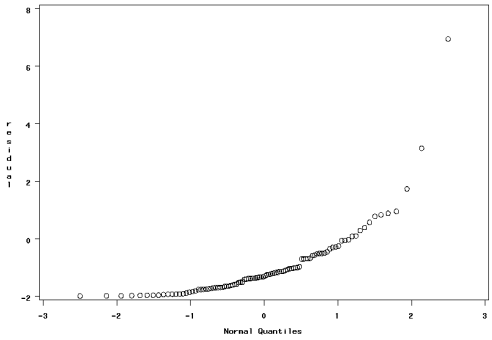


Figure 8.22 and Figure 8.23 give a typical histogram and normal Q-Q Plot when the error terms have an asymmetric "right skewed" distribution, that is the right tail of the distribution is more spread out than the left tail.

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.

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 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

Figure 8.24: Residual Histogram: Left Tail Distribution

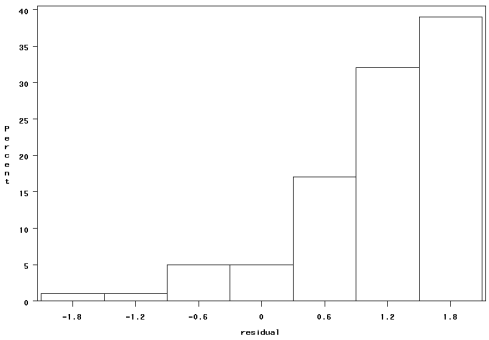


Figure 8.25: Residual Q-Q Plot: Left Tail Distribution

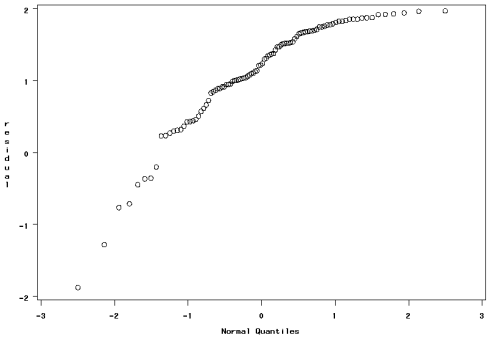


Figure 8.26: Histogram of Residuals for Battery Lifetime Data

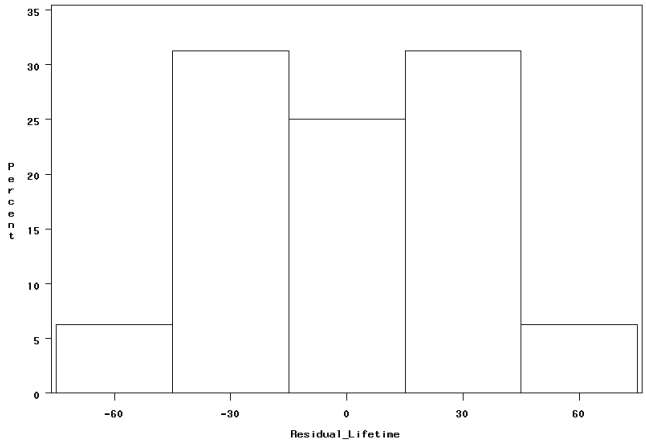
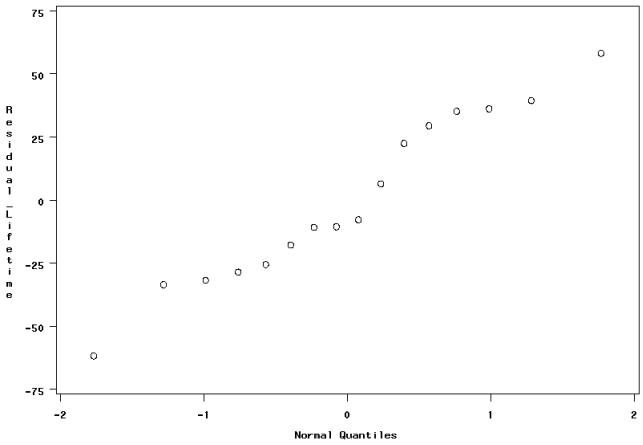


Figure 8.27: QQPlot of Residuals for Battery Lifetime Data



of error variances.

8.4.2 Checking Normality of the Errors for the Paper Towel Example

Figures 8.28 and 8.29 provides 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

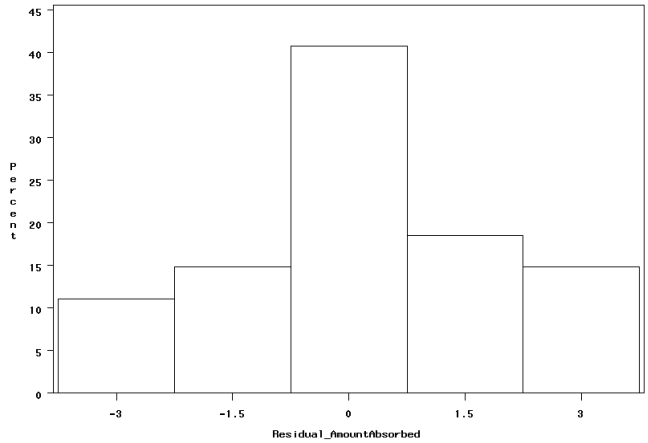


Figure 8.29: QQPlot of Residuals for Paper Towel Absorption Data

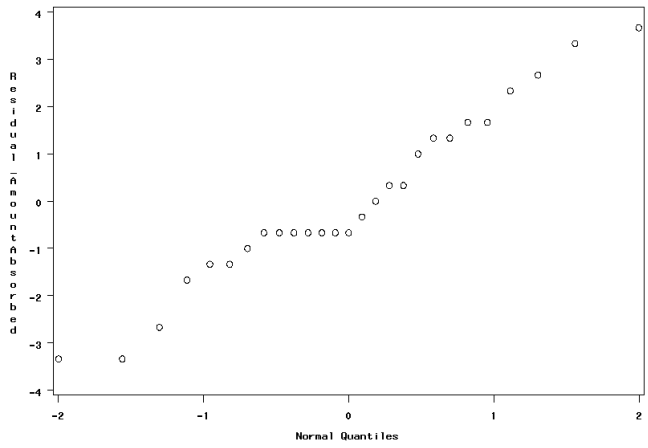


Figure 8.30: Histogram of Residuals for Auditor Proficiency Data

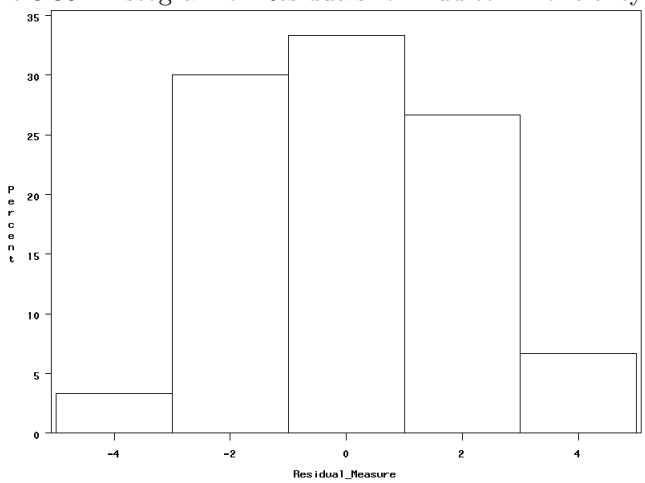
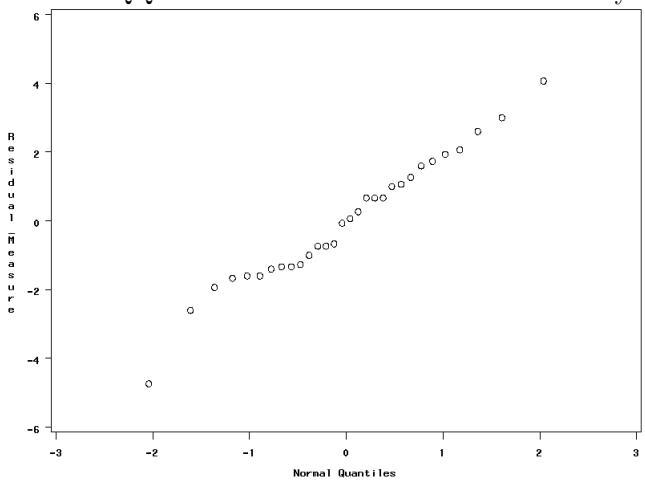
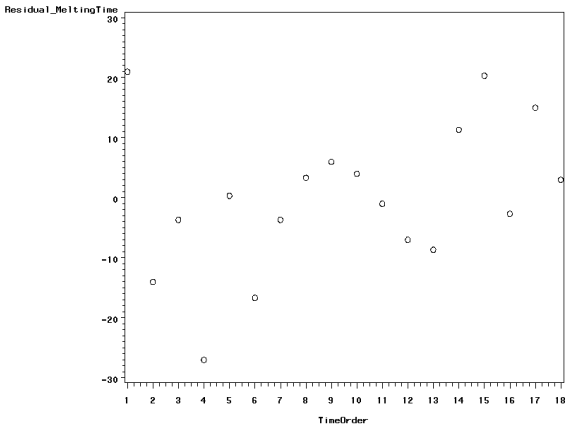


Figure 8.31: QQPlot of Residuals for Auditor Proficiency Data



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.

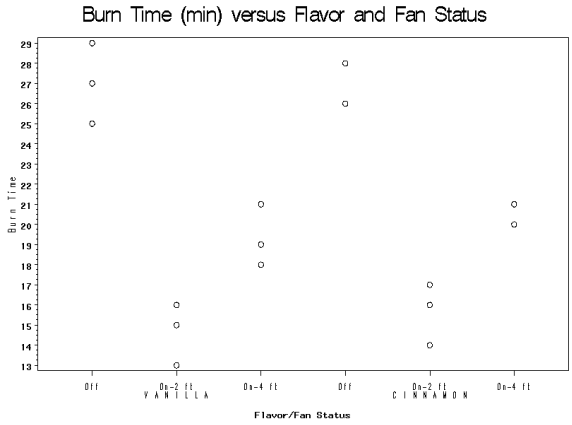


8.2 An incense burning experiment was run in Fall of 2008 by David Gately to study the effects of fan status (off, 2 feet from incense, 4 feet from incense) and flavor of incense stick (vanilla, cinnamon) on the amount of time (to the nearest minute) it took the stick of incense to burn out. The data are reported in the following table. The experimental design was completely random 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.

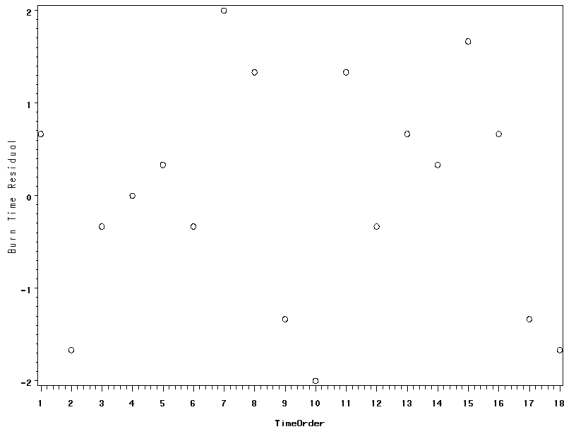
Fan Status	Flavor	Burning Time	TimeOrder	Predicted	Residual
On2	Vanilla	15	5	____	____
On2	Vanilla	16	11	____	____
On2	Vanilla	13	18	____	____
On2	Cinnamon	14	2	____	____
On2	Cinnamon	17	8	____	____
On2	Cinnamon	16	14	____	____
On4	Vanilla	19	6	____	____
On4	Vanilla	21	15	____	____
On4	Vanilla	18	17	____	____
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).

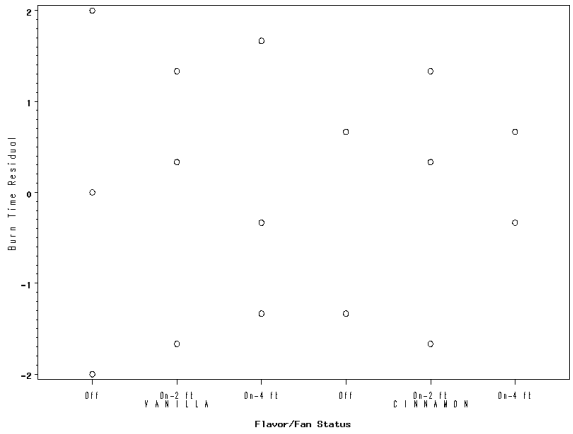
b. Consider the following plot of burn times versus treatment. Can this plot be used to check any of the assumptions about the error terms? Explain in the context of this data.



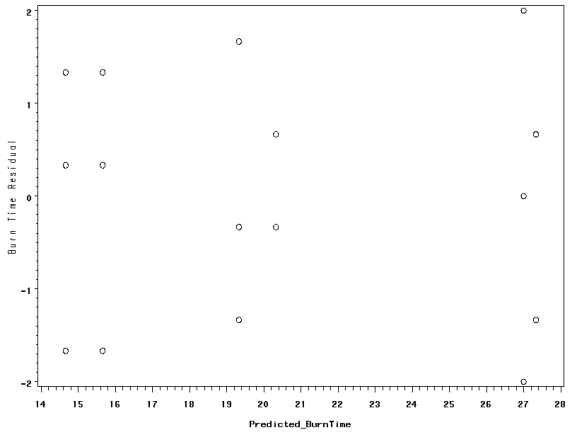
c. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



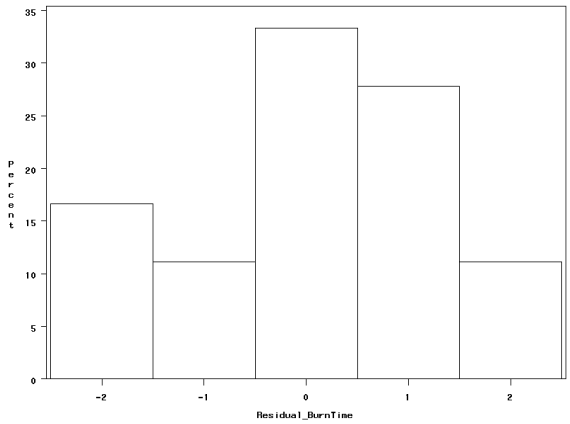
d. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



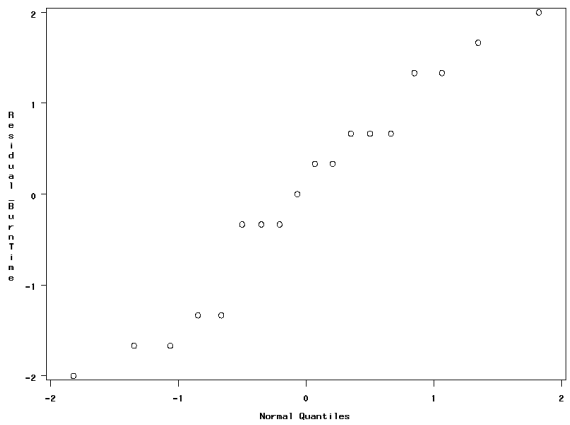
e. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



f. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



g. Consider the following residual plot. What assumption is this plot used to check? Comment on the assumption for this data.



Chapter 9

Split Plot Designs

The simplest of the split plot designs is a two factor design. We have studied the completely randomized two factor design 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 B are applied to subunits of the units assigned to 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, subgroup 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 split plot design the levels of A are assigned to larger "whole units" and the levels of B are assigned to smaller "subunits" of the whole units. In the education experiment above the whole units are the classes of students and the groups of students within each class are the "subunits." In an agricultural experiment the whole units are often large plots of land and the subunits are subdivisions of the large plot, or split plots. Hence the name split plot design.

9.1 Arrangement of Whole Units

The levels of A can be assigned to the whole units in a completely randomized design or the whole units might be blocked and the levels of A assigned to

the whole units within each block. Regardless of the treatment design for 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 Whole Units Arranged Completely at Random

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. Within each of the classes 2 groups of students are formed. These groups within each of the classes form the subunits of the experiment. Then the two levels of factor B (library usage or not) are assigned at random to the two subgroups within each class. Note that each whole unit (a class) is also a block of split units (2 subgroups in the class).

9.1.2 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 blocked into 5 blocks each with 2 plots. The arranging 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 plots (whole units) within each block. Thus the whole units in this experiment are arranged in blocks of two. Each plot 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 the same crop planted on all 5 x 6 = 30 subplots.

This example gives one reason why the split plot design is used. Sometimes larger units are simply required for the levels of one factor A and smaller units can be used for the levels of the other factor B. Here irrigation method would require large plots of land because of the equipment while type of fertilizer can be applied to smaller plots.

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 arranged 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 factor A, $j = 1, ..., b$, with b being the number of levels of 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 at the i^{th} level of the factor A, k^{th} whole unit receiving or "nested" within the i^{th} level of A, getting j^{th} level of the factor B.
- μ 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.
- α_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.
- β_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
- ϵ^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 ϵ^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 σ^2_w and that the split unit errors are independent normal random variables each with mean 0 and common variance σ^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 does hypothesize that the observations on Y for the split units within each whole unit are correlated since those observations have a common factor, that being a common whole unit.

9.2.2 The ANOVA Table

The ANOVA table is derived in a manner similar to how the ANOVA table is derived in 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

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		σ_s^2

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 ANOVA table 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. Also the values Q_1, Q_2, and Q_3 in the EMS column are, respectively, functions of the A effects, B effects, and AB effects, which are zero if the corresponding effect is 0. The column EMS gives the expected or population average mean squares.

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 *As a class project John Szarka and Zamda Lumbi in 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*

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

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 rolls 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. The design structure for the whole units is completely randomized. 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 section is baked. The split plot 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 $i = 1(white), 2(wheat), a = 3(bread)$ indexing type of flour, $j = 1(5min), 2(10min), b = 3(15min)$ indexing baking time, and $k = 1, 2, n = 3$ indexing the roll made with a particular flour, 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.
- μ is the grand mean of height change averaged over a population of rolls, all levels of flour type, and all levels of baking time.
- α_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.
- β_j is the true effect of the j^{th} level of baking time on height change of bread
- $\alpha\beta_{ij}$ is the true interaction effect on height change of the i^{th} level of flour type and the j^{th} level of baking time.
- ϵ^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 "roll" errors, $\epsilon^w_{k(i)}$, are independent normal random variables each with mean 0 and common variance σ^2_w and that the "biscuit" errors, ϵ^s_{ijk}, are independent normal random variables each with mean 0 and common variance σ^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.

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 evidence of interaction between type of flour and baking time.

The ANOVA table for the baking experiment is given in Table 9.4. Note that 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. However both the effects of type of flour and baking time are significant at the 0.05 level.

Tukey-Kramer confidence intervals are used to make pairwise comparisons of the different types of flour and and different baking times.

Figure 9.1: Plot of Height Increase versus Flour/Baking Time

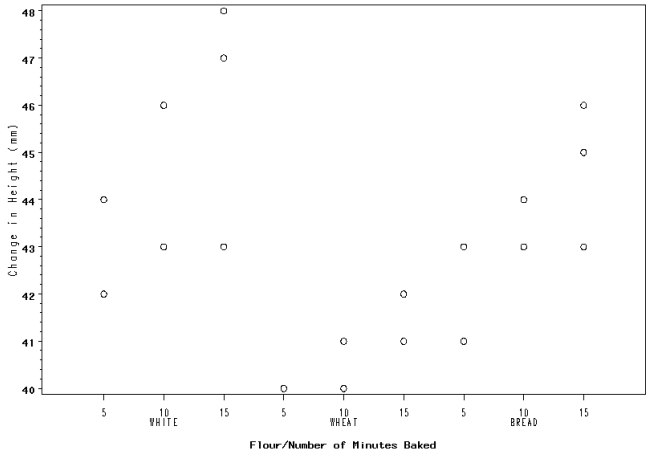


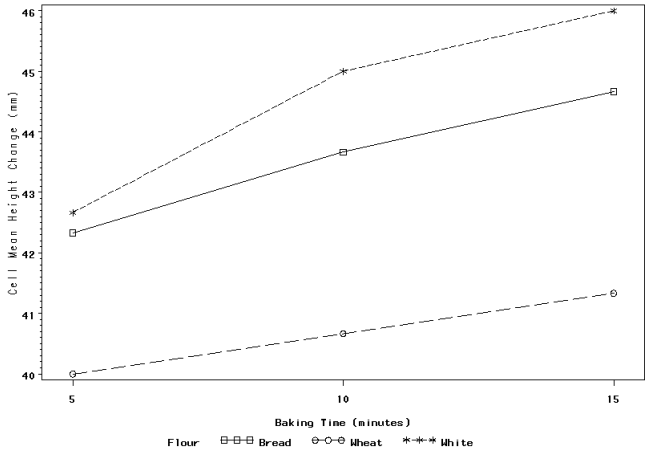
Table 9.3: Height Change Means: Grasses Example

	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	
				$\overline{y}_{...} = 42.9$

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		
Baking Time	2	24.96	12.48	16.85	0.0003
Flour*BakeTime	4	3.48	0.87	1.17	0.3701
Error (Piece)	12	8.89	0.74		

Figure 9.2: InteractionPlot



The Tukey-Kramer confidence intervals with overall confidence level $1-\alpha$ for the levels of the whole plot factor A 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 $\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}$ and $q_{\alpha;\nu,t}$ is the upper α probability point from the Studentized range distribution. Here ν 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, type of flour.

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:

$$(43.6 - 40.7) \pm (3.1)(0.92) \quad (44.6 - 40.7) \pm (3.1)(0.92) \quad (44.6 - 43.6) \pm (3.1)(0.92)$$

Thus the Tukey-Kramer simultaneous 95% confidence intervals are:

$$\begin{array}{rcccl} 0.1 & \leq & \mu_{3\cdot} - \mu_{2\cdot} & \leq & 5.7 \\ 1.1 & \leq & \mu_{1\cdot} - \mu_{2\cdot} & \leq & 6.7 \\ -1.8 & \leq & \mu_{1\cdot} - \mu_{3\cdot} & \leq & 3.8 \end{array}$$

The mean change in height for the Bread and White flours are greater than the mean change in height for the Wheat flour. There is not enough evidence of a difference in mean height change between the White and Bread flours. These conclusions are supported by an overall 95% confidence level.

The Tukey-Kramer confidence intervals with overall confidence level $1-\alpha$ for the levels of the split plot factor B 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 $\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdot j'\cdot}$ and $q_{\alpha;\nu,t}$ is the upper α probability point from the Studentized range distribution. Here ν 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 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 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 $\mu_{\cdot 2} - \mu_{\cdot 1}$, $\mu_{\cdot 3} - \mu_{\cdot 1}$, and $\mu_{\cdot 3} - \mu_{\cdot 2}$ comparing the baking times 5 and 10, 5 and 15, and 10 and 15 minutes are:

$$(43.1 - 41.7) \pm (2.67)(0.41) \quad (44.0 - 41.7) \pm (2.67)(0.41) \quad (44.0 - 43.1) \pm (2.67)(0.41)$$

The Tukey-Kramer simultaneous 95% confidence intervals are:

$$\begin{array}{rcccl} 0.3 & \leq & \mu_{\cdot 2} - \mu_{\cdot 1} & \leq & 2.5 \\ 1.2 & \leq & \mu_{\cdot 3} - \mu_{\cdot 1} & \leq & 3.4 \\ -0.2 & \leq & \mu_{\cdot 3} - \mu_{\cdot 2} & \leq & 2.0 \end{array}$$

Thus mean height change is greater at both the 10 and 15 minute baking times when compared to the 5 minute baking time. There is not enough evidence of a difference in mean height change at the 10 and 15 minute baking times.

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 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 for the fitted model. There appears to be no patterns and thus the assumptions of homogeneity of error variance appears to be satisfied approximately.

Figure 9.3: Histogram of Split Plot Residuals: Baking Experiment

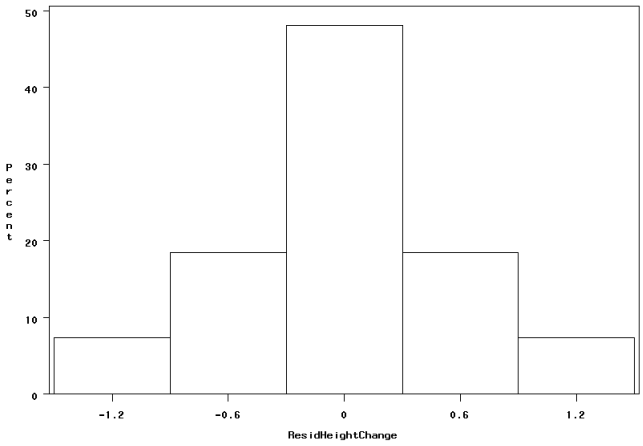
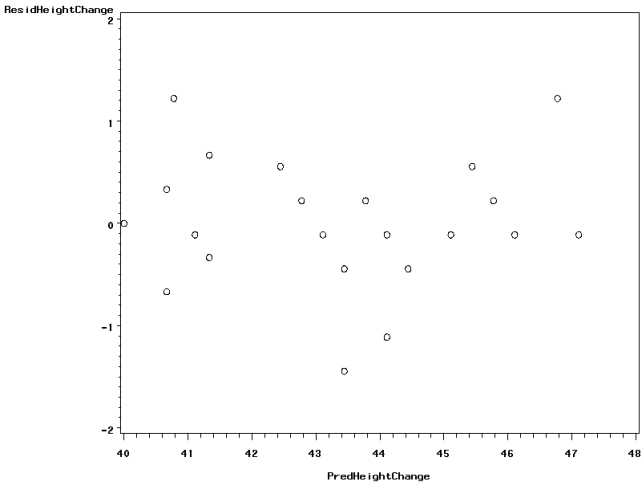


Figure 9.4: Plot of Split Plot Residuals versus Predicted Change: Baking Experiment



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.
- μ 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.
- α_i is the true effect of the i^{th} level of the factor A on the response variable
- ρ_k is the true effect of the k^{th} level of the blocking factor
- ϵ^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.
- β_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
- ϵ^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 σ^2_w and that the split unit errors are independent normal random variables each with mean 0 and common variance σ^2_s. It is also assumed that a whole unit error is independent of a split unit error.

9.3.2 The ANOVA Table

The ANOVA table for the split plot design where the whole units are arranged in blocks is given in Table 9.5. 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 interaction effects, which are zero if the effect is 0.

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

Table 9.5: ANOVA Table: Split Plot Design - Whole Units Blocked

Source of Variation	df	SS	MS	F	EMS
Blocks	r - 1	SSBlocks	MSBlocks		
A	a - 1	SSA	MSA	MSA/MSE_w	$\sigma_s^2 + a\sigma_w^2 + Q_1$
$Error_w$	$(a-1)(r-1)$	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	$MSAB/MSE_s$	$\sigma_s^2 + Q_3$
$Error_s$	$a(r-1)(b-1)$	SSE_s	MSE_s		σ_s^2

Table 9.6: Quality Data for Chocolate Cake Experiment

		Recipe								
Replication/Block		R1			R2			R3		
	Temp	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

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 the computer.

9.3.3 Example

Example 9.2 *This example is based on an experiment described in Cochran and Cox [4]. 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 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 9.6. 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 units are the three portions of a batch of cake batter prepared at a particular large 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 $i = 1(R1), 2(R2), a = 3(R3)$ indexing recipe, $j = 1(175), 2(195), b = 3(215)$ 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
- μ is the grand mean of the quality characteristic averaged over a population of cakes, all levels of recipe, and all levels of temperature.
- α_i is the true effect of the i^{th} level of recipe on quality
- ρ_k is the true effect of the k^{th} replication
- ϵ^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.
- β_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
- ϵ^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 σ^2_w and that the portion errors are independent normal random variables each with mean 0 and common variance σ^2_s. It is also assumed that a batch error is independent of a portion error.

A plot of quality versus recipe and oven temperature is given in Figure 9.5 Recipe does not appear to have an effect on quality. Oven temperature appears to affect the quality.

Quality treatment, marginal, and grand means are provided in Table 9.7.

An interaction plot is given in Figure 9.6. There is no strong evidence of interaction between recipe and baking temperature.

Figure 9.5: Scatterplot of Quality Versus Recipe/Temperature

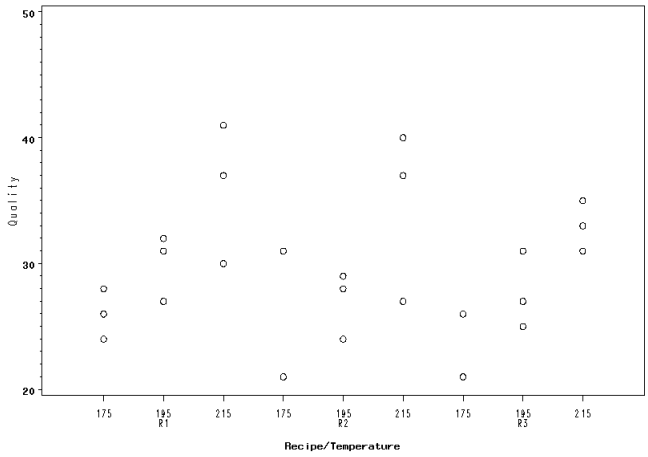


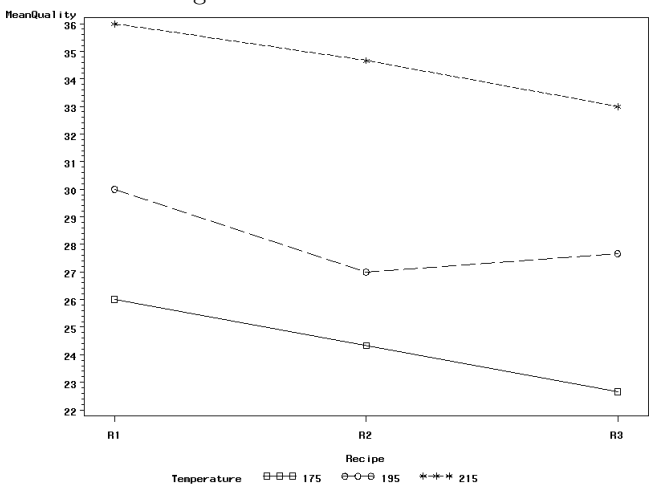
Table 9.7: Quality Means: Chocolate Cake Example

	Baking Temperature (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	
				$\overline{y}_{...} = 29.0$

Table 9.8: 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		

Figure 9.6: Interaction Plot



The ANOVA table for this example is is given in Table 9.8. Note that there is no evidence of interaction between recipe and temperature ($F = 0.15, P - value = 0.9571$). The effects of recipe are not significant ($F = 0.82, P - value = 0.5038$) while the effects of temperature are significant ($F = 26.03, P - value < 0.0001$).

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 $\overline{y}_{i\cdot\cdot} - \overline{y}_{i'\cdot\cdot}$ and $q_{\alpha;\nu,t}$ is the upper α probability point from the Studentized range distribution. Here ν 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 (sample) 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:

$$\begin{array}{rcrcl} -6.2 & \leq & \mu_{1.} - \mu_{2.} & \leq & 10.2 \\ -5.3 & \leq & \mu_{1.} - \mu_{3.} & \leq & 11.1 \\ -7.3 & \leq & \mu_{2.} - \mu_{3.} & \leq & 9.1 \end{array}$$

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

$$\overline{y}_{.j.} - \overline{y}_{.j'.} \pm \frac{q_{\alpha;\nu,t}}{\sqrt{2}} \sqrt{MSE_s} \sqrt{\frac{1}{an} + \frac{1}{an}}$$

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 α probability point from the Studentized range distribution. Here ν 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.

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 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 $\frac{3.77}{\sqrt{2}} = 3.67$. Thus the endpoints for the intervals for differences in temperatures $\mu_{.2} - \mu_{.1}$, $\mu_{.3} - \mu_{.1}$, and $\mu_{.3} - \mu_{.2}$ comparing the temperatures 175 and 195, 175 and 215, and 195 and 215 are:

$$(28.2 - 24.3) \pm (2.67)(1.43) \quad (34.6 - 24.3) \pm (2.67)(1.43) \quad (34.6 - 28.2) \pm (2.67)(1.43)$$

The Tukey-Kramer simultaneous 95% confidence intervals are:

$$\begin{array}{rcrcl} 0.08 & \leq & \mu_{.2} - \mu_{.1} & \leq & 6.6 \\ 6.5 & \leq & \mu_{.3} - \mu_{.1} & \leq & 14.1 \\ 2.6 & \leq & \mu_{.3} - \mu_{.2} & \leq & 10.2 \end{array}$$

All pairwise comparisons of mean quality are significant.

A check is made of the assumptions of normality and homogeneous error variances associated with the split plot errors. Figure 9.7 gives a histogram of the split plot residuals from the model. Normality appears to be satisfied approximately.

Figure 9.8 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 assumptions of homogeneity of error variance appears to be satisfied approximately.

Figure 9.7: Histogram of Residuals from Cake Experiment

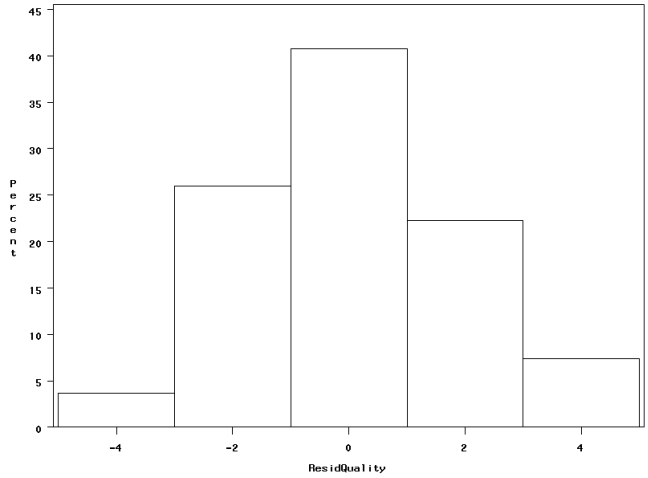
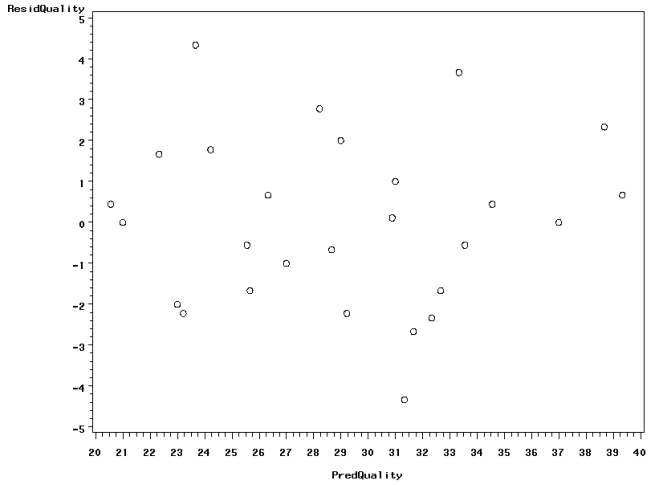


Figure 9.8: Scatterplot of Residuals versus Predicted for Cake Baking Experiment



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 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*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 at each of 70, 75, and 80 degrees. 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. Give a model for this experiment and describe the terms in the model including the error terms.

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. Give a model for this experiment and describe the terms in the model including the error terms.

9.3 As a class project Casey Gundersen in 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 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 a model for this experiment and describe the terms in the model including the error terms.

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 significant.

g. From part (f) there is no evidence of interaction between oven temperature and ice type. Thus test for oven temperature and ice type main effects. Use a significance level of 0.05 for each type. Make appropriate pairwise comparisons using the Tukey-Kramer multiple comparison procedure 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.

9.4 Milliken and Johnson [14], 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. 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 a model for this experiment and describe the terms in the model including the error terms.

e. Use a statistical computing program to obtain an ANOVA table for the data.

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. Make appropriate pairwise comparisons using the Tukey-Kramer multiple comparison procedure with an overall confidence level of 0.95.

Bibliography

[1] Agresti, A., C. Franklin. 2007 *Statistics: The Art and Science of Learning from Data* Pearson.

[2] Chance Magazine, 2000.

[3] G. Cobb. 1997. *Introduction to Design and Analysis of Experiments.* Springer

[4] Cochran, W.G., G.M. Cox. 1950 *Experimental Designs* Wiley.

[5] Dean, A., D. Voss. 1999. *Design and Analysis of Experiments* Springer

[6] DeVeaux, R., P. Velleman, D. Bock 2005. *Stats Data and Models* Pearson

[7] Peck, R., C. Olsen, J. Devore. 2005. *Introduction to Statistics and Data Analysis* Second Edition.

[8] Peck, R., J. Devore. 2008. *Statistics The Exploration and Analysis of Data* Sixth Edition.

[9] Johnson, R., Sui

[10] Kuehl, R.O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*, Second Edition

[11] Kutner, M.H., C.J. Nachtsheim, J. Neter, W. Li. 2005. *Applied Linear Statistical Models*, Fifth Edition.

[12] Littel, R.C., W.W. Stroup, R.J. Freund. 2002. *SAS for Linear Models*, Fourth Edition.

[13] McClave, J.T., T. Sincich. 2003. *A First Course in Statistics*, Eighth Edition.

[14] Milliken, G.A., D.E. Johnson. 1992. *Analysis of Messy Data, Volume I: Designed Experiments*

[15] Moore, D., G. McCabe. 2003. *Introduction to the Practice of Statistics.* Freeman.

[16] Oehlert, G.W. 2000. *A First Course in Design and Analysis of Experiments.* Freeman

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 α probability points for the Student t distribution

	Table entries are $t_{\alpha;\nu}$, where $P[t > t_{\alpha;\nu}] = \alpha$							
	α							
ν	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
∞	.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
∞	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
∞	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
∞	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
∞	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 α 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
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250
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
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
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	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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

Table A.8: Upper α 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. 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
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
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
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
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

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

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

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

B.2 Chapter 2

2.1 $\overline{y} = 1.25$, $s = 0.2$, $s_{\overline{y}} = \frac{s}{\sqrt{n}} = \frac{0.2}{\sqrt{48}} = 0.03$

Sample standard deviation $s = 1.25$ measures variation of weight gains of individual pigs in the sample around the sample mean $\overline{y} = 0.2$. 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 populatin mean weight gain.

2.2 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

2.3 a. 1.711

b. 0.80

2.4 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

2.5 $\overline{y} = 244.3$, $s = 12.4$, $t = 1.21$, $P - value = 0.2508 > 0.05$, 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.

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.

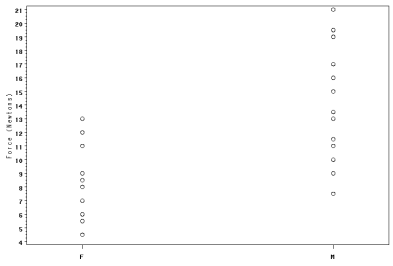
3.6 a. Paired samples design - reusing. Corresponding to each squirrel are two time periods, one when FT twigs are given and one when NFT twigs are given.

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 block 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

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

	Source of Variation	Df	SS	MS	F	P-value
	Grand Mean	1	1728			
a.	Treatments	4	85	21.25	5.06	< 0.01
	Error	25	105	4.2		
	Total	30	1918			

4.2 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
- α_i is the true effect of the i^{th} type of drink on melting time y_{ij}
- ϵ_{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 $\alpha_i's = 0$

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 ϵ_{ij} are independent, normally distributed each with mean of 0 and common standard deviation σ.

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, $P < 0.0001 < 0.05$.

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, $6.6 < \mu_1 - \mu_3 < 11.0$, significant

b. 99% percent confident that all 3 intervals from part(a) are simultaneously correct.

c. narrower

5.3 $t_{0.05/2;\nu=20} = 2.086$, $q^*/\sqrt{2} = 2.95/\sqrt{2} = 2.086$

B.6 Chapter 6

6.1 a. $\hat{\alpha}_1 = -1$, $\hat{\alpha}_2 = 1$

b. $\hat{\beta}_1 = -2.175$, $\hat{\beta}_2 = -0.025$, $\hat{\beta}_3 = 3.025$, $\hat{\beta}_4 = -.825$

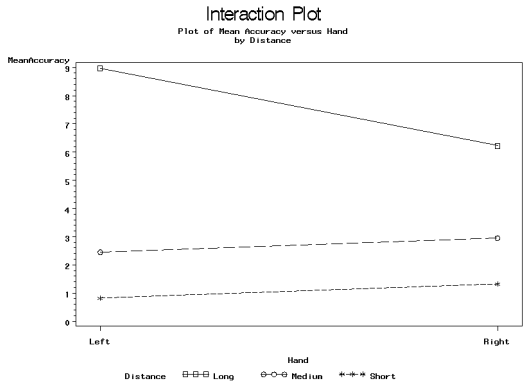
c. $\hat{\alpha\beta}_{11} = 1.85$, $\hat{\alpha\beta}_{21} = -1.85$ $\hat{\alpha\beta}_{12} = 0$, $\hat{\alpha\beta}_{22} = 0$ $\hat{\alpha\beta}_{13} = -0.65$, $\hat{\alpha\beta}_{23} = 0.65$ $\hat{\alpha\beta}_{14} = -1.2$, $\hat{\alpha\beta}_{24} = 1.2$

d. MSAB $= \frac{11.175}{3} = 3.725$, F $= \frac{3.725}{3.625} = 1.03$, $1.03 < F_{0.05;3,16} = 3.24$, 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.

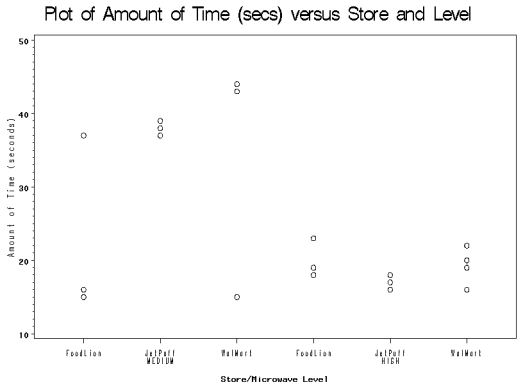


b. Test of interaction between hand and distance not significant at the 0.10 level ($F = 1.20, P - value = 0.3184, \nu_1 = 2, \nu_2 = 24$
Test of Distance Effects significant at 0.05 level $F = 15.74, P - value < 0.0001, \nu_1 = 2, \nu_2 = 24$. For i = 1 (long), i = 2 (short), and i = 3 (short), $\overline{y}_{1.} = 7.60, \overline{y}_{2.} = 2.70, \overline{y}_{3.} = 1.08$,

$$\begin{array}{rcccl} 1.88 & \leq & \mu_{1.} - \mu_{2.} & \leq & 7.92 \\ 3.50 & \leq & \mu_{1.} - \mu_{3.} & \leq & 9.55 \\ -1.40 & \leq & \mu_{2.} - \mu_{3.} & \leq & 4.65 \end{array}$$

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. $y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$ 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
   - $\mu_{..}$ = true grand mean amount of time averaged over all brands and levels
   - α_i is the true effect of the i^{th} brand on amount of time y_{ijk}
   - β_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}
   - ϵ_{ijk} is the effect of extraneous variables on amount of time y_{ijk}.
   - ϵ_{ijk} are independent normal random variables, each with mean 0 and variance σ^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

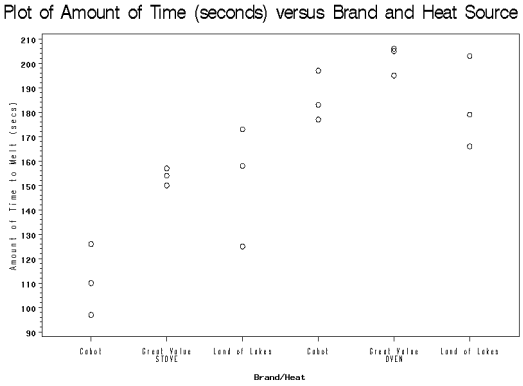
   ii. Interaction is significant ($F = 3.93, P = 0.0384 < 0.10$)
   Comparison of Brands when Setting = Medium (j=1):

$$\begin{array}{rcrcr} 3.8 & \leq & \mu_{21} - \mu_{11} & \leq & 30.7 \\ 2.1 & \leq & \mu_{31} - \mu_{11} & \leq & 29.0 \\ -15.2 & \leq & \mu_{31} - \mu_{21} & \leq & 11.7 \end{array}$$

Comparisons of Brands when Setting = High (j=2):

$$\begin{array}{rcrcr} -16.0 & \leq & \mu_{22} - \mu_{12} & \leq & 11.0 \\ -13.7 & \leq & \mu_{32} - \mu_{12} & \leq & 13.2 \\ -11.2 & \leq & \mu_{32} - \mu_{22} & \leq & 11.2 \end{array}$$

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.



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
   - α_i is the true effect of the i^{th} brand on amount of time y_{ijk}
   - β_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}
   - ϵ_{ijk} is the effect of extraneous variables on amount of time y_{ijk}.
   - ϵ_{ijk} are independent normal random variables, each with mean 0 and variance σ^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):

$$\begin{aligned} -16.0 &\leq \mu_{21} - \mu_{11} \leq 48.6 \\ -35.3 &\leq \mu_{31} - \mu_{11} \leq 29.3 \\ -19.3 &\leq \mu_{31} - \mu_{21} \leq 13.0 \end{aligned}$$

Comparisons of Brands when Heat = Stove (j=2):

$$\begin{aligned} 10.4 &\leq \mu_{22} - \mu_{12} \leq 75.0 \\ 8.7 &\leq \mu_{32} - \mu_{12} \leq 73.3 \\ -34.0 &\leq \mu_{32} - \mu_{22} \leq 30.6 \end{aligned}$$

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 = 400/28 = 14.3; F = 13.3/14.3 = 0.93 < 2.45, not significant.

d. MSA = 310/3 = 103.3, MSE = 400/28 = 14.3, F = 103.3/14.3 = 7.2 > 2.95, significant

e. SSB = 100, MSB = 100/2 = 50.0, MSE = 400/28 = 14.3, F = 50.0/14.3 = 3.5 > 3.34, significant

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. Time elapsed since college graduation is an extraneous variable that would presumably affect proficiency score. Type A blocking.

b. Experimental units are 30 subjects (grouped by time elapsed since graduation).

c. Differences in ability of 3 persons within each block. Different testing conditions for three persons within a block.

7.5 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
- ρ_i is the true effect of the i^{th} block on the burning time y_{ij}
- τ_j is the true effect of the j^{th} color on the burning time y_{ij}
- ϵ_{ij} is the effect of extraneous variable on the burning time y_{ij}

	Source of Variation	Df	SS	MS	F	P-value
	Color	3	12398.4	4132.8	2.77	0.0713
f.	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.

B.8 Chapter 8

8.1 Yes, 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.

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.

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{B.1}$$

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.
- μ is the grand mean of growth averaged over a population of chambers, all levels of temperature, and all levels of strain of petunia.
- α_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.
- β_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.
- ϵ^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.

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.

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{B.2}$$

where $i = 1, 2, 3$ indexes 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.
- μ 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.
- α_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.
- β_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.
- ϵ^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.

9.3 a. Whole plot factor is oven temperature. Whole plot EU is oven run/session. Characteristics of oven run/session such as slight variations 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 plot factor is Type of Ice Cube (bottle, tap, and salt). Split plot experimental unit is ice cube. Extraneous variables include size of cube, temperature variability within parts of oven, etc.

d.

$$y_{ijk} = \mu + \alpha_i + \epsilon^w_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{B.3}$$

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.
- μ 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.
- α_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
- β_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.
- ϵ^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.

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	43717.7		
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 $i = 1(250), i = 2(300), i = 3(350)$

$$\begin{aligned} -6.9 &\leq \mu_{1\cdot} - \mu_{2\cdot} \leq 183.1 \\ 44.1 &\leq \mu_{1\cdot} - \mu_{2\cdot} \leq 234.1 \\ -44.0 &\leq \mu_{2\cdot} - \mu_{3\cdot} \leq 146.0 \end{aligned}$$

Evidence of Ice Type main effects (F = 13.53, P-value = 0.0008)
Tukey-Kramer pairwise comparisons of ice cube type with $j = 1(tap), j = 2(bottle), j = 3(salt)$

$$\begin{aligned} -15.8 &\leq \mu_{\cdot 1} - \mu_{\cdot 2} \leq 102.5 \\ 55.1 &\leq \mu_{\cdot 1} - \mu_{\cdot 3} \leq 173.4 \\ 11.7 &\leq \mu_{\cdot 2} - \mu_{\cdot 3} \leq 130.0 \end{aligned}$$

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:

$$y_{ijk} = \mu + \alpha_i + \rho_k + \epsilon^w_{ik} + \beta_j + \alpha\beta_{ij} + \epsilon^s_{ijk} \tag{B.4}$$

where

- Y_{ijk} is the observation on yield at the i^{th} fertilizer, k^{th} block, and j^{th} wheat variety
- μ is the grand mean of yields averaged over a population of blocks, all levels of fertilizer, and both wheat varieties.
- α_i is the true effect of the i^{th} level of fertilizer on yield
- ρ_k is the true effect of the k^{th} level of block
- ϵ^w_{ik} 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.
- β_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
- ϵ^s_{ijk} 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.

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)