Lecture 4

Last time

We went from density curves all the way to calculating probabilities with a normal distribution.

What was all that stuff?



Well, depending on how the data is distributed (left-skewed, symmetric, wide dispersion=high variation in observations), there are many different density curves to describe it. A density curve is a smooth curve that approximate the trends in the data.

What are the two properties of all density curves?

We talked about the empirical rule that allows us to calculate probabilities of occurrences that are within one, two and three standard deviations away from the mean.

What about if we want to know the probability of occurrences that are .5 standard deviations away from the mean? Do we just give up? No.

- (1) We need the underlying population to have a normal distribution.
- (2) We need to know the mean and standard deviation.

Why?

Then, we can standardize any observation of any normal distribution and use the probabilities of the standard normal distribution to find our answer!!

Example:

We assume that the weight of population of Harris and Ft. Bend Counties has a normal distribution with μ =112.55 and σ =21.36.

 $Z = \frac{90 - 112.55}{21.36} = -1.06$

What does -1.06 mean?

We can use Table A to find the probability associated to Z=-1.06. It is .1446 so 14.46% of people weigh below 90 pounds.

Samples and Populations

Now, we're going to change our focus to collecting our data in a way that allows us to make correct inferences to the population from out sample. The main goal is to collect a sample that closely resembles the population. We know it will never be exactly like the population, the only way to do this is to take a **census** which collects data on every single person in the population (most of the time this is impossible). This means that all of the summaries (graphical: histogram, pie chart or numerical: \bar{X} , s^2) will change depending on the sample we have in hand. This change in different samples is called **sampling variability**.

Example:

Consider our BFRSS data set that consists of 20,000 observations and now let that be the population under study.

We can create a **random sample** of size 50 and then look at a histogram of BMI measurements for men.

What do we expect to see?



These graphs show our sampling variability. Each plot, each descriptive statistic, are slightly different depending on the sample that was brawn. If we are open-minded about our inability to get a perfect sample then we can measure the sampling variability and accurately approximate the difference: this is why our standard deviation is so important.

1 How to sample

A Simple Random Sample (SRS consists of n individuals chosen from a population of size N so that each possible sample has an equal chance of being selected.

Example:

It's windy and snowing outside and you and your roommates have just ordered a pizza with a coupon that's only valid when you pick it up at the restaurant. Of course, no one is volunteering to go, so you write each person's name on a slip of paper that is the same size, color and texture and put the names into a hat. You then draw one name. This is the same idea behind random sampling because each person from the population has the same chance of being chosen.

Example:

Five trees in an apple orchard of 15 trees will be selected and tested for disease. take a simple random sample of trees. The trees are laid our in the orchard as seen below. First, label the individuals.

*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Then, using Table B: Random Number Table, pick any line and start reading off numbers in pairs. Get 5 numbers between 1 and 15. These 5 numbers will be our random sample of trees.

A Simple Random Sample is a type of **probability sample** which is a sample that gives each member of a population a known chance (between 0 and 1) to be selected. This probability or chance of being selected is equal for all individuals in an Simple Random Sample.

A **Stratified Random Sample** (not SRS) first divides the population into similar groups of individuals called strata and then takes a Simple Random Sample (SRS) of the individuals within each strata.

Example:

If the A&M student body is the population, then possible strata could be gender, race, class or county of residence.

Example:

The city-council planning board has put together a package of new roads, road maintenance, and new bridges to be built. They want the county's opinion on this proposal. The population they want to sample is the group of all voting age residents of the county. There are 10,000 living in the county. They plan to sample 100 individuals. They divide the population into two stratum:

- rural residents who live outside the city limits

- urban residents who live inside the city limits

There are 3,000 rural residents and 7,000 urban residents. To get proportional representation from each group, 30 rural residents and 70 urban residents could be chosen at random from each strata.

A stratified random sample, if done correctly, can be more representative and more informative than an SRS with statistics from each stratum.

A **Systematic Sample** is used, say, when we have a list of 100 subjects and we want a sample of size twenty, we can divide the list into twenty smaller lists (of equal size) and then select one unit from the first list (the kth one, say) and then select the kth unit from the subsequent lists; in general, we select every kth observation.

A Multistage Sample is a combination of sampling schemes (SRS, stratified, systematic, cluster).

Example:

If the A&M student body is the population, I first stratify by major, then within each major, I select every 10th name from an alphabetical list of each major.

Example:

A survey of households is made. They are stratified by house and apartment, and then by north and south side of town. Each group is broken down into clusters of either 4 or 8 units. An SRS of clusters is taken, and every individual within that cluster is examined.

Sampling Issues

Undercoverage - subjects of a certain demographic are omitted from the sampling procedure

Example:

Random digit dialing concerning presidential performance omits individuals that have no phone or only a cell phone

Nonresponse- when subjects do no cooperate or are not available

Example:

I select every 5th customer that comes into Blockbuster and the 35th person tells me to get lost and leaves.

Response Bias- when question may cause respondents to lie or choose a response that the "interviewer wants"; in general, anything where the responses exhibit unusual bias

Example:

- Do you smoke marijuana? (respondent may lie)

- Do you suffer from persistent diarrhea? (respondent may be embarrassed)

- Do you think abortion should be illegal so that innocent lives are saved? (second part of the question exhibits bias against those who are not pro-life)

Question Wording - confusing or leading questions can influence a respondent's answer and can completely change the outcome of a survey.

Example:

Which question is better to ask if you want to know how many people believe that George W. Bush was elected fairly?:

(A) In light of the debate over Florida's electoral college votes, was George W. Bush rightfully elected?

(B) With many votes left uncounted and recounts of votes in several counties abandoned due to partian interference, do you believe that George W. Bush was rightfully elected in 2000?

Real world sampling

Election day 2004:

Exit polls are a kind of survey, like the CDCs BRFSS. Instead of calling people, interviewers are asked to approach voters as they leave their polling places As you may recall, based on these samples, we were getting lots of reports of a Kerry lead. What happened?

The good people at Mitofsky International and Edison Media Research train their exit pollsters to sample voters. In January of last year they released a report illustrating what we call **sampling bias**.

Interviewing Rate

(Problems) generally (get) worse as the interviewing rate increases. This occurs at polling locations where a large number of people are voting, either because our sample precinct is large or because other precincts in addition to our sample precinct may be voting at the same polling place. The increased (problems) in these precincts could suggest that some interviewers do not follow the interviewing rate exactly. As the interviewing rate increases so does the potential for interviewers to exercise more of their own judgment on whom they will approach in order to participate in the exit poll. However, the data also show (errors) in the Kerry direction still exists even in precincts where the interviewer was instructed to ask every single voter to participate in the exit poll (and the interviewer had no option in the selection of the respondent). In precincts with an interviewing rate of 1, there was still (an error) in the Kerry direction of almost 4 points. Again, this indicates that a portion of the (error) is coming from differential nonresponse.

Swing states

The (error) was greater in the more competitive "swing" states.... This indicates that voters in the swing states (who were exposed to more paid advertising and media coverage than voters in non-swing states) were less likely to respond to the exit poll: but among those who did, more likely to be Kerry voters.

Interviewer age

Older interviewers had (fewer problems) than the youngest interviewers. They also had better completion rates. This does not necessarily mean that the younger interviewers did poorly at their task. It indicate that in this election voters were less likely to complete questionnaires from younger interviewers.

Interviewer education

The (errors) decreased slightly when the interviewer had more education. However, the (problems were the greatest) among those with post-graduate education. They had a significantly greater overstatement of Kerry than any other group. Those with High School education or less had a lower overstatement of Kerry but a higher absolute error.

Taken from http://www.exit-poll.net/index.html