Chapter 5 Notes Guide
5.1 Designing Samples

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: ____________________ involves studying a part in order to gain information about the whole.

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* HK 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 UHS students, we *estimate* the true average height is 67.1 inches.

Sampling method refers to the process used to choose the sample from the population. Poor sampling methods can produce misleading conclusions.

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. Let's consider, the Gettysburg Address by Abraham Lincoln.

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

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 is a potential problem.

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!

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.

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: 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 that 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.

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.

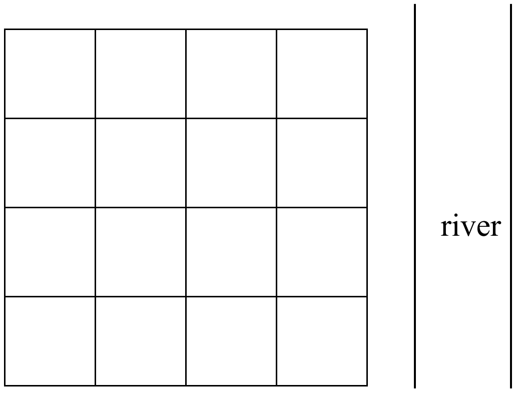
Suppose we wanted to get a stratified random sample of UHS to answer a question about assemblies. Since sophomores and juniors may have different views than seniors, we want to make sure each group is properly represented in our sample.

Suppose there are 800 sophomores, 700 juniors, and 500 seniors. If we wanted to take a stratified random sample of size 100, how many of each class should be included?

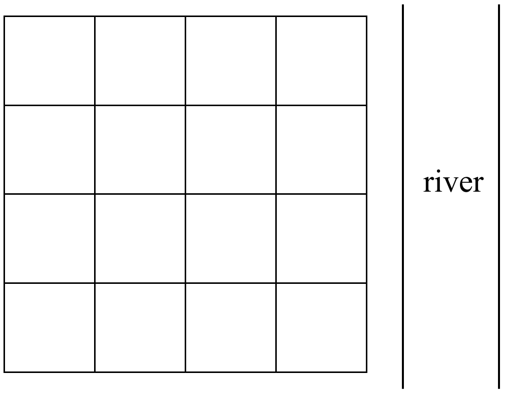
Once we determine the number of subjects to select from each stratum, we take a SRS within each stratum.

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.

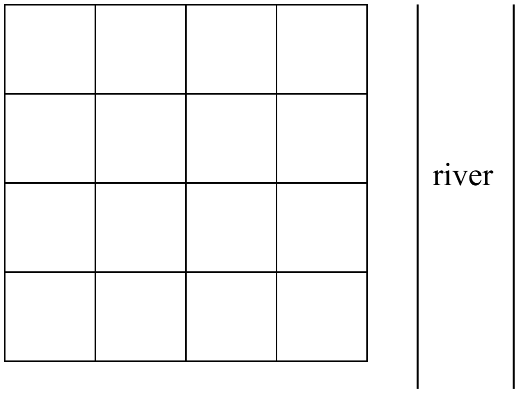
Using a random number generator, pick a simple random sample (SRS) of 4 plots. Place an X in the 4 plots that you choose.



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?

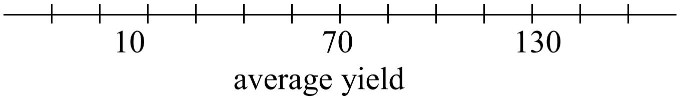
Now, its time for the harvest! The numbers below are the yield for each of the 16 plots. For each of your three samples above, calculate the average yield.

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

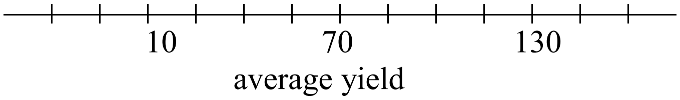
river

Graphing the results:

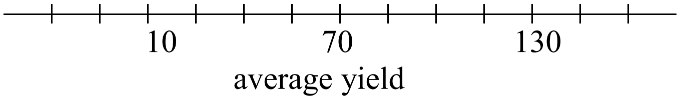
Simple Random Sample:



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:

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.

Other Sampling Methods

___________________________: Sometimes it is easier to select groups from a population then 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 kth individual in the sequence is included in the sample. A sample selected in this way is called a "1 in k" systematic sample.

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

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.

Experiments and Observational Studies

ARTICLE: “ADHD linked to lead and Mom’s smoking”
http://www.nbcwashington.com/news/health/ADHD_Linked_To_Lead_and_Mom_s_Smoking.html

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. In an experiment we study the specific factors we are interested in, while controlling the effects of lurking variables.

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.

The combination of specific levels from all the factors that the experimental unit receives is known as its ______________________.

A recent study declared that people who go to church have longer life expectancies than people who don't go to church.

- Do you think this was an observational study or an experiment? Explain.

- Assuming there is an association between church attendance and longer lives, can we conclude that going to church is the cause?

Section 5.2: Designing Experiments

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 Coke 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 are 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.

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.

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 and 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 thestudy). 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?

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).