Samples, populations, and random sampling

I. Samples and populations.

Suppose we tried to figure out the weights of everyone on campus. How could we do this?

Weigh everyone. Is this practical? Possible? Accurate?

Try counting every word in your textbook. You think you might do better by estimating?

Your text mentions trying to weigh every grasshopper in Kansas.

Again is this even possible?

A recent (?) debate in congress and the supreme court dealt with this issue as it applied to the census.

Obviously we need to do this differently:

Pick some people/pages/grasshoppers at random (more on this soon) and hope that they represent, in some way, the populations.

Figure out what you want based on this sample.

Advantages:

Easier, less expensive, possibly more accurate.

Often it is impossible to "measure" a whole population, even if we wanted to.

So, we ESTIMATE what we want about our population. This is sometimes also called *statistical inference* - making conclusions about a population based on a sample.

A. Populations:

We need to define this carefully. Suppose we were trying to figure out peoples' hair color. What is our population?

- GMU?

- United States?

- Asia?

- Europe?

- World?

Once we know, we can figure out how to get a sample. Trying to figure out hair color of folks up in Norway isn't going to work if we sample Asia.

Some examples:

Suppose we want to get an idea of blood types in England.

Sampling 3,696 people in England is probably a good way of trying to figure out the overall proportion of blood types. Our population is the proportion of blood types in England.

Feeding gerbils to cats. Suppose someone was interested in trying (for whatever reason) to figure out how many gerbils a cat can eat at one time.

The person takes 15 hungry cats, and starts feeding them gerbils in a lab. The number of gerbils each cat eats within an hour is written down. What is the population?

The population is the number of gerbils a cat can eat under conditions similar to this experiment.

Note that the population here is not "naturally occurring", but rather something set up in a lab.

Most often what we are trying to do is to make conclusions about populations where we can't measure everything. We are not often interested in conclusions about small groups such as the:

Height of people in this class

The specific number of words on a particular page of your text

Weight of grasshoppers in my sample from Kansas.

Usually we're really interested in might be:

Height of people on campus

The number of words in the text

The average weight of grasshoppers in Kansas

II. Estimates and parameters.

In general, we usually don't know the "true" average or standard deviation of a population.

Can we really measure the height of everyone on campus?

The "true" population mean (which we don't know) is symbolized with the Greek letter mu or "μ".

Similarly, the "true" population standard deviation is symbolized by the Greek letter sigma or "σ".

We don't know what these are, so we estimate them with the sample mean and sample standard deviation:

$\bar{y}$ *estimates* μ (*mu*)

s estimates σ (*sigma*)

How good are our estimates? We'll learn that a bit later.

"True" population characteristics are often symbolized by Greek letters and termed "parameters".

Sample characteristics (or "statistics" which we use to estimate parameters) are often symbolized by Latin letters.

But not always:

If we're looking at a proportion, we might consider the proportion of people infected with AIDS, "*p*".

p is the true unknown proportion.

we try to estimate *p* with *p*-hat, or $\hat{p}$.

some people do use π (pi - may not show on the web) for proportions, but that letter is generally thought to be "taken" by 3.14159.....

(Despite saying that Greek letters are always used as parameters, your text does this the same way as described here).

The "^" (hat) symbol always means "estimate" in statistics, but a lot of the other stuff varies depending on who writes the book (and even sometimes in the same book!)

so $\hat{p}$ estimates *p*, the true proportion.

IV. Random sampling.

Problem - we need to make sure our sample is representative of the population.

A possibility:

- number every item in the population

- pick *n* random numbers

- sample those items that match the random numbers

The idea:

- pick the sample in such a way so that each item in the population has the same probability of being picked.

- If I pick item *x*, it does not influence the probability of picking item *y*.

Random numbers:

1) for a small sample, use a random number table (see table B41, p. 856).

- try to pick a "random" starting point in the table.

- pick the appropriate number of digits

  - if you have 150 items in your population, you should pick three digit numbers.

- ignore any number that doesn't fit your sample

- if you have 150 items, ignore numbers larger than 150.

- continue until you have however many numbers you need.

  - you can pick the second, third, etc. numbers simply by moving to the left, right, up or down. It's irrelevant.

2) for bigger samples, use computer generated random numbers. R will easily generate random numbers for you (look up the runif command from the command line).

Here's an example using R ("#" means everything following that is a comment; you can just copy and past the following if you want to try it):

```
# generate a sequence of 35000 numbers;
y <- seq(1:35000)
# take a look at what you got;
# (it'll scroll by quickly);
y
# Now get a sample of 150 numbers;
# make sure to use the replace statement;
sampy <- sample(y,150, replace = FALSE)
# Sort them;
sampy <- sort(sampy)
# Now take a look at them;
sampy
```

- [an aside - computer generated random numbers are actually "pseudo-random". There are books written on this topic, including one by my major advisor!]

- The random number generator in older versions of Excel is one of the worst ones we know about (the numbers don't appear truly random). Microsoft has been told about this, but, until recently, hasn't addressed this.

  Even in recent versions, no one seems to know for sure what Microsoft is doing.

Other comments.

You want to make sure you actually do correct random sampling.

The population needs to be carefully defined, and then one needs to sample randomly from it.

A few brief comments on sample size:

In general, a larger sample is better.

You should have a large enough sample so that inferences (=conclusions) about the population are believable.

We'll learn a bit more about this later.

There are actually other ways of taking a sample (systematic sampling, for example), but we don't have time to go into these.

There are whole texts written on sampling.

All of them will have some kind of random component, though it may not be apparent.

Sometimes it's difficult to take a real "random" sample.

How would you do this for our grasshoppers???

(Opportunistic sampling)