1. Sampling

Definition 1.1.

The _____in a statistical study is the entire group of individuals about which we want information.
A _____is a part of the population from which we actually collect information used to draw conclusions about the whole.
_____refers to the process of choosing a sample from the population.

2. Bad Sampling Methods

Definition 2.1.

The design of a statistical study is _____if it systematically favors certain outcomes.
A_____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 _____sample consists of people who choose themselves by responding to a general appeal. People with strong opinions are more likely to respond and will be overrepresented in the sample.

Example 2.2.

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?

- Population:
- Sample:
- Type of sample:
- Bias?

Example 2.3.

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?

- Population:
- Sample:
- Type of sample:
- Bias?

Example 2.4.

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?

- Population:
- Sample:
- Type of sample:
- Bias?

3. Simple Random Samples

Definition 3.1.

A _____(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 in the sample actually selected.

Example 3.2.

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

Definition 3.3.

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.

Algorithm 3.4 (Generate SRS by using table of random digits)**.**

(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 and values that have already been used.
(4) Ignore spaces.
(5) Option used in this class: If you reach the end of a line without enough digits for your label, ignore the end of that row and continue with the next line.

LINE	RANDOM DIGITS
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

Example 3.5.

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 3.6.

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

If there were 1405 students, what would the labels look like?

A school has 2452 students. Starting at line 122 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:

4. Cautions About Sample Surveys

Definition 4.1.

A _____occurs when some groups in the population are left out of the process of choosing the sample.

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

(Reminder: A 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 in the sample actually selected.)

Example 4.2.

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

Example 4.3.

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

Example 4.4.

All new owners of a certain brand of car that were bought at a particular dealership in town during the past month had their names put into a bowl. At the end of the month, the dealership selected 50 names and sent each a long questionnaire concerning their new car. Is this a SRS? Why or why not? What kind of bias may be present?

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.

Definition 5.1.
An experiment deliberately _____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 by controlling other influences.

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

How can we deal with confounded variables?
Use a _____that does not receive the treatment.

A _____experiment uses a control group. If this group is picked at random, you have a randomized comparative experiment. Otherwise, you just have a comparative experiment.

An _____experiment lacks a control group.

Example 5.2.
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 5.3.
Design an experiment to test if traditional lecture or self-paced is better for a particular math class.

Watch out for the _____! It is the effect of a dummy treatment on the response of the subjects. This is often in the form of a "sugar pill".

In a _____experiment, neither the experimental subjects nor the observers know which treatment the subjects are given.

Example 5.4.
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 students from getting sick?

Example 5.5.
Design an experiment to see if vitamins decrease the number of sick days for students living in a dorm.

Even if you are using an SRS, results can vary due to different subjects that are chosen. One SRS may over-represent some group just by chance.
An observed effect so large that it would rarely (less than 5 percent of the time) occur by chance is called _____.

6. Experiments Versus Observational Studies

Definition 6.1.

An _____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 _____study is an observational study that records slowly developing effects of a group of subjects over a long period of time

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

Example 6.2.

A 10-year look at 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.

- Is this an experiment or study?
- If an experiment, is it controlled or uncontrolled?
- If it is a study, is it prospective or retrospective?

Example 6.3.

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.

- Is this an experiment or study?
- If an experiment, is it controlled or uncontrolled?
- If it is a study, is it prospective or retrospective?

7. Inference: From Sample to Population

Definition 7.1.

_____refers to methods used for drawing conclusions about an entire population on the basis of data from a sample. This is one of the main uses of statistics. Statistical inference will only be valid if the data is from a random sample or a randomized comparative experiment.

A _____is a fixed (and usually unknown) number that describes a population.

A _____is a number that describes a sample. It is known when we have taken a sample, but it can change from sample to sample. It is often used to estimate an unknown parameter.
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 7.2.
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 percent support the law. If the actual proportion of the community's residents who support the law is 72 percent, what is p and what is $\hat{p}$?

Is the difference in these values most likely a result of random chance or sampling bias?

Definition 7.3.
The _____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.

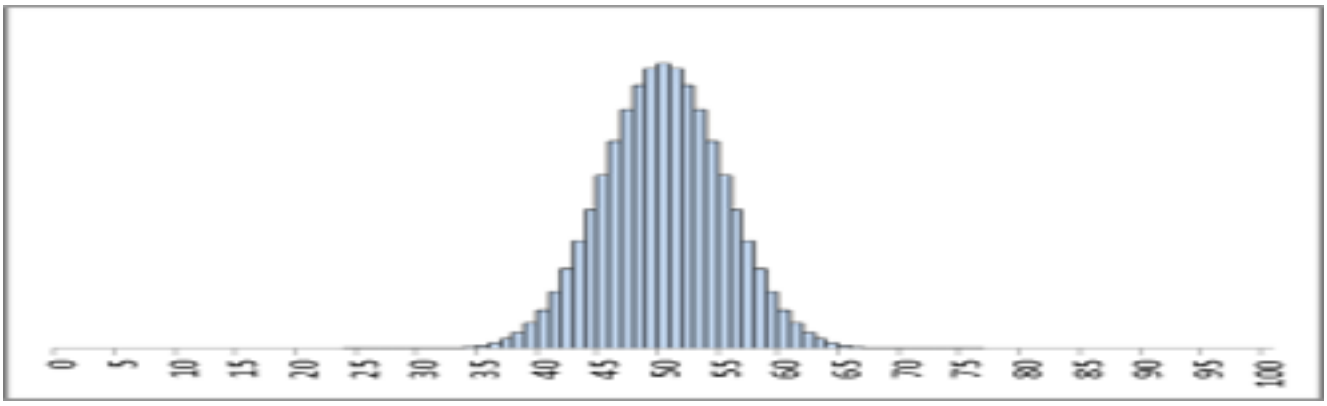
Example 7.4.
We have a population that has a 50 percent chance of voting for party X. One hundred ($n = 100$) voters were surveyed at random to ask if they would vote for party X. The results were that 44/100 would vote for party X. Of course, party X didn't like the results. Another survey of one hundred voters was taken. This time, 57/100 would vote for party X. These results are quite varied and were repeated 8 more times with other groups of one hundred voters. The results of all 10 surveys (one hundred people were in each survey) were proportions

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

who would vote for party X.

The mean of these samples tells us that the proportion of people who would vote for party X (based on these samples) is 0.504. The standard deviation of the samples is 0.062218.

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:

Theorem 7.5 (Sampling Distribution of a Sample Proportion)**.**
Choose an SRS of size n from a large population that contains population proportion p of successes. Let $\hat{p}$ be the **sample proportion of successes**, *expressed as $\hat{p} = \frac{\text{count of successes in the sample}}{n}$ Then:*

- *Shape: For large sample sizes ($n \geq 30$), the sampling distribution of $\hat{p}$ is approximately normal.*
- *Center: The mean of the sampling distribution of $\hat{p}$ is p.*
- *Variability: The standard deviation of the sampling distribution of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$*

Example 7.6.
For a population with $p = 0.4$, if we were to take samples of $n = 100$ over and over, the mean of this sampling distribution would be $\hat{p} = 0.4$.
(Note: Not every sample will have a proportion of 0.4, but on average, the proportion will be 0.4.)
If $p = 0.4$ and $n = 100$, the theorem predicts a spread in our sampling distribution of

$$\sigma = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{100}} = \sqrt{\frac{0.24}{100}} = \sqrt{0.0024} \approx 0.049$$

Example 7.7.
A population has 25 prcent who are smokers. 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?

8. Confidence Intervals

Unfortunately, we don't always know the actual value of p for the population. If we take a sample and find a statistic, how sure can we be that this statistic is "close" to the actual population parameter? We can construct confidence intervals using the method below.

We can construct a 95% percent confidence interval using the formulas given below. If we took lots of samples and constructed 95% confidence intervals around them, 95% of those confidence intervals would contain the true population mean (so 5% of those confidence intervals would NOT contain the true population mean).

Theorem 8.1. *Choose an SRS of size n from a large population that contains an unknown proportion p of successes. A* 95% *confidence interval for p is approximately*

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The $\pm$ *sign is read "plus or minus," so, for example,* 0.5 ± 0.2 *yields two numbers:* $0.5 - 0.2 = 0.3$ *and* $0.5 + 0.2 = 0.7$. *This can be written as an interval: (0.3,0.7)*

Definition 8.2.

In the above formula, **he margin of error** is the term $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Example 8.3.

Suppose a study was done of a random sample of 500 voters and 55% said they would vote for party X. Construct a 95% confidence interval for the actual percentage of adult voters who would vote for party X. What is the margin of error?

Example 8.4.

A random sample of 200 people found that 25% of them smoke. Construct a 95% confidence interval for the actual percentage of adults who smoke. What is the margin of error?

Example 8.5.
If the 95% confidence interval for some parameter is determined to be from 42% to 46%, what is the margin of error?

Example 8.6.
What is the formula for approximating a 99.7% confidence interval for p?

Example 8.7.
A national poll asked 2500 adults whether they were satisfied with their jobs. 1040 of them said they were. Construct a 99.7% confidence interval for the actual percentage among all adults who are satisfied with their job. What is the margin of error?

Example 8.8.
A survey will be done to determine if young adults read novels. If 60% of all young adults do read novels, determine the minimum number of people to be surveyed so that the margin of error will be 3%, assuming you want a 95% confidence interval.