CHAPTER 7: DATA FOR DECISIONS

How can we produce data that can be trusted? How much can we trust the data we find?

7.1 Sampling

The ***population*** in a statistical study is the entire group of individuals about which we want information.

A ***sample*** is a part of the population from which we actually collect information used to draw conclusions about the whole. ***Sampling*** refers to the process of choosing a sample from the population.

7.2 Bad Sampling Methods

A ***convenience sample*** is a sample of individuals who are selected because they are members of a population who are the most convenient to reach. Usually this type of sample cannot be trusted to be representative of the entire population.

A ***voluntary response sample*** consists of people who choose themselves by responding to a general appeal.

The design of a statistical study is ***biased*** if it systematically favors certain outcomes.

EXAMPLE
In a study of the shopping habits of adults, we asked 250 people as they exited a grocery store about their total purchase. What is the population? What is the sample? What kind of sample is it? Is there any possible bias?

EXAMPLE
In order to determine if students on a college campus are in favor of a tuition hike to pay for expanded parking services, a member of the student senate surveys 25 people in a commuter parking lot. What is the population? What is the sample? What kind of sample is it? Is there any possible bias?

EXAMPLE
On October 12, 2011 a quick poll on CNN.com asked "Has the BlackBerry outage affected you?" What is the population? What is the sample? What kind of sample is it? Is there any possible bias?

7.3 Simple Random Samples

A ***simple random sample (SRS)*** of size *n* consists of *n* individuals from the population chosen in such a way that ever set of *n* individuals has an equal chance to be in the sample actually selected.

EXAMPLE
In a class of 25 students, every students name is placed in a box and 5 names are drawn at random. Is this a SRS?

A ***table of random digits*** is a list of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with the following two properties:

1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9
2. The entries in the table are independent of each other.

LINE		RANDOM DIGITS				
101	08705	42934	79257	89138	21506	26797
102	00755	39242	50772	44036	54518	56865
103	35486	59500	20060	89769	54870	75586
104	87788	73717	19287	69954	45917	80026
105	51052	25648	02523	84300	83093	39852
106	88988	12439	73741	30492	19280	41255

To use the table of random digits to generate a SRS, do the following:

1. Give each member of the population a numerical label of the same length.
2. Read from the table strings of digits of the same length as the labels.
3. Skip values that are not in the range.
4. Ignore spaces.

EXAMPLE

Use the random digits table, beginning at line 103, to choose a sample of four people from the following list:

01 Adams	06 Ford	11 Kramer	16 Post
02 Brown	07 Goodman	12 Loomis	17 Quayle
03 Cook	08 Harris	13 Martin	18 Rogers
04 Davis	09 Inez	14 Norton	19 Stevens
05 Elliot	10 Jones	15 O'Hare	20 Thompson

First person: _____ Second Person: _____

Third person: _____ Fourth Person: _____

EXAMPLE

If a class has 100 students, how long do the labels for an SRS need to be?

If there were 405 students, what would the labels look like?

EXAMPLE

A school has 2452 students. Starting at line 120 in the table below, choose a random sample of 4 students.

120		2	7	0	3	1		0	3	8	9	7		1	6	7	3	8		3	1	4	5	3		0	7	5	4	5
121		3	5	1	8	6		0	3	9	5	1		6	8	2	0	8		7	3	4	6	0		7	5	3	1	4
122		3	2	2	7	4		6	7	4	9	2		2	1	6	2	5		3	0	2	9	8		1	5	8	5	5
123		9	7	8	8	6		3	1	4	8	0		9	6	6	1	1		3	9	0	3	1		3	1	5	2	5
124		4	0	1	3	5		2	2	6	0	9		7	1	8	7	5		7	3	4	3	3		1	2	8	3	8
125		8	7	5	3	8		7	4	6	3	3		4	0	0	0	2		7	4	4	7	9		8	8	1	1	3
126		5	1	3	4	9		3	9	8	8	5		2	9	9	9	5		3	7	8	5	8		1	8	3	1	3
127		7	0	7	1	8		4	0	9	4	1		2	8	7	0	6		7	5	5	1	0		0	5	8	3	2
128		9	0	2	3	4		7	4	9	8	3		3	7	7	3	2		3	7	0	2	4		4	1	7	1	8
129		0	0	9	6	2		9	3	9	5	8		4	6	9	8	5		9	4	9	8	9		3	0	2	2	1
130		2	7	2	1	9		6	7	2	6	0		8	2	7	4	0		1	8	9	4	6		2	9	1	7	0

Students picked: _____

7.4 Cautions About Sample Surveys

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.

EXAMPLE
A political survey is done via telephone by calling land lines between 6P and 8P. Is this a SRS? Why or why not? What kind of bias may be present?

EXAMPLE
A survey of dining services is done at lunchtime at a busy cafeteria. Is this an SRS? Why or why not? What kind of bias may be present?

EXAMPLE
A long questionnaire is sent out to owners of a certain brand of new cars. Is this a SRS? Why or why not? What kind of bias may be present?

EXAMPLE
From http://www.sciencenews.org/sn_arc99/3_6_99/bob1.htm

For the year 2000 enumeration, the Bureau of the Census plans to mail out more than 90 million questionnaires and deliver millions more by hand to households across the nation before Census Day, April 1. An army of census takers will next fan out across the country aiming to track down the significant number of people who fail to return forms.

Past experience suggests that, despite such a huge effort, the population count will still be incomplete. In 1990, the Census Bureau officially recorded 248,709,873 people. Evidence from other surveys and demographic analyses indicated that the population was closer to 253 million and those not counted were mostly children, people from racial and ethnic minorities, and poor residents of both rural and urban areas.

"Coming out of the 1990 census, we recognized that you can't count everyone by direct enumeration," says Barbara Everitt Bryant, Census Bureau director during that head count.

In January, however, the Supreme Court ruled against the use of statistical sampling methods to obtain population figures for determining how many of the 435 seats in the House of Representatives go to each of the 50 states. At the same time, the court upheld a law that mandates the use of statistically adjusted figures, when feasible, for all other purposes, such as distributing $180 billion in funds for federal programs and determining congressional and state district boundaries.

7.5 Experiments

Observe and describe are not effective ways to determine if a response variable is really responding to an explanatory variable. To determine if correlation really is causation an experiment is needed.

An ***experiment*** deliberately imposes a treatment on individuals in order to observe their responses. The purpose of an experiment is to study whether the treatment *causes* a change in the response.

EXAMPLE
Students in a college math class are allowed to choose if they would like to attend a traditional lecture class or do a self-paced online class. At the end of the semester the final exam grades are compared and it is found that the average in the traditional lecture group was higher than the self-paced group. Can you say that the traditional lecture is more effective than the self-paced method?

EXAMPLE
Students in a dorm were offered free vitamins one semester and the number of sick days of all the students was tracked. At the end of the semester, the average number of sick days for the students who took the free vitamins was lower than those who did not take the vitamins. Can we conclude that the vitamins keep the students from getting sick?

Variables, whether intentionally part of a study or not, are said to ***confounded*** when their effects on the outcome cannot be distinguished from each other.

How can we deal with confounded variables? Use a ***control group*** that does not receive the treatment.

Watch out for the placebo effect! The ***placebo effect*** is the effect of a dummy treatment on the response of the subjects.

In a ***double-blind*** experiment neither the experimental subjects nor the observers know which treatment the subjects are given.

EXAMPLE
Design an experiment in to test if traditional lecture or self-paced is better for a particular math class.

EXAMPLE
Design an experiment to see if vitamins decrease the number of sick days for students living in a dorm.

Note that the results can vary due to different subjects that are chosen when the experiments are done.

An observed effect so large that it would rarely (less than 5% of the time) occur by chance is called ***statistically significant***.

7.6 Experiments Versus Observational Studies

An ***observational study*** is a passive study of a variable of interest. The study *does not attempt to influence* the responses and is meant to describe a group or situation.

A ***prospective study*** is an observational study that records slowly developing effects of a group of subjects over a long period of time.

A ***retrospective study*** is an observational study that uses interviews or records to collect information about past behaviors in two or more groups.

A ***controlled study*** is a study that has a control group.

An ***uncontrolled study*** is a study that lacks a control group.

EXAMPLE
A 10-year study of low-birth-weight babies is performed to determine if birth weight affects IQ and performance in elementary school. Children are identified in hospitals at birth and their performance is tracked until they are 10 years old. This type of study is

A group of 100 students is randomly chosen and divided into two groups. One group is taught typing using a set of new materials and the other using traditional methods. After instruction, typing speeds are compared to determine if the new materials improve learning. This type of study is

7.7 Inference: From Sample to Population

Statistical inference refers to methods used for drawing conclusions about an entire population on the basis of data from a sample. A ***confidence interval*** is one type of inference method.

Statistical inference will only be valid if the data is from a random sample or a randomized comparative experiment.

A ***parameter*** is a fixed (and usually unknown) number that describes a population.

A ***statistic*** is a number that describes a sample.

If the parameter for the proportion of successes is called p, then the corresponding statistic for the proportion of successes is called $\hat{p}$.

EXAMPLE
A survey is sent to 100 employees at a community hospital asking if they support a law requiring motorcycle riders to wear helmets. The results indicate 88% support the law. If the actual proportion of the community's residents who support the law is 72%, the difference is most likely a result of what? What is p and what is $\hat{p}$?

The ***sampling distribution*** of a statistic is the distribution of values taken on by the statistic in all possible samples of the same size from the same population.

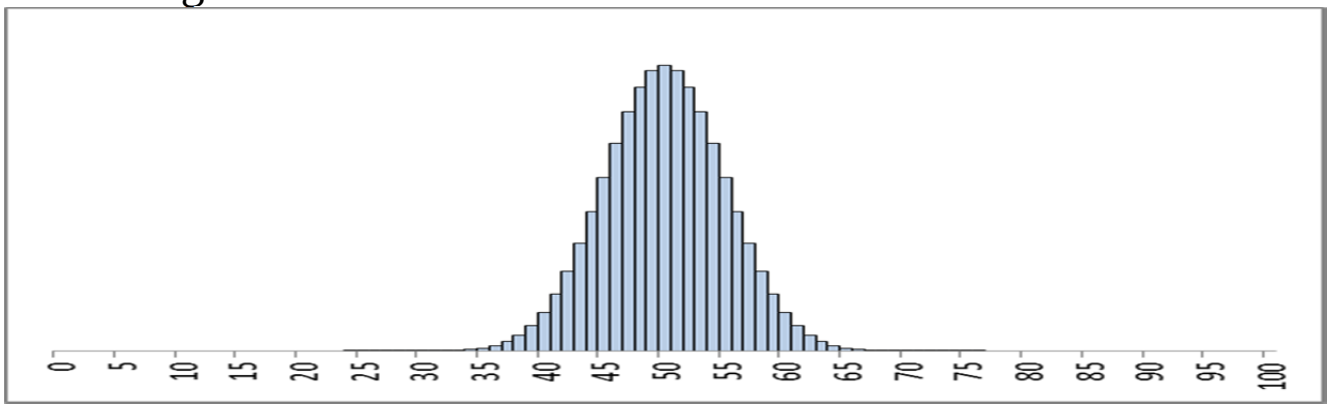
EXAMPLE

We have a population that has a 0.5 chance of voting for the X party. One hundred (n = 100) voters were surveyed at random to ask if they would vote for the X party. The results of the first 10 surveys were

0.44, 0.47, 0.51, 0.46, 0.51, 0.61, 0.57, 0.48, 0.57, 0.42

who would vote for the X party. Display this information in a histogram. Find the mean and standard deviation for the proportion of people who would vote for the X party.

If this experiment was repeated many times, the results would look something like



If it was repeated even more times, the curve would have an even smaller spread. This leads us to a theorem about sampling distributions:

Choose a SRS of size *n* from a large population that contains population proportion *p* of successes. Let $\hat{p}$ be the sample proportion of successes.

- The shape of the sampling distribution of $\hat{p}$ will be approximately normal if *n* is 30 or more.
- The mean of the sampling distribution of $\hat{p}$ is *p*.
- The spread in the sampling distribution of $\hat{p}$ is

$$\sqrt{\frac{p(1-p)}{n}}$$

For a population with $p = 0.5$, the theorem predicts $\hat{p} = 0.5$ for samples.

If n=100, the theorem predicts a spread in our sampling distribution of

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{100}} = \sqrt{\frac{.25}{100}} = \sqrt{0.0025} = 0.05$$

EXAMPLE

A population has a probability $p = 0.25$ of smoking. A random sample of 200 people was asked if they smoked or not. What would you expect the results of the sample to look like if this experiment was repeated many times? If one sample returned the result that no one smoked, would you believe it?

7.8 Confidence Intervals

EXAMPLE
We have a population that has a 0.5 chance of voting for the X party. One hundred voters were surveyed at random to ask if they would vote for the X party. In what range should 95% of our results fall?

EXAMPLE
A population has a probability $p = 0.25$ of smoking. A random sample of 200 people was asked if they smoked or not. In what range should 95% of our results fall?

In almost all large random samples, the parameter p will be close to the statistic $\hat{p}$, so we will use $\hat{p}$ as the guessed value for the unknown p.

A ***95% confidence interval*** is an interval obtained from the sample data by a method in which 95% of our samples will produce an interval containing the true population parameter.

Chose an SRS of size n from a large population that contains an unknown proportion p of successes.

The 95% confidence interval for p is approximately $\hat{p} \pm 2\sqrt{\frac{p(1-p)}{n}}$

The term $2\sqrt{\frac{p(1-p)}{n}}$ is known as the ***margin of error***.

EXAMPLE

If the confidence interval is determined to be from 42% to 46%, what is the margin of error?

EXAMPLE

A national poll asked 2486 adults whether they were satisfied with their jobs, and 45% said they were. Estimate a 95% confidence interval for the actual percentage among all adults.

EXAMPLE

A survey will be done to determine if young adults read novels. Assuming that that half of the young adults do read novels, determine the minimum number of people to be surveyed so that the margin of error will be 5%.