Math 107
Beginning Statistics Notes

You might hear statements like this in the press:

1) 4 out of 5 dentists prefer crest
2) 52% of the American people believe President Obama is doing a good job
3) The average salary of an NBA player in the 2011-2012 season was 4.5 million dollars per year.

People have used the word statistics to describe this type of information – when they do so they are using this term to mean that statistics are the data that describe or summarize something. We want to use the OTHER definition of statistics – namely that statistics is the science of collecting, organizing and interpreting data

Like all sciences statistics has its own set of terms and processes – so here are some:

Definitions:

Population: set of all items being studied (could be the set of all people or things)
Sample – A sub collection of the population
Census – collecting information about EVERY item in the population. Most of the time we can not (or it is very impractical to) do a census – we must do a sample
Population parameters – are specific characteristics of the population that are being studied
Sample Statistics are numbers or observations that summarize the raw data

There is usually a five step process in a statistical study

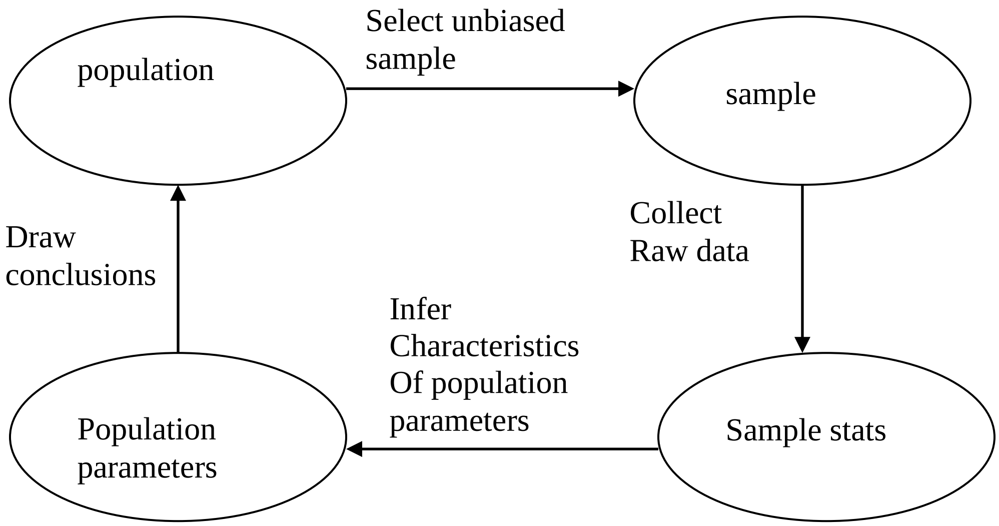
1. Pose the question to be studied – identify the population
2. Choose a sample and collect raw data
3. Create Sample Statistics
4. Infer characteristics of the population from the sample statistics
5. Draw conclusions

Some Examples are in order:

1) Suppose you wish to study if SFCC Students prefer coke or pepsi

   Your population is ALL SFCC students – there are about 5000 of these so taking a census (surveying ALL of them) is impractical. So we take a sample (more later on how to do this) and collect the raw data (do they prefer pepsi or coke?). We might calculate a sample statistics which might be the percentage of people who prefer pepsi over coke. Suppose we get 56% prefer pepsi in our sample. While this does not mean 56% prefer pepsi in the population (ALL SFCC students) we can probably give some guidelines on how many do prefer (we can estimate the population parameter) and draw conclusions

This process can be represented in a diagram:



Example: The National Highway safety commission crash tests 25 2011Mercury Sable wagons at speeds less than 5 mph to determine the average cost of repair. The wagons are taken from 6 different Mercury manufacturing plants throughout the country

Population:
Population Parameter;
Sample:
Sample Statistic:

Example: Mr. Wildman decides to quit his day job and run for governor of Washington. He conducts a name recognition poll of 250 potential Washington voters and find his name is recognized by 10% of them

Population;
Population Parameter:
Sample:
Sample Statistics:

Realize that this 10% comes from the sample – this does not mean that 10% of the population knows who I am – but usually (it is beyond the course to see how) we can ESTIMATE the population parameter (for example we might say between 7% and 13% recognize the name)

Example: An airline executive would like to know if there has been an increase in frustration among air travelers with respect to the increased security requirements. Explain how to apply the basic steps of the statistical study to this question

HOW TO SAMPLE

Simple Random Sampling – choose samples of items in such a way so that every sample of a give size has an equal chance of being selected (or each member of the population has an equal chance of being selected)

Example: Computer generates student id numbers and use this sample to see what students think of their advisor

Systematic Sampling: Select every 5th or 10th or kth member of the population

Example: A restaurant might attach to the bill of every 10th diner a survey where they can call a number and express their opinion in return for this they are often given a free drink or meal

Convenience Sample: We select a sample that is easy to collect or convenient. Usually this method of sampling results in worthless data – a bias!

Example: Ask the students in your class if they prefer coke or pepsi

Stratified Sampling – Use this method when you are concerned about differences within subgroups or strata in the population. Identify subgroups and then draw a simple random sample from each

Example: Suppose you want to know the attitudes US citizen have about gun control. You might be concerned that you get in your random sample a group that is entirely made up of urban residents. What you could do is divide the population into urban and rural residents and take a sample from each in the same proportion as the general population. For example if 60% of US citizen are urban and 40% rural. You might select 100 people in your sample with 60 being urban and 40 rural.

PROBLEMS:

Determine the method of sampling used:

1) In order to complete a student survey for accreditation, SFCC uses a computer program to randomly create 100 student identification numbers. These students are asked to be in the sample

2) You are doing a survey to investigate racial prejudice. You classify the united states population into 5 different racial groups: White, Hispanic, African-American, American Indian and Asian. You take a random sample from each group (in the same proportion as the population) and use these people for your sample

3) Red Lobster asks every 20th diner to fill out a survey concerning the food and service at their restaurant

4) You are interested in whether students prefer coke or pepsi, so you ask 20 students in your English class

5) You manufacture light bulbs and are interested in how many bulbs in each batch of 1000 bulbs are defective. You test each 10th bulb off the assembly line to determine if it is defective

6) A university is interested in determining if Student Activities Funds are spent effectively. The student population is divided into undergraduate students and graduate students and a random sample is taken from each group to construct the entire sample

7) You design two T-shirts and are interested in whether others would buy the shirts. You go to the mall and ask people seated in the food court their opinions of the shirts

8) To conduct a phone survey, a company uses a computer to randomly generate telephone numbers and uses these numbers as their sample

TYPE OF STATISTICAL STUDIES

What is the average neck size of Yellowstone Bears? To study this question you would have to take random sample of Yellowstone Bears (good luck ☺) and measure their individual neck sizes. This is an example of an observational study – where researchers observe or measure characteristics of the sample but do not attempt to influence or modify these characteristics

Suppose you have developed a new drug to control the symptoms of diabetes. You might take a random sample of 40 diabetes sufferers and divide them into two groups – the first group would receive your new drug (this group is called the treatment group) and the second group called a control group would receive a placebo – this is something that lacks the active ingredient of a treatment being tested but is identical in appearance to the treatment. The presence of a placebo is to help with the placebo effect – where patients improve only because they believe they are receiving treatment. This is an example of an experiment - where researchers divide the sample into two or more groups, each is treated differently.

To insure against some additional bias – usually blinding is used. A single blind experiment is one in which the participants do not know whether they are in the control group or treatment group. A double blind experiment is where both participants and the experimenters do not know who belongs to which group

A good way to think about the difference between an observational study and an experiment, is that in an experiment the researchers determine who is in the treatment group and in an observational study researchers don't determine who is in the treatment group.

There is an **association** between the explanatory variable and the response variable if a change in one of the variable corresponds with a change in the other variable. For example, towns with more churches tend to have more bars. There is an association between churches and bars. However, the presence of churches does not cause more bars! If the change in the explanatory variable causes a change in the response variable, we say there is **causality.**

A well designed experiment can determine whether there is causality whereas most observational experiments can only determine associations.

In the church/bar example, there is something else happening that is causing the association. Such variable are called **lurking variable** or **confounding variables.**

Estimating A Parameter with a Large Sample

At SFCC in the year 2011, 38% of the 6508 students were the first in their family to attend college.

a) What is the number of students who were first in their family to attend college?
b) A random sample of 1588 students was taken. 578 were first generational students. What is the sample proportion of first generational students? Does the sample proportion closely estimate the population proportion?
c) Another random sample of 350 students was taken. In this sample 147 students were first generational students. What is the sample proportion of first generational students? Does the sample proportion closely estimate the population proportion?
d) Why are the answers different?

The difference of the sample proportion and the population proportion is called the sampling error.

Definition: Sampling error is the error involved in using a sample to estimate information about a population due to randomness in the sample.

Thinking back to **The Law of Large Numbers** from probability, it should make sense that in general if the sample is chosen “wisely”
i) the larger the sample, the smaller the sampling error
ii) the larger the sample, the less likely we are to encounter “bad luck”

MARGIN OF ERROR AND CONFIDENCE INTERVALS

Suppose you have the following results from a poll

48% of the voters approve of candidate X with a margin of error of plus or minus 3%

What does this mean?

Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter; however, the interval computed from a particular sample does not necessarily include the true value of the parameter. When we say, "we are 99% confident that the true value of the parameter is in our confidence interval", we express that 99% of the hypothetically observed confidence intervals will hold the true value of the parameter. After any particular sample is taken, the population parameter is either in the interval, realized or not; it is not a matter of chance. The desired level of confidence is set by the researcher (not determined by data).

Certain factors may affect the confidence interval size including size of sample, level of confidence, and population variability. A larger sample size normally will lead to a better estimate of the population parameter.

So what does “48% of the voters approve of candidate X with a margin of error of plus or minus 3% points” mean at a 95% confidence level?

If we conducted this poll 100 times, we would expect our sample parameter to be within 3% points of the actual population parameter about 95 times and more than 3% points away from the true parameter about 5 times. That is, about 5% of the time, are estimation won’t be “close” to the true parameter just due to plain bad luck.