Basic Statistical Concepts

Population - the collection of all items of interest to a researcher.

Sample - a subset of the population which we gather information on. A common sample is a SRS (Simple Random Sample). A **simple random sample (SRS)** of size *n* consists of *n* individuals from the population chosen in such a way that every set of n individuals has an equal chance to be selected.

Descriptive statistics - summarize information contained in a sample.

Statistical inference - generalize from the sample to the population.

Statistic - a numerical summary of a sample. e.g., the sample mean is a statistic. A statistic is random. Its distribution is called the **sampling distribution**.

Descriptive statistics

Frequency distribution and histogram: summarize quantitative data.

sample mean $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

sample median: the midpoint of the ordered data.

sample variance: $s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$,

sample standard deviation: $s = \sqrt{s^2}$.

Example

Table 2.7 Particulate Emissions for 65 Vehicles

1.50	0.87	1.12	1.25	3.46	1.11	1.12	0.88	1.29	0.94	0.64	1.31	2.49
1.48	1.06	1.11	2.15	0.86	1.81	1.47	1.24	1.63	2.14	6.64	4.04	2.48
1.40	1.37	1.81	1.14	1.63	3.67	0.55	2.67	2.63	3.03	1.23	1.04	1.63
3.12	2.37	2.12	2.68	1.17	3.34	3.79	1.28	2.10	6.55	1.18	3.06	0.48
0.25	0.53	3.36	3.47	2.74	1.88	5.94	4.24	3.52	3.59	3.10	3.33	4.58

Table 2.9 Relative Frequency Distribution for Particulate Data

Class	Frequency	Relative Frequency
0.00–0.99	9	0.138
1.00–1.99	26	0.400
2.00–2.99	11	0.169
3.00–3.99	13	0.200
4.00–4.99	3	0.046
5.00–5.99	1	0.015
6.00–6.99	2	0.031

Histogram

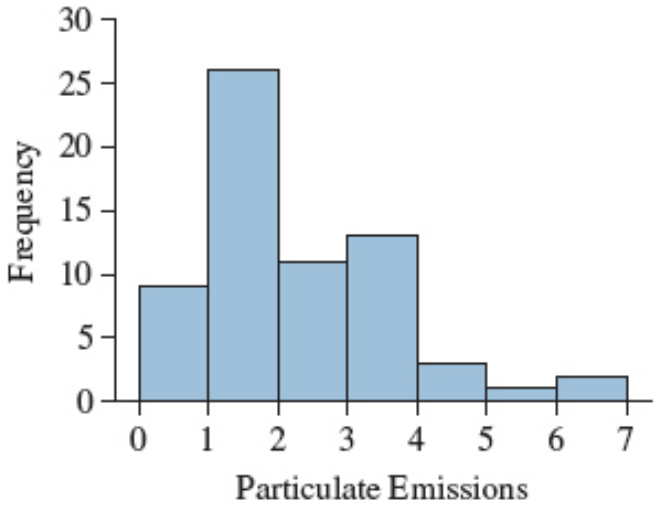

Figure 2.5 Frequency histogram for the frequency distribution in Table 2.10

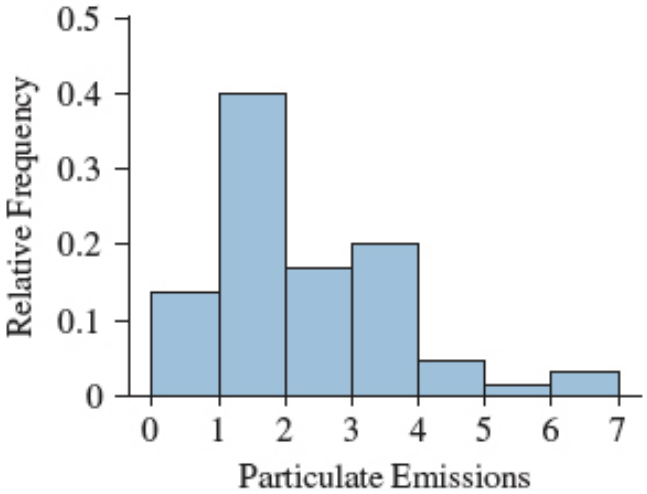

Figure 2.6 Relative frequency histogram for the relative frequency

Experiment and observational study

Experiment: A study in which the conditions are assigned to individuals for the purpose of seeing the effects of these conditions on some characteristic. The assigned conditions are called **treatments**. The characteristic is called the **response**. The individuals are called the **experimental units**.

Observational study: A study in which the conditions are not assigned/controlled but simply observed.
Example: observe the effect of smoking on cancer incidence by observing smokers and nonsmokers.

Variables of an experiment

Response variable: The outcome variable. Mainly quantitative.

Factor: the variable whose effect on the response is of interest. Can be categorical or quantitative. The **treatments** are the different levels of the factor. Some experiments have two (or more) factor under study. **In a two factor study, the treatments are the combinations of the levels of the two factors**.

Extraneous variable: not of main interest but believed to be associated with the response variable.
e.g. In an experiment studying the effect of different fertilizers on the tomato yield (in pounds), extraneous variables include variety of tomatoes, amount of water or sunlight the plant receives and the soil fertility.
The extraneous variables are called **confounding variables(confounders)** when their effects are mixed with the effects of the factor(s) of interest.

An experiment

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 caloriediet, and twenty to a high calorie diet. The number of months that the mice lived was recorded.

This study is an experiment.
Response variable: lifespan of a mouse measured in months.
Factor: diet
Treatments: three treatments-low, medium and high colorie diet.
Experimental units: the 60 mice.

Another experiment

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

The placebo group is a **control group** which often gets a standard or a sham treatment and is uesd as a basis for comparison.

Two-factor experiment

An experiment was conducted to examine the effects of external distractions (none, constant, changing) and type of words (fruit, mixed, nouns) on the ability to memorize words. There are nine (3 x 3) combinations of external distraction and type of word. Thirty-six subjects were assigned at random to the 9 combinations, with four subjects per combination. Each subject studied his/her randomly assigned word list of 30 words under his/her randomly assigned distraction type for a fixed amount of time.

The **response variablle**: the number of words correctly remembered out of 30.

The **experimental units**: the 36 subjects.

The two **factors**: type of external distrctions and type of words

The **treatments**: the 9 combinations of external distraction and type of word: e.g., none distraction and fruit type, constant distraction and mixed type, etc.

Experimental error: variation in values of the response variable for identically treated experimental units. sources: natural variation in experimental units, inability to identically treat the units in the same group, measurement error. **Analysis of Variance** or ANOVA compares the difference between different treatment groups to the experimental error.

Key to a good experiment: “control” the variation resulting from extraneous variables to reduce the experimental error.
e.g. variety of tomato: use the same variety.

It is hard to control the effects of extraneous variables in an observational study. Treatment groups may differ in other ways beside the factor of interest.

Blinding

Blinding the study–the subjects are blind to the treatments to which they are assigned.
Double blinding–neither the subjects or data collectors know the treatment assignment. That is ideal. Often used in medical studies.
The purpose of blinding is to remove the effect of knowledge of the treatment on the response variable.

Principles of Experiment

Randomization: use randomization to assign experimental units to the treatments. Treatment groups should be balanced with regard to extraneous variables.
Blocking: Group experimental units into blocks with similar values on an extraneous variables then randomly assign treatments within each block. Goal: eliminate the effect of the extraneous variable.
Direct **control**: Use experimental units that have the same value on some extraneous variable. That can limit the scope of the conclusion.
Replication: assign a series of independent experimental units to a treatment.

Interested in study the effect of a drug at two dose levels (low, high). 100 men and 100 women available for study.
Gender is a possible confounder.

Direct control: use only women for study.
50 women: low dose
50 women: high dose
Conclusion only appiles to women.

Blocking:
block 1: 100 women. Compare two doses among 100 women.

block 2: 100 men: compare two doses among 100 men.

conclusion: applies to both men and women.

Exercise

One study looked at the perceived benefits of arthroscopic surgery on osteoarthritis by giving some patients a real knee operation while others underwent a sham surgery. Patients were assigned at random to either arthroscopic surgery or a sham surgery. One response variable was the speed of walking after surgery.

a. Is this an experiment or an observational study? Explain.
b. What is the factor? What is the response variable?
c. What is the purpose of one group receiving a sham study?

Solutions

a. Experiment. The two treatments are assigned to different patients.
b. Factor: surgery type. Response variable: speed of walking after surgery.
c. To control for the placebo effect. A placebo effect means often patients respond to any treatment, even a sham treatment. If the surgery really is beneficial then it should perform better than the sham study group.

Exercise

A survey of 232 elderly patients who had recently undergone heart surgerywas 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 orcomfort from religion lived longer than those who said they did not.

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.

Solutinos

a. The factor of interest is whether or not a patient derived strength or comfort from religion. The response variable is the lifespan of a patient.
b. This is an observational study because the two conditions of the factor (derived strength or did not derive strength) were not assigned to the patients.
c. Some confounding variables could include being active or not in social life, finanical status, marital status etc. (e.g., patients who derived strength from relegion may be more socially active and this may contribute to a longer lifespan).