AP STATS NOTES Section 4-1

Collecting Data

In order to better understand the characteristics of a population, statisticians and researchers often use a sample from that population and **make inferences** based on the summery results from the sample. Polling is an example of sampling from the population in order to get a better idea of the characteristics of a population. Because we make inferences about a population from the sample, it is very important that the sample is collected appropriately and that it is **representative** of the population being studied. The following is a list of possible sample designs and some of the advantages and disadvantages of each:

1) **Convenience sampling** – Uses subjects that are readily available.
   Advantage: Easy and less costly to collect
   Disadvantage: Not representative of the population
   Example: In order to get an idea of how students think of the new school policy, the principal stands outside the library and asks a few students their opinions.

2) **Voluntary Response Sample** – A sample obtained by allowing subjects to themselves decide whether or not to respond. (A.K.A. Self –selected survey)
   Advantage – Easy to collect
   Disadvantage – Over represents people with strong opinions.
   Example: After the State of the Union speech, ABC tells its audience to call a 1-900-555-1234 if they thought the speech was good and 1-900-555-7890 if they thought the speech was bad (there is a $0.50 charge for the call).

3) **Systematic random sampling** – randomly select an arbitrary starting point, and then select every kth member of the population

   Advantage: Every member has an equal probability of being selected
   Disadvantage: Not every sample of size n has an equal chance of being selected
   Example: HP Selects every 200^{th} computer off the assembly line and inspects it for quality control.

Continued on the next page…

4) **Simple Random Sample (SRS)** – consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance of being the sample actually selected. This is often the best and most appropriate way to collect data for a sample.
Advantages – Easy to accomplish using a table of random digits; likely to produce samples that are good representatives of the population.
Disadvantage – None (could be cost prohibited)
Example: In order to determine how happy students are with their education at DHS, the principal assigns each student a number from 1 to 850 (the number of students at the school) and then uses a random number generator to choose 50 numbers between 1 and 850. He then surveys all the students with the chosen numbers.

5) **Stratified random sampling** – Divide the population into groups of similar individuals (strata) then select an SRS within each strata. Combine the SRSs from each strata to form your full sample.
Advantage: Can produce more exact information (especially in large populations) by taking advantage of the fact that individuals in the same strata are similar to one another.
Disadvantage: Not appropriate unless strata are easily defined.
Example: In order to get a better idea of what DHS athletes thought about homecoming last year, the director divides all athletes into the teams they play for, and then selects a random sample from each sports team. His full sample consists of aggregating the random samples form each team.

6) **Cluster Sampling (Multi-stage sampling)** – Divide the population into sections (clusters) then randomly choose a few of those clusters, and select every member of the clusters chosen.
Advantage – Don't need a list of entire population
Disadvantage – More variability between samples depending on how clusters are determined.
Example – A psychologist at the University of Pennsylvania collects a sample by first dividing up the students into their respective schools (Wharton, engineering, nursing, arts and sciences) then by the departments that their major is in, and then she selects a few departments at random and surveys every student within those chosen departments.

Term:

Sampling Frame – the list from which the sample is drawn

Randomization Methods

Method 1: Roll a Die or 2 Dice

Method 2: Put all the items in a hat and draw them out randomly

Method 3: Use a random number table (Table D in AP Guide)

Method 4: Use technology to randomize items

AP Stats Calculator Program > "3 random numbers"

Option 1: Random Number given a number range to stay within. Results in 1 number.

Option 2: Random List given a number range to stay within AND the total number of random digits desired. Results in as many numbers as you desire.

Option 3: Shuffle a set of numbers within a range.

Sources of Bias

Samples are **biased** if they are systematically not representative of the desired population.

Under-coverage: Occurs when some groups in the population are left out of the process of choosing a sample.

Example: Because they are generally fearful of government intrusion, many immigrants from Latin America did not return their census questionnaire during the 1990 census.

Non-response: Occurs when an individual chosen for a sample can't be contact or refuses to respond. Non-response is a big problem in mail surveys.

Example: The DHS administration sends out 100 survey questions to a sample of DHS parents in order to gage their attitudes toward the school. Only 23 surveys are returned. We have a non-response rate of 77%.

Response Bias: Caused by the behavior of the respondent or the interviewer

Untruthful answers: people give untruthful answers for several reasons:

1) Sensitive questions

Example: How often do you run red lights?

2) Socially acceptable answers

Example: Do you use corporal punishment with your children?

3) Telling the interviewer what he or she wants to hear.

Example: One year after the Detroit race riots of 1967, interviewers asked a sample of black residents in Detroit if they felt they could trust most white people, some white people, or none at all. When the interviewer was white, 35% answered "most"; when the interviewer was black, 7% answered "most"

The fix: secret ballots, anonymous surveys, "sensitive question" techniques.

Ignorant people **–** People will give silly answers just so that they won't appear like they know nothing about the subject.

Example: In a study educators were asked how they would rank Princeton's undergraduate business program. In every case, it was rated among the top 10 departments in the country, even though Princeton doesn't offer an undergraduate business major.

Lack of memory: giving a wrong answer simply because respondent doesn't remember the correct answer.

Example: Students were asked to report their grade point averages. Researchers then determined the actual GPA's. Over 17% of the students reported a GPA that was .4 or more above their actual average, and about 2% reported a GPA more than .4 below their actual GPA. (more inflated their GPA's!)

Timing: When a survey is taken can have an impact on the answers.

Example: in January, the National Football League reported a poll that revealed football as the nation's favorite sport (this is at the time of the Super Bowl)

Phrasing of questions: Subtle differences in phrasing make large differences in the results.

Example:

a) Should the president have the line-item veto to eliminate waste? 97% said "yes"

b) Should the president have the line item veto? 57% said "yes"

Errors

When drawing a sample, two types of errors may occur:

Sampling Error: The difference between a sample result and the true population result. This error results from chance variation.

> *Example*: Place 50 red and 50 green balls in a bag. Mix the balls thoroughly and randomly sample 30 balls. In your sample you find that 12 balls are red and 18 are green. Your sample result (12:18 = 2:3) is different than the true population ratio of 50:50 which is 1 to 1. This difference is due to sampling error. Virtually any experiment involving a sample will have sampling error. We can minimize sampling error through various statistical techniques; the most obvious is to increase the sample size.

Non-sampling error – Occurs when the sample data are incorrectly collected, recorded or analyzed. Such an error results from an error other than chance sample fluctuations. Usually occurs when the sample is selected in a non-random fashion with obvious sources of bias.

> *Example*: In order to gage student opinion on a new grading policy, an administrator stands outside the library during common time and asks a sample of 50 students if they agree with the new policy. The administrator finds that 25 out of the 50 students sampled agree with the new policy. When the entire student body is asked their opinion, however, the results were 30% in favor 70% against. The difference between the sample percentage (25/50 = 50%) and the true population percentage is due to non-sampling error, because the sample was collected in such a way that a lot of bias was involved (convenience sampling).

Answering a free response question on how to conduct a poll:

1. Describe marking the population
2. Describe what randomization tool you will use (table, die, hat, calculator, etc.)
3. Describe the method of selection (interpret how the random numbers become subjects)
4. State how repeats will be handled.
5. Answer the question and ensure that the selection is a Simple Random Sample.

Homework for Section 4-1

A: 1,3,5,7,9,11

B: 17,19,21,23,25

C: 27-29, 31, 33, 35

X: Read the How to Conduct a Survey by Gallop and write down 24 things you learned (4 per page x 6 pages)

Section 4-2 Experiments

Terms

Experimental Units – The things on which the experiment is done.
Subjects – When the experimental units are human beings
Treatment – A specific experimental condition applied to the units.
Factors – The explanatory variables in an experiment

Question: What is the number one advantage of an experiment over an observational study?
Answer: In principle, experiments can give good evidence for causation.

Two other advantages of an experiment:

1.) We can study the specific factors we are interested in while controlling the effects of lurking variables.
2.) Experiments also allow us to study the combined effects of several factors.

Are the following an experiment or an observational study?

1.) A medical team examines the records of 5 large hospitals and compares the survival times of those cancer patients who had surgery versus those who had chemotherapy.

2.) In a gym class, the effect of exercise on blood pressure is studied by requiring that half of the students walk a mile each day while the other students run a mile each day.

3.) The relationship between weights of bears and their lengths is studied by measuring bears that have been anesthetized.

4.) People who smoke are asked to halve the number of cigarettes consumed each day so that any effect on pulse rate can be measured.

Determine if it is an observational study or an experiment, and then identify the explanatory and response variables in each situation.

1.) One effect of alcohol is a drop in body temperature. To study this effect, researchers give several amounts of alcohol to mice, and then measured the change in each mouse's body temperature.

2.) A study if done to try and find the correlation between verbal and math SAT scores. The scientist wants to use the verbal score to predict the math score.

3.) Some breast cancer patients were given each a new treatment. The patients were closely followed to se how long they lived following surgery.

4.) To find out how well a child's height predicts their age a study was done where they measured the heights of a group of children at age 6, wait until they are 16 and then measure their heights again.

Lurking Variable

A variable that is not among the explanatory or response variables in a study but that may influence the response variable.

Confounding

Confounding is when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other. We cannot state the effect of the explanatory variable on the response variable because there is another variable that could also have affected the response variable.

When a lurking variable is not addressed in the design of the experiment, then the results are confounded.

How to design an Experiment

Keep in mind the 3 critical elements of experimental design:

1. Control
2. Randomization
3. Replication

- Control for lurking variables. Use a comparative design and ensure that the only difference between the groups is the treatment administered. A **Control Group** is treated identically in all respects to the group receiving the treatment except that the members of the control group do not receive the treatment. **Placebos**: There is a proven phenomenon called the placebo effect. Patients receiving Placebo tablets which have no active drug ingredient (e.g. a sugar tablet) may experience a certain beneficial effect.

- **Randomization** is the most important element of any experiment. It must be incorporated either in the selection process of experimental units and/or the distribution of experimental units into treatment and control groups. You can use your calculator, the random digit table or names out of a hat or flipping a coin to randomize an experiment.

- **Replication** means the study should be repeated on a large number of subjects. This ensures that the results are due to the treatment and not to variation with the subjects.

Statistically Significant

An observed effect is so large that it would rarely occur by chance is called statistically significant.

Blinding

It is usually best if the subject does not know whether they are receiving the treatment or not. This practice is called Blinding. Sometimes it is also best if the experimenter does not know which subject is receiving the treatment and which is not. This will remove any potential bias in the way the experimenter reports his findings. Experiments in which both the subject and the administrator of the experiment do not know who receives the treatment are called **Double Blind**.

Block Designs

A **Block** is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.
In a **randomized block design**, the random assignment of the experimental units to treatments is carried out separately within each block.

The purpose of blocking is to reduce variation for the response variable.

Consider the final analysis between the explanatory variable (x) and the response variable (y) (scatterplot, etc.), these variable are NOT what are blocked. There is another variable that is assumed to impact the response variable (y). Therefore the population is split into blocks (groups) based on the other variable. We can now say we have <u>controlled</u> for the other variable.

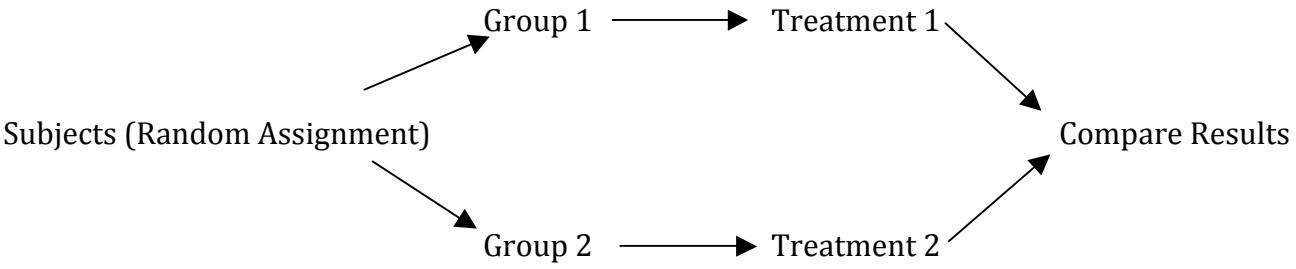
AP Statement when asked about blocking:

The researchers should block based on _____ because it would lead to less variation
how to block
in _____.
response variable

What is the difference between Stratification and Blocking?

- Both separate subjects into groups.
- Stratification is used in surveys. (Stratification – think of quotas required).
- Blocking is used in experiments

The following is a generic experimental design using blocking



Example: A fitness instructor believes that a certain exercise regimen will increase upper body strength. He recruits students to test his theory by having them do as many pushups as they can after they complete the training. How would a block design improve this experiment?

Matched Pair Designs

These are experimental designs in which either the same individual or two matched individuals are assigned to receive the treatment and the control. In the case where an individual receives both the treatment and the control, the order in which this happens should be random. And the experiment should be conducted as a Double Blind experiment.

Example: A medical researcher is interested in testing a new medication for poison ivy. He decides to conduct a clinical trial on 250 volunteers who are allergic to poison ivy. He purposefully rubs poison ivy on their calf, then after the rash appears, he gives half of the volunteer's calamine lotion, and the other half he gives his new medication. How can this experiment be conducted in a matched pair design?

Homework for Section 4-2

D: 37-42, 45, 47, 49, 51, 53
E: 57, 63, 65, 67, 69, 71
F: 73, 75, 77, 79, 81, 85

Section 4-3 Inference

Inference

Drawing conclusions based on the data/information.

Two key questions:

1. Were the individuals randomly selected?
   If yes, then you can make inferences about the population. (The sample reflects the larger group)

2. Were the treatments randomly assigned?
   If yes, then you can make inferences about cause and effect. (The results in the response variable were caused by the explanatory variable)

Homework for Section 4-3

G: 91-98, 102-108