Chapter 3 - Relationships

Chapter 3 Producing Data

Read: Chapter 3 of the text , then read the notes and try the WEBCT assignment questions. If you need more practice, try the practice questions with answers available on the web.

Chapter Three Introduction

Experimental designs for producing data have had a great impact on scientific enquiry. Randomization and sampling have specified how scientists should collect data for valid inference making. The foundations of inference are introduced here using the concept of a sampling distribution for a statistic.

Listed below are a number of different types of statistical investigation. They begin with a low and end with a high level of evidence for making statistically based conclusions.

View the Video-Experimental Design

I) <u>Observational Study</u>

The data is simply collected with no intervention by investigator in how data was produced.

<u>Example</u>: Consider an investigation to study the relationship between high school grades and university grades.

The researcher collects high school and 1st year GPA's from everybody in the class.

The reliability of the conclusions drawn is suspect as conclusions may be highly dependent on the particular data set or class.

However, the data still provide 'evidence' for determining the relationship between high school and university grades.

II) Sampling study

The data are collected via sampling from the population to which conclusions will apply.

However, at the start we must define precisely the **population** = the set of all people, animals, things etc. which we are interested in studying.

Denote the population by π.

Example: Consider a study to investigate the relationship between graduating high school GPA and 1st year GPA for all university students in Ontario over the past 5 years

We define the population as π = set of all university students in Ontario over the past 5 years.

Next define the **variable** of interest.

A **variable** **"X"** is something the researcher is interested in defined on **π.**

We define the following two variables X and Y,

X = graduating high school GPA
Y = 1st year university GPA

A variable X defined on a population has a **distribution**.

We can obtain the distribution of X by conducting a **census**.

A census is when the researcher obtains X, the variable of interest, for all members of the population.

Typically this is not possible. The alternative is to select a subset of the population (**sample**) and then use the empirical (**sample**) distribution of the variable to "estimate" the population distribution.

Randomization

How do you select the sample?

In order to avoid **selection effects** or **bias** researchers use **simple random sampling (SRS).**

A key ingredient of SRS is **randomization.**

Randomization is a process where each sample of size "n" has an equal probability of being selected. Here is one way of obtaining a SRS.

Suppose the population is labelled 1, 2, 3, ... K. Where K is the total # in the population. Put the K chips, labeled 1, 2, 3, ... K, in a bowl.

Mix the chips thoroughly and draw a sample of size "n" **without replacement**.

This process ensures that each subset of n has the same probability of being selected.

SPSS EXAMPLE ONE

Researchers usually use a computer to obtain a SRS. SPSS example one shows how to use the computer to randomly draw a sample. A random number table could also be used.

View the Video – Blocking and Sampling

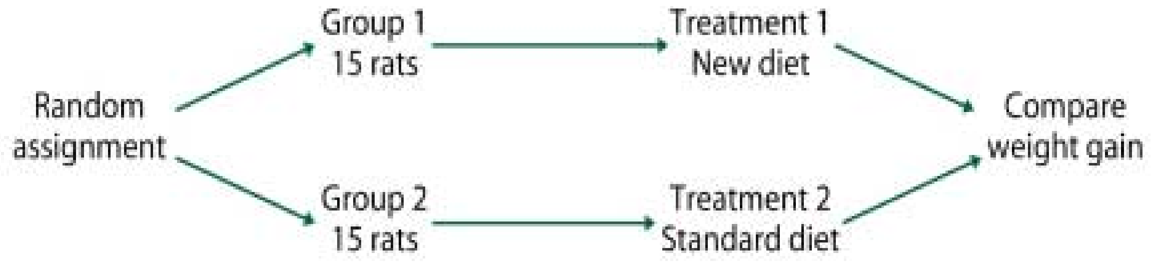


Figure 3.2 Outline of a randomized comparative experiment.

III) Experiment

In an experiment the researcher has defined a population **π** and variables **X** and **Y** defined on **π.** In an experiment we want to determine whether or not changes in X cause changes in Y.

If we only did an observational or sampling study then we could never be sure that changes in other variables were not causing X and Y to change simultaneously.

The key ingredient for an experiment is that the investigator can **assign** values of predictor variable X to the elements of the population.

The fact the investigator can **control** the values of the predictor variable X is crucial for establishing cause and effect.

Experimental design

Experimental design is the process of designing experiments so that 'cause and effect' can be established with a minimal use of resources. For example:

The researcher selects the values of X to investigate in her experiment, say $X_1, X_2, ..., X_{m.}$ The **"m"** values of X are called **treatments**. Next, determine the number of times say n, each treatment will be applied. When each treatment is applied n times we have balance.

Then randomly select **m** x **n** elements from π**.** The elements of the population are called experimental units.

Next, randomly allocate X_1 to n experimental units, X_2 to n experimental units etc. This process of random allocation is called **randomization**.

The researcher obtains the values of Y, our dependent variable, from the experimental units.

We now have samples from each of the conditional distributions of Y given $X = X_1$, $X = X_2$, ..., $X = X_m$ and we can compare these conditional distributions to determine if a cause-effect relationship exists.

Example 1: Our population $\pi =$ students in class. We want to know which is better open or closed book tests.

Let X = 1 open book}
= 0 closed book}

The two values of X are our treatments. Our dependent variable is Y = grade on the test.

Next, randomly select 30 students from class and randomly allocate 15 to X = 1 and 15 to X = 0.

Now, compare the distribution of marks between the 2 groups (open vs. closed book). In other words compare the conditional distribution of Y given X to see if there are any differences.

Example 2: smoking and lung cancer in humans

We can't carry out an experiment (not morally or reasonably possible). Hence, it is very, very difficult to determine cause and effect.

However, there is lots of 'evidence' (animal studies, correlational studies, etc.) however that smoking does cause cancer.

Remember although experiments establish cause and effect, sometimes the amount of evidence is so substantial that we conclude cause and effect without an experiment.

View the video-Samples and Surveys

Sampling Distributions

Suppose π is a population and X is a variable defined on π. Suppose we generate a sample of size n and obtain X_1, X_2, ..., X_n.

Typically we are interested in characteristics of the distribution of X, such as the mean, median, quartiles, standard deviation, etc.

We can estimate these using X_1,...X_n.

Example:
Estimate the population mean of X by the sample mean

$$\overline{\mathbf{X}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum \mathbf{X}_{\boldsymbol{i}}$$

Note the sample characteristics will vary as the sample X_1, X_2… X_n varies. The sample estimates are called **statistics.**

Example: The sample mean "$\bar{x}$" is a statistic. Suppose we take N samples of size "n" from π and for each sample compute $\overline{\mathbf{x}}$ obtaining $\overline{\mathbf{x}_1}, \overline{\mathbf{x}_2}, \ldots \overline{\mathbf{x}_N}$. We can then form the empirical distribution of : $\overline{\mathbf{X}}$.

Fact: As N gets large we obtain the **sampling distribution of the mean** based on a sample of n from π.

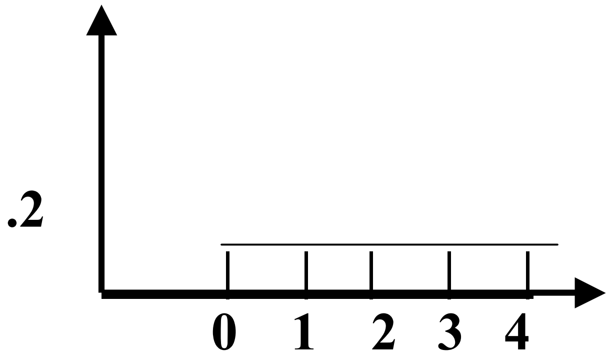
.

Example: Suppose we are studying families and have a population of K=100,000 families and X is a variable denoting the number of children per family taking values {0,1,2,3,4}.

The distribution of X is given in the table below:

X	0	1	2	3	4
P	.2	.2	.2	.2	.2

We have 20% or .2 of families have 4 children, 20% have 3 children, etc. The mean of the population is 2.0 children. Note the population distribution is flat (see below) a very non-normal distribution.



<u>Question</u>:

Describe the sampling distribution of the mean based on samples of size 10? We use SPSS to generate 25 samples of size 10. For each sample calculate the mean number of children per family and plot this value. The short form is:

Generate N=25 samples of size n=10

Click on the SPSS example below to see how this is done and read the output.

SPSS EXAMPLE TWO

Note the mean of the sampling distribution is very close to the population mean (2.0). The normal probability plot indicates that the distribution is approximately normal.

SPSS EXAMPLE THREE

Consider the following example taken from your text. The population has a population proportion p=.6. 1000 samples of size 100 are drawn from this population. For each sample we calculate the sample proportion. A histogram of the 1000 sample proportions is calculated. Note, the mean of the distribution is close to .6 and the overall shape of the sampling distribution is normal. Note that in all of the examples above we start from a population that is very non normal.

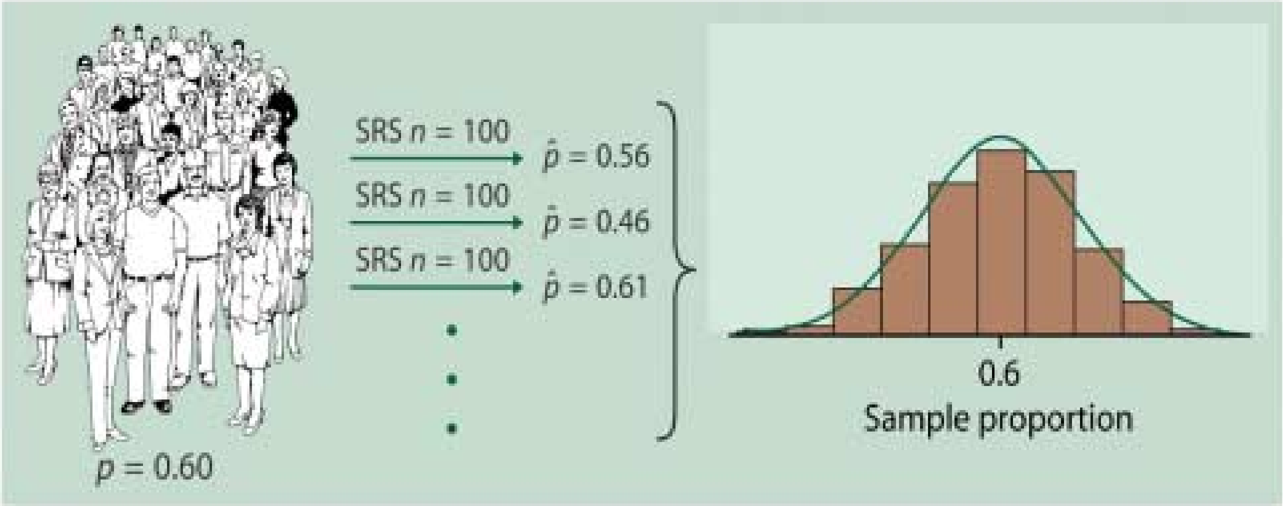


Figure 3.6 The results of many SRSs have a regular pattern. Here, we draw 1000 SRSs of size 100 from the

same population. The population proportion is p = 0.60. The histogram shows the distribution of the 1000 sample proportion p (hat)

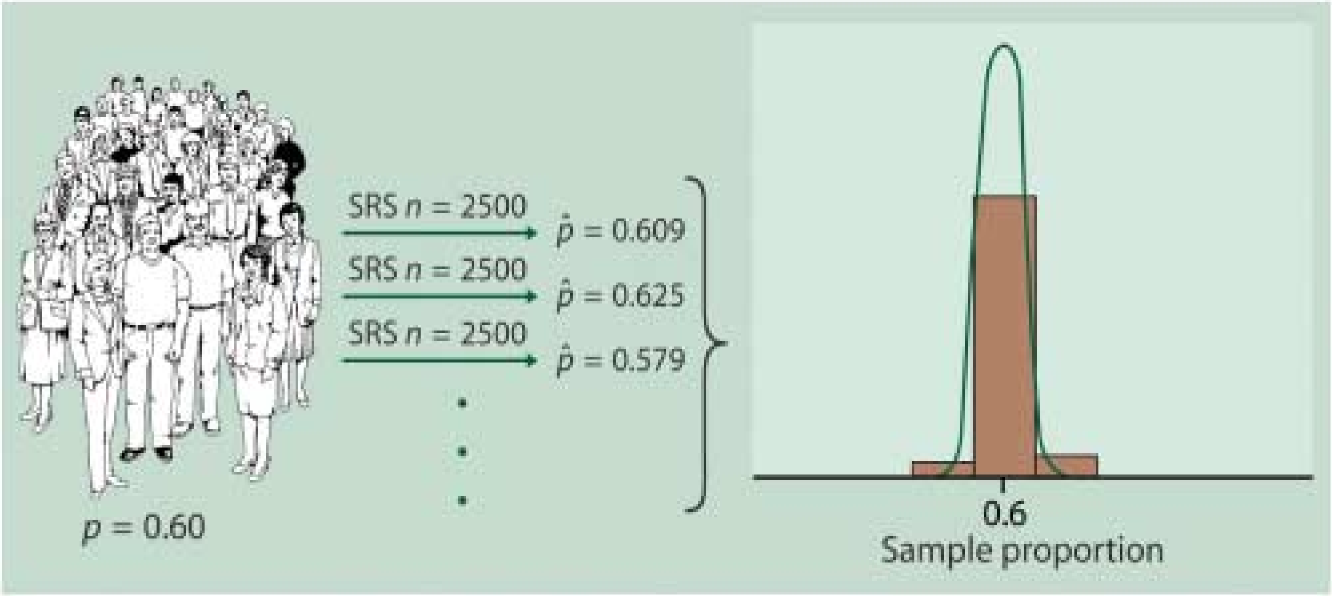


Figure 3.7 The distribution of sample proportions p(hat) for 1000 SRSs of size 2500 drawn from the same population as in Figure 3.6. The two histograms have the same scale. The statistic from the larger sample is less variable or has smaller variance.

When we use the computer to simulate drawing many samples from the population and calculate a mean for each sample, the plot of the means is approximately a normal distribution. This normal distribution, which we call the sampling distribution, has the same mean as the population but with a standard deviation that is smaller than the population. As the sample size increases the standard deviation continues to shrink, in other words, the distribution becomes more concentrated about the mean. Amazing! Look below which graph has the larger sample size and consequent smaller variance?

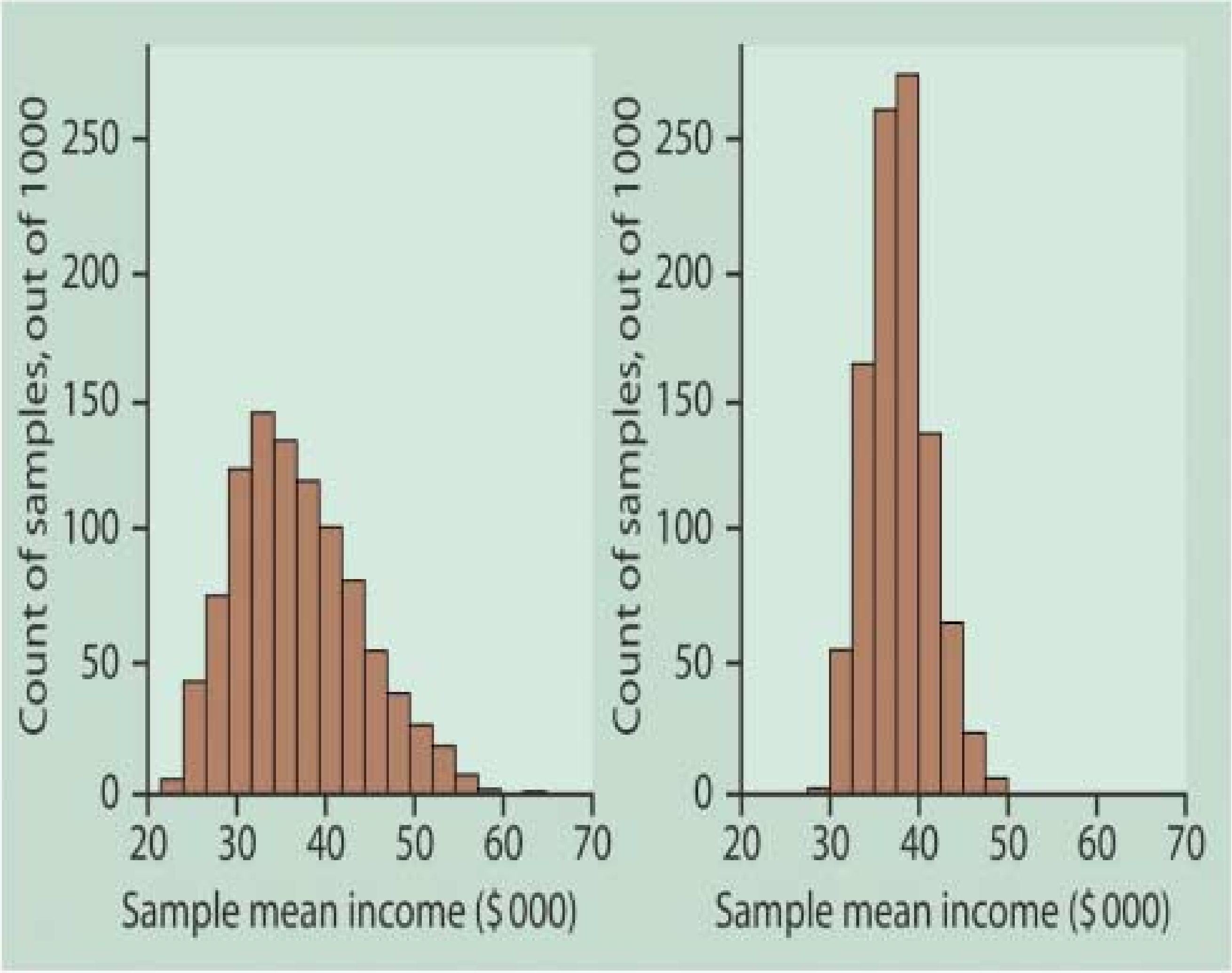


Figure 3.11 The distribution of sample means x(bar) for SRSs of size 25 (left) and of size 100 (right) from the same population.

Chapter 3 Summary

In an experiment treatments are imposed on **experimental units**. Each treatment is a combination of **levels** of the x variables which we call **factors**.
The design of an experiment refers to the choice of treatment using the basic principals of **control**, **randomization** and **replication**. **Randomization** makes an experiment unbiased as well as more sensitive.
A **parameter** refers to an unknown in the population whereas a **statistic** refers to a number computed from the data. A statistic has a **sampling distribution** that shows how the statistic varies in repeated sampling.

Remember: Tips for Success

1) Read the text.
2) Read the notes.
3) Try the assignment.
4) If needed, try the exercise questions.
5) Try the simulations and view the videos if you need more help with a concept.
6) Try the self tests for practice on each chapter of the text at **www.whfreeman.com/ips**
7) Steady Work = SUCCESS !!!