

## **Thursday, August 6: Chapter 12: Sample Surveys**

An Introductory Activity: The Sexual Discrimination Problem: Recently, a large company had to downsize and fire 10 employees. Of these 10 employees, 5 were women. However, only 10 of the 30 employees at the company were women. This discrepancy has led the women who were fired to file a sexual discrimination lawsuit. Do they have a legitimate claim?

Def: \_\_\_\_\_ is the science of collecting, analyzing, and drawing conclusions from data.

Def: The \_\_\_\_\_ is the entire collection of individuals or objects about which information is desired.

Def: When you study an entire population, it is called a \_\_\_\_\_.

Def: A \_\_\_\_\_ is a subset of the population, selected for study in some prescribed manner.

Def: A \_\_\_\_\_ is a list of individuals from which the sample is drawn.

Def: A \_\_\_\_\_ is a characteristic of an entire population, such as the average height of *all* CDO students or the proportion of *all* US citizens who approve of Barack Obama. Unfortunately, we cannot know the value of a parameter without taking a census.

Def: A \_\_\_\_\_ is an estimate of a parameter based on a sample from the population. For example, based on a sample of 50 CDO students, we *estimate* the true average height is 67.1 inches.

One application of statistics is to determine the “readability” of various books and articles. One simple way to do this is to estimate the average word length. Consider, the Gettysburg Address by Abraham Lincoln, one of the most famous speeches of all time:

### **Lincoln's Gettysburg Address**

***Directions: Use 5 words of your choice to estimate the average length of a word in the speech below.***

Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation or any nation so conceived and so dedicated can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead who struggled here have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us the living rather to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us--that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion--that we here highly resolve that these dead shall not have died in vain, that this nation under God shall have a new birth of freedom, and that government of the people, by the people, for the people shall not perish from the earth.

1	Four	55	We	109	cannot	163	for	217	they
2	score	56	are	110	dedicate,	164	us	218	gave
3	and	57	met	111	we	165	the	219	the
4	seven	58	on	112	cannot	166	living,	220	last
5	years	59	a	113	consecrate,	167	rather,	221	full
6	ago,	60	great	114	we	168	to	222	measure
7	our	61	battlefield	115	cannot	169	be	223	of
8	fathers	62	of	116	hallow	170	dedicated	224	devotion,
9	brought	63	that	117	this	171	here	225	that
10	forth	64	war.	118	ground.	172	to	226	we
11	upon	65	We	119	The	173	the	227	here
12	this	66	have	120	brave	174	unfinished	228	highly
13	continent	67	come	121	men,	175	work	229	resolve
14	a	68	to	122	living	176	which	230	that
15	new	69	dedicate	123	and	177	they	231	these
16	nation:	70	a	124	dead,	178	who	232	dead
17	conceived	71	portion	125	who	179	fought	233	shall
18	in	72	of	126	struggled	180	here	234	not
19	liberty,	73	that	127	here	181	have	235	have
20	and	74	field	128	have	182	thus	236	died
21	dedicated	75	as	129	consecrated	183	far	237	in
22	to	76	a	130	it,	184	so	238	vain,
23	the	77	final	131	far	185	nobly	239	that
24	proposition	78	resting	132	above	186	advanced.	240	this
25	that	79	place	133	our	187	It	241	nation,
26	all	80	for	134	poor	188	is	242	under
27	men	81	those	135	power	189	rather	243	God,
28	are	82	who	136	to	190	for	244	shall
29	created	83	here	137	add	191	us	245	have
30	equal.	84	gave	138	or	192	to	246	a
31	Now	85	their	139	detract.	193	be	247	new
32	we	86	lives	140	The	194	here	248	birth
33	are	87	that	141	world	195	dedicated	249	of
34	engaged	88	that	142	will	196	to	250	freedom,
35	in	89	nation	143	little	197	the	251	and
36	a	90	might	144	note,	198	great	252	that
37	great	91	live.	145	nor	199	task	253	government
38	civil	92	It	146	long	200	remaining	254	of
39	war,	93	is	147	remember,	201	before	255	the
40	testing	94	altogether	148	what	202	us,	256	people,
41	whether	95	fitting	149	we	203	that	257	by
42	that	96	and	150	say	204	from	258	the
43	nation,	97	proper	151	here,	205	these	259	people,
44	or	98	that	152	but	206	honored	260	for
45	any	99	we	153	it	207	dead	261	the
46	nation	100	should	154	can	208	we	262	people,
47	so	101	do	155	never	209	take	263	shall
48	conceived	102	this.	156	forget	210	increased	264	not
49	and	103	But,	157	what	211	devotion	265	perish
50	so	104	in	158	they	212	to	266	from
51	dedicated,	105	a	159	did	213	that	267	the
52	can	106	larger	160	here.	214	cause	268	earth.
53	long	107	sense,	161	It	215	for		
54	endure.	108	we	162	is	216	which		

Which method is better? Why?

In this activity, what was the population parameter? What was the sample statistic?

When a statistician is using a sample to estimate something about a population, there are 2 potential problems.

Def: \_\_\_\_\_ occurs when our estimates are consistently too high or consistently too low. Bias can be a major problem when conducting a sample survey. To eliminate selection bias, we need to let chance do the choosing! When we chose which words to use, our eyes were drawn to the larger words and our samples were therefore biased.

Def: The \_\_\_\_\_ of an estimate refers to the range of values that the estimate can take in repeated sampling. Even when we all used an unbiased method for choosing the sample, there were many different estimates. Obviously, it would be better if we could all get the same correct answer!

How can we reduce the amount of variability in the estimates of the mean word length?

### Bias vs. Variability:

Suppose we are trying to estimate a population mean ( $\mu$ ) by taking a sample and finding the sample mean ( $\bar{x}$ ). Describe the following distributions of  $\bar{x}$ :

Note: The size of the \_\_\_\_\_ has little effect on the variability of an estimate, especially when the sample is a small proportion of the population.

**HW #1: Get info sheet signed, Suggested Reading (270-275), Problems page 291 (25, 26, 30)**

## **Monday, August 10: Chapter 12: Bias in Surveys**

Def: \_\_\_\_\_ (often called undercoverage bias) is introduced when some part of the population is systematically underrepresented in the sample.

Selection bias also occurs when volunteers self-select themselves for a sample. People who voluntarily respond to surveys tend to have different and stronger opinions than the rest of the population. This is often called \_\_\_\_\_.

In all sampling procedures, it is very important that every member of the population be given an equal chance to be chosen for the sample! Random sampling is the best way to make sure this happens.

It is important to carefully compare the population of interest (the population we want to know about) with the population of inference (the population we can actually make conclusions about based on the subjects chosen in the sample). If there is selection bias, then these populations are not the same!

For example, suppose we wanted to estimate the average GPA of students at CDO. If we took a random sample of students on campus from 7:00-8:00 am, describe the population of interest and population of inference. Will our estimate be biased?

Def: \_\_\_\_\_ (or measurement bias) occurs when our method of collecting the data tends to produce values that systematically differ from the true population value in some way.

ex: using a faulty instrument:

ex: wording of questions:

ex: characteristics of the interviewer:

ex: human nature:

ex: order of questions:

Def: \_\_\_\_\_ occurs when responses are not actually obtained from subjects chosen for the sample.

Very few surveys, if any, have a 100% response rate, but every effort should be made to make this rate as high as possible. Personal interviews have a better response rate, but are more costly than mail or phone surveys. In all three methods, it is important to follow up on subjects who do not respond the first time rather than sample more people.

Note: Increasing the sample size is usually a good idea, but if there is bias present, even a very large sample will probably be worthless.

Read article: *Polling's Dirty Little Secret*

**HW #2: Flip internet poll, Suggested Reading (282-285), Problems page 290 (16, 18, 20, 21)**

*Internet Poll Idea from Kyrie Martin.*

## **Tuesday, August 11: Chapter 12: Sampling Methods**

As we discovered with the Gettysburg Address, it is very important to \_\_\_\_\_ members of the sample to avoid selection bias. There are many random sampling procedures, the most basic being a simple random sample.

Def: A \_\_\_\_\_ (SRS) of size  $n$  is a sample from the population that is selected in a way that ensures that every member of the population has an equal chance of being selected \_\_\_\_\_. Every sample of size  $n$  has the same chance of being chosen.

For example, to select a SRS of size 4 from this class, we could write each name on a slip of paper, mix them up, and select 4 names. In this way, each member of the population has the same chance of being chosen, as does each possible group of size 4.

Or, I could use my roll sheet as a sampling frame. To choose a SRS, I would assign each member a number, and then use random number generator to select the sample.

Note: when choosing a sample in this way, occasionally the same number will be selected twice. However, in most cases, statisticians do not want to use the same person more than once. This is called \_\_\_\_\_ because after a person is selected, he is not replaced in the sampling frame.

What are some advantages to using a SRS?

What are some disadvantages to using a SRS?

Suppose that a class is half boys and half girls. To get a sample of size 4 from this class, we could write the name of each boy on a slip of paper, mix them up, and select 2. Do the same for the girls. Why isn't this a SRS?

Def: \_\_\_\_\_ is a method of random sampling which seeks to reduce the variability of a SRS by selecting a random sample from each subgroup of the population. This guarantees that each subgroup, or stratum, is properly represented in the overall sample.

Note: To be most effective, the members of each stratum should be as similar as possible with regard to the question of interest and very different than the members of the other strata.

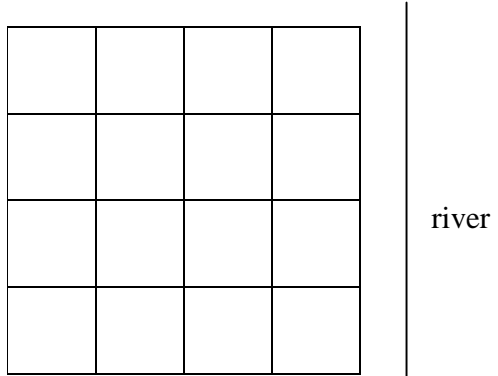
Suppose we wanted to get a stratified random sample of CDO to answer a question about assemblies. Since freshman may have different views than seniors, sophomores have different views than juniors, etc., we want to make sure each group is properly represented in our sample.

Suppose there are 600 freshman, 500 sophomores, 500 juniors, and 400 seniors. If we wanted to take a stratified random sample of size 100, how many of each class should be included?

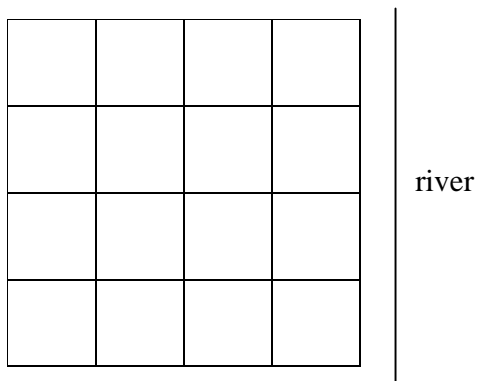
Once we determine the number of subjects to select from each stratum, we take a SRS within each stratum.

Suppose we wanted to estimate the yield of our corn field. The field is square and divided into 16 equally sized plots (4 rows x 4 columns). A river runs along the eastern edge of the field. We want to take a sample of 4 plots.

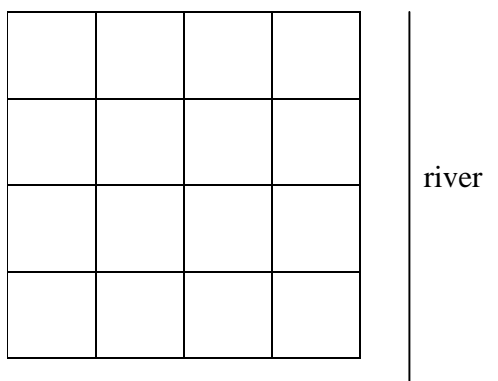
Using a random number generator, pick a simple random sample (SRS) of 4 plots. Place an X in the 4 plots that you choose.



Now, randomly choose one plot from each horizontal row. This is called a stratified random sample.



Finally, randomly choose one plot from each vertical column. This is also a stratified random sample.



Which of the 3 methods above do you think will be the most effective? Why?

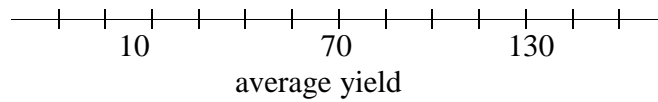
Now, its time for the harvest! The numbers below are the yield for each of the 16 plots. For each of your three samples above, calculate the average yield.

4	29	94	150
7	31	98	153
6	27	92	148
5	32	97	147

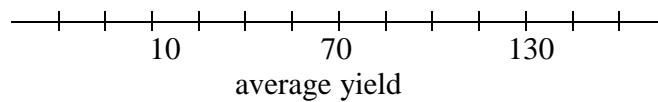
river

### Graphing the results:

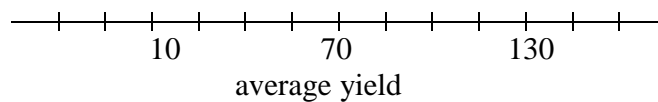
Simple Random Sample:



Stratified by Row:



Stratified by Column:



What happened?

When we stratified by columns, the \_\_\_\_\_ of our estimate was greatly reduced. Also note that each plot was centered in the same place, suggesting all 3 methods are \_\_\_\_\_.

Why does this work?

With a SRS, it is possible that I randomly choose 4 plots near the river (giving an estimate that is way too high) or that I choose 4 plots far from the river (giving an estimate that is way too small). However, when I use each column as a stratum, I am guaranteed to get one plot close to the river (high yield), one plot far from the river (low yield), etc. This guarantees that we will have a representative sample with respect to the river.

When should we stratify?

If you think there are groups within the population who may be \_\_\_\_\_ with regard to the question of interest, you should take an appropriately sized simple random sample from each group.

In our example, we should anticipate that the river will have an effect on the yield of the plots. Thus, since the plots near the river are similar to each other (but different than the rest of the plots) stratifying by columns is the best method.

ex: population: United States adults                      topic: affirmative action

possible strata:    non-effective strata:

ex: population: CDO    topic: Pride Time

possible strata:    non-effective strata:

Note: The reason why we stratify is to get a representative sample and reduce the variability that is possible in a SRS. The purpose is NOT to compare the results between strata, although this is a secondary benefit.

What are the advantages to this method?

- It helps to ensure that the sample is representative of the various subgroups within the population. No group will be over- or under-represented.
- If strata are chosen correctly, stratifying reduces the variability that is possible in a SRS of the same size. Thus, we can either keep the sample size the same and have more precision OR keep the same precision and reduce the sample size (and costs).

What are the disadvantages to this method?

- We need a sampling frame which includes the entire population as well as characteristics about each member to use when stratifying. This could be difficult when the population is large.
- The statistical analysis is more difficult with a stratified random sample.
- In some cases it is difficult to obtain stratified random samples since information about the population needs to be known in advance. For example, if random digit dialing is used, researchers will often start by asking demographic questions to determine which stratum a subject belongs to. Then, they will keep calling until there are enough subjects in each strata.

**HW #3: Suggested Reading (276-278) Problems page 291 (23, 24, 27, 28)**

## **Thursday, August 13: Chapter 12: Other Sampling Methods**

\_\_\_\_\_ : Sometimes it is easier to select groups from a population than it is to select individuals themselves. For example if we wanted to survey approximately 100 freshmen, we could randomly select 3 freshmen English classes and use all the members of those classes. This is much more efficient than a simple random sample since all of the people selected will be together in the same place.

Cluster sampling involves dividing the population of interest into non-overlapping subgroups, called \_\_\_\_\_. Clusters are then selected at random, and all individuals in the selected clusters are included in the sample.

Since whole clusters are selected, the ideal situation occurs when each cluster mirrors the characteristics of the \_\_\_\_\_. However, since this is rarely the case, it is wise to choose as many clusters as you can afford.

Be careful not to confuse clustering and stratification. Even though both involve dividing the population up into subgroups, both the way in which the subgroups are sampled and the optimal strategy for creating the subgroups are different. In \_\_\_\_\_ sampling, we sample from every stratum, whereas in \_\_\_\_\_ sampling, only selected whole clusters are included in the sample. Because of this difference, to increase the chance of obtaining a sample that is representative of the population, we want to create \_\_\_\_\_ (similar) groups for strata and \_\_\_\_\_ (reflecting the variability in the population) groups for clusters.

In many cases, multiple sampling methods can be combined.

\_\_\_\_\_ : Systematic sampling is a procedure that can be employed when it is possible to view the population of interest as consisting of a list or some other sequential arrangement. A value  $k$  is specified (for example,  $k = 50$  or  $k = 200$ ). Then one of the first  $k$  individuals is selected at random, after which every  $k$ th individual in the sequence is included in the sample. A sample selected in this way is called a "1 in  $k$ " systematic sample.

Note:  $k = \frac{\text{population size}}{\text{sample size}}$

As long as there are no repeating patterns in the population, systematic sampling works reasonably well. The potential danger is that if there are such patterns, systematic sampling can result in an unrepresentative sample.

\_\_\_\_\_ : It is often tempting to resort to this form of sampling—using an easily available or convenient group to form a sample. Results from such samples are rarely informative, and it is a mistake to try to generalize from a convenience sample to any larger population.

*AP QUESTION:*

**HW #4: Suggested Reading (278-281) Problems page 290 (11, 13, 15, 17, 19)**

## **Monday, August 17: Chapter 13: Experiments and Observational Studies**

ARTICLE: “ADHD linked to lead and Mom’s smoking”

Based on this article, can we conclude that smoking or lead exposure *causes* ADHD?

When it is impossible to tell which of 2 or more factors is causing a change in the response variable, we say the factors are \_\_\_\_\_.

Studies like this one are called \_\_\_\_\_ because researchers don’t assign subjects to do one thing and other subjects something else. In an observational study, we CANNOT conclude that changes in the explanatory variable *cause* changes in the response variable because of the presence of confounding variables.

Is there any way we can show that smoking causes ADHD?

An \_\_\_\_\_ investigates how a response variable behaves when the researcher manipulates one or more factors to determine if changes in those factors *cause* changes in the response variable.

The primary difference between an experiment and an observational study is the way in which the groups are formed. If groups are formed based on the choices of the subjects, then a study is observational. If a researcher assigns groups at random, then the study is an experiment.

If humans are being experimented on, they are called \_\_\_\_\_. Other individuals (tomato plants, mice, loads of dirty laundry) are commonly referred to as \_\_\_\_\_. An experimental unit is the smallest unit to which a treatment is applied.

The specific values that the experimenter chooses for a factor are called the \_\_\_\_\_ of the factor.



## **Tuesday August 18: Chapter 13: Designing Experiments**

*Please Don't Look Ahead in the Notes!!*

Suppose we wanted to design an experiment to see if caffeine affects pulse rate.

What is the explanatory variable (factor)?

What is the response variable?

Who will be the experimental units?

Here is an initial plan:

- measure initial pulse rate
- give each student some caffeine
- wait for a specified time
- measure final pulse rate
- compare final and initial rates

What are some problems with this plan?

Some problems, such as telling a joke while waiting for the caffeine, can be easily solved by including a \_\_\_\_\_ which does not receive caffeine. In our experiment, we can accomplish this by using 2 \_\_\_\_\_ of caffeine: no caffeine and some caffeine. For example, we could assign each member one of two \_\_\_\_\_: Regular Coke or Caffeine Free Coke.

Why don't we give Coke to one group and nothing to the other group?

Often times applying *any* treatment can create a change in the response variable. For example, when a child gets hurt, they feel better when their wound is kissed or covered with a band-aid, even though neither of those treatments actually take away the pain.

In our study, if only one group got a treatment, the fact that they were chosen to receive free soda might make their pulse increase before the caffeine even hits their bloodstream!

The \_\_\_\_\_ occurs when subjects in an experiment know they are receiving a treatment. This knowledge may cause a change in the response variable which \_\_\_\_\_ the effect of the treatment. In other words, we will not know which caused the change in the response variable: the explanatory variable or the placebo effect.

Def: A \_\_\_\_\_ is a treatment known to have no effect, administered so that all groups experience the same conditions. In this case, caffeine-free Coke is a placebo.

Having every subject receive a similar looking treatment ensures that the placebo effect will treat both groups the same. Then, any difference between their pulse rates can be attributed to the \_\_\_\_\_ (factor) and not the excitement of being in an experiment.

Of course, it is essential that the subjects do not know which treatment they are receiving! When a person doesn't know who is receiving which treatment, that person is \_\_\_\_\_.

There are two classes of individuals who can influence the results of an experiment:

- those who could influence the results (subjects, treatment administrators, etc.)
- those who evaluate the results

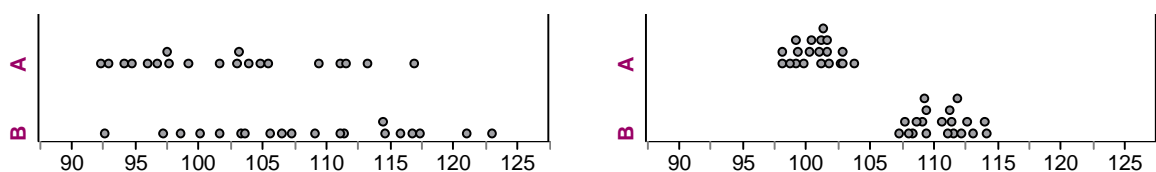
When every individual in one of these classes is blinded, the experiment is called \_\_\_\_\_. If every individual in both classes is blinded, then the experiment is \_\_\_\_\_.

Can our experiment be run in a \_\_\_\_\_ manner?

**Key Principles of a Good Experiment:** THE BIG IDEA--Our goal when designing an experiment is to make the treatment groups as similar as possible, with the exception of the treatments. Then, if there is a change in the response, it can be attributed to the explanatory variable (factor) and not any other extraneous variables.

An \_\_\_\_\_ is one that is not of interest in the current study but is thought to affect the response variable. We need to be aware of extraneous variables for two reasons:

1. Extraneous variables have the potential to become confounding variables.
  - For example, sugar is an extraneous variable since it may affect pulse rates. If one treatment group was given regular Coke (which has sugar) and the other treatment group was given caffeine free Diet Coke (which has no sugar), then sugar and caffeine would be confounded. If there was a difference in the average pulse rates of the two groups after receiving the treatments, we wouldn't know which variable caused the change, and to what extent. To prevent sugar from becoming a \_\_\_\_\_ variable, we need to make sure that both treatment groups get the same amount of sugar.
2. Extraneous variables create extra variability in the response variable, making it harder to estimate the effect of the treatment
  - For example, the rate at which the subjects drink the soda is an extraneous variable since it may affect pulse rates. If we let subjects drink the soda at any rate they want, the changes in pulse rates will probably be more variable than if we made sure each subject drank the soda at the same rate.
  - The two sets of plots below have approximately the same centers, but much different amounts of variability. Which set shows the second center is higher more convincingly?



**Principle #1:** \_\_\_\_\_ means holding extraneous variables constant for all treatment groups so that their effects are not confounded with the explanatory variable. This eliminates these variables as sources of variability.

If we do not control these extraneous variables by making them the same for all treatment groups, they could confound the effects of the caffeine on pulse rates or create extra variability in pulse rates.

**Principle #2:** \_\_\_\_\_ is random assignment of subjects to treatments to ensure that the experiment doesn't systematically favor one treatment over the other.

What about all of the other extraneous variables we do not think of? What about the variables we cannot directly control or block for? What if a critic asks "what about this variable?"

If we randomly assign subjects to treatments, this should \_\_\_\_\_ (but not eliminate) the effects of these variables since their effects should be spread equally between the treatment groups.

Note: We must ALWAYS randomize since there will always be extraneous variables we do not consider. Randomizing guards against what we don't know and prevents people from asking "But what about this variable?"

How do we randomize?

**Principle #3:** \_\_\_\_\_ means ensuring that there is an adequate number of observations in each treatment group.

If each treatment group only had one experimental unit, then we would not be able to conclude that any changes in the response are due to the treatments. It is also possible that some characteristic of the unit was the cause of the change.

Increasing the \_\_\_\_\_ makes randomization more effective. The more subjects we have, the more balanced our treatment groups will be. For example, if we have 10 subjects and only 2 have a certain unknown characteristic, it is quite likely that both of those subjects will end up in the same treatment group simply by chance.

However, if we have 100 subjects and 20 have the characteristic, it is very unlikely for all 20 to end up in the same group. There is a much better chance that the groups will be close to balanced (10/10, 9/11, 11/9, etc.) when the sample size is larger.

Note: Replication can also refer to repeating the experiment with different subjects. This can help us feel more confident applying the results of our experiment to a \_\_\_\_\_.

SUMMARY: With control, blocking, randomization, and replication, each treatment group should be nearly identical, and the effects of extraneous variables should be the same in each group. Now, if changes in the explanatory variable are associated with changes in the response variable, we can conclude that it is a cause-and-effect relationship.

**HW #6: SR (296-300) Problems page 315 (25abc, 28, 34)**

**Thursday, August 20: Chapter 13: More Designing Experiments / The Caffeine Experiment**

Not all experiments have \_\_\_\_\_ or use a \_\_\_\_\_, as long as there is comparison. For example, if you are testing a new drug, it is usually compared to the currently used drug, not a placebo. Also, you can do an experiment to compare four brands of paint without using a placebo.

There are also ethical issues to consider when doing experiments:

The results of an experiment are called \_\_\_\_\_ if they are unlikely to occur by random chance.

For example, if caffeine really has no effect on pulse rates, then the average pulse rate of the two groups should be \_\_\_\_\_. However, because the results will vary depending on which subjects are assigned to which group, the averages will probably differ slightly. Thus, whenever we do an experiment and find a difference between two groups, we need to determine if this difference occurred because of \_\_\_\_\_ or because there really is a difference in the treatments.

The \_\_\_\_\_ refers to the type of inferences (conclusions) that can be drawn from a study. The types of inferences we can make (inferences about the population and inferences about cause-and-effect) are determined by two factors in the design of the study: how the subjects were selected from the population and how the subjects were assigned to groups.

		Allocation of Subjects to Groups	
		Randomized	Not Randomized
Selection of Subjects from Population	Random	Inferences about the population and inferences about cause and effect can be made	Inferences about the population can be made but not about cause and effect. Some observational studies are in this category.
	Not Random	Inferences about cause and effect can be made, but not about the population (only those in the study). Most experiments are in this category.	No inferences about the population or about cause and effect can be made. Some observational studies are in this category.

**Examples from Dan Teague, NCSSM**

**Suppose a dentist wants to know if a daily dose of 500 mg of vitamin C will result in fewer canker sores in the mouth than taking no vitamin C.**

Case 1) The dentist, working through the local dental society, convinces all of the dental patients in town with appointments the first two weeks in December to be subjects in an experiment. He divides them into two groups, those who take at least 500 mg of vitamin C each day and those who don't. He

then asks them how often they have canker sores in their mouth and checks their patients records to see who has complained about canker sores. He compares the proportion of those who take vitamin C daily and complain of canker sores with the proportion of those who don't take vitamin C and complain of canker sores. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

Case 2) A dentist, working through the local dental society, convinces all of the dental patients in town with appointments the first two weeks in December to be subjects in an experiment. He randomly assigns half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months. At the end of this time he determines the proportion of each group that has suffered from canker sores during those three months. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

Case 3) The dentist, working through the local dental society, selects a random sample of dental patients in town and convinces them to be subjects in an experiment. He divides them into two groups, those who take at least 500 mg of vitamin C each day and those who don't. He then asks them how often they have canker sores in their mouth and checks their patients records to see who has complained about canker sores. He compares the proportion of those who take vitamin C daily and complain of canker sores with the proportion of those who don't take vitamin C and complain of canker sores. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

Case 4) The dentist, working through the local dental society, selects a random sample of dental patients in town and convinces them to be subjects in an experiment. He randomly assigns half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months. At the end of this time he determines the proportion of each group that has suffered from canker sores during those three months. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

*AP QUESTION 1999 #3 (APPLES AND DENTISTS)*

**HW #7: SR (300-304), Problems page 316 (32, 37, 43)**

## **Monday, August 24: Chapter 13: Blocking in Experiments**

**Principle #4:** \_\_\_\_\_ is when subjects are divided into homogeneous groups (blocks) based on some extraneous variable and then separated into different treatment groups.

What if men react differently to caffeine than women?

How can we eliminate this source of variability?

Blocking in experiments is similar to stratification in sampling.

- Blocking reduces a source of variability, just like stratifying.
- Blocks should be chosen like strata: the units within the block should be similar, but different than the units in the other blocks. You should only block when you expect that the blocking variable is associated with the response variable.

What are some other extraneous variables that we can block for?

You should try to make the blocks as small as possible. Ideally, the size of the block should be the same as the number of treatments. For example, if there are 3 treatments, then there should be 3 subjects in each block.

If each block has only 2 subjects, then the subjects are called a \_\_\_\_\_.

How can we assign treatments in a matched pair?

If you do not use blocking when dividing the subjects, the result is a \_\_\_\_\_  
\_\_\_\_\_

If you incorporate blocking in your design, it is called a \_\_\_\_\_  
(every subject is assigned to a block based on some characteristics and the members of the block are randomly assigned to the different treatments).

**HW #8: SR (304-309, skip “Lurking or Confounding”), Problems page 315 (23, 38, 39)**

**Tuesday, August 25: Matched Pairs Cholesterol Activity / Review**

*Matched Pairs Activity*

*AP QUESTION*

**HW #9: Problems page 315 (26abc, 30, 42), page 290 (12, 21, 29)**

**Thursday, August 27: Review chapters 12-13**

*AP QUESTION*

**Note: Optional Review Assignment online with solutions!**

**Monday, August 31: Test chapters 12-13**