8th Class **6/17/10**

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."--H.G. Wells

review main result from Ch. 6:

Normal Approximation Rule, a.k.a. Central Limit Theorem

$\overline{X}$ is distributed approximately normal with parameters $(\mu, \frac{\sigma}{\sqrt{n}})$

while the population distribution may or may not be normal, but has parameters (μ, σ)

repercussion of this theorem: can answer questions about how close the sample mean $\overline{X}$ is to the population mean μ by using the standard normal table after transforming $\overline{X}$:

$$Z = \frac{\overline{X} - \mu}{SE}$$

Normal Approximation Rule for Proportions:

as n increases, the sampling distribution of P, the sample proportion, becomes concentrated around π and becomes more and more of a normal distribution, regardless of the underlying shape of the population distribution

the standard error of P is $\sqrt{\frac{\pi (1 - \pi)}{n}}$, which goes to 0 as n gets large

so P is distributed with parameters $(\pi, \sqrt{\frac{\pi (1 - \pi)}{n}})$

therefore, we can again use standard normal tables to describe how close the sample mean P is to the population proportion π by using the standard normal table after transforming P:

$$Z = \frac{P - \pi}{SE}$$

note that proportions are simply means for (0,1) (dummy) variables

Discuss sampling distributions
[do M&Ms experiment]
this is one way of building up a sampling distribution; another way one can sometimes do it is with the Monte Carlo technique: sampling of a population by matching each person with a serial number and selecting serial numbers randomly using a random digit generator on the computer or a random digit table if doing it by hand; then calculating $\overline{X}$ over and over again to build up a sampling distribution of $\overline{X}$. This sampling distribution will settle in around the population mean μ. This technique can be used to calculate the sampling distribution of other sample statistics as well, such as the median.

This can be applied to large or small populations; in sampling with replacement, only relative frequencies matter. The size of the population N is irrelevant.

[discuss use of table of random digits in the book]
[discuss pseudo-random number generators--start from a seed number--useful if you want to use the same "random" number sequence over again]
[show clip from Numb3rs season 2 episode "Double Down" 32:24-34:47 re how machines, including computers, can't generate truely random numbers]
[discuss true random number generators--show them from the course webpage]
[Dilbert cartoon]

one caveat and one extension on Ch. 6 material:

--for small samples, the central limit theorem doesn't really apply, so the normal approximation is not going to be very good.

--note that standard error is reduced when sampling is done without replacement

SE reduction factor in case where sampling without replacement is done:

$\sqrt{\frac{N - n}{N - 1}}$ where N = population size and n = sample size

so SE of $\overline{X}$ in this case $= \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}}$

useful applications from Ch. 6: discuss regression to the mean and the importance of considering sample size in evaluating results
[overheads and handout]

start Ch. 7:

review concepts of population and sample.

essential to remember that the population mean μ and variance σ^2 are constants (though generally unknown). They are called population <u>parameters</u>. By contrast, the sample mean $\overline{X}$ and sample variance s^2 are random variables, where $\overline{X}$ is distributed approximately $N(\mu, \frac{\sigma^2}{n})$.

to summarize, a random sample is a random subset of the population, in which relative frequencies $\frac{f}{n}$ are used to compute $\overline{X}$ and s^2. These random variables are examples of <u>statistics</u>, or <u>estimators</u>. In a population, probabilities p(x) are used to compute μ and σ^2. These fixed constants are examples of <u>parameters</u>, or <u>targets</u>.

This chapter deals with outlining desirable properties of estimators.

What does it mean for an estimator to be <u>unbiased</u>?
very important idea in statistics

U is an unbiased estimator of θ if $E(U) = \theta$

so an estimator V is called <u>biased</u> if $E(V) \neq \theta$

and bias is defined as the difference between the expected value of the estimator and the population parameter:

$$\text{bias} \equiv E(V) - \theta$$

What does it mean for an estimator to be efficient? Defined in relative terms to other estimators by ratio of their variances, for two unbiased estimators U and W:

$$\text{Efficiency of estimator U compared to estimator W} \equiv \frac{\text{Var(W)}}{\text{Var(U)}}$$

So in comparing unbiased estimators to decide which is better, simply pick the more efficient one; i.e., the one with minimum variance. But suppose you are comparing a biased estimator to an unbiased one, or two biased ones to each other. We may want to use the criteria of minimizing some combination of bias and variance. One criterion that is widely used is mean squared error (MSE):

$$\text{for estimator V, its MSE} \equiv E(V - \theta)^2$$

this can be shown to be equal to the linear combination of variance plus squared bias:

$$\text{MSE} = \text{Var}(V) + [\text{Bias}(V)]^2$$

as follows:

$$E(V - \theta)^2 = E[V^2 - 2V\theta + \theta^2] \qquad \text{by multiplying it out}$$

$$= E(V^2) - 2E(V)\theta + \theta^2 \qquad \text{distributing the E operator}$$

$$= E(V^2) - [E(V)]^2 + [E(V)]^2 - 2E(V)\theta + \theta^2$$

$$\text{adding and subtracting } [E(V)]^2$$

$$= \{E(V^2) - [E(V)]^2\} + \{[E(V)]^2 - 2E(V)\theta + \theta^2\} \quad \text{grouping terms}$$

$= \{Var(V)\} + [E(V) - \theta]^2$ first term is variance formula; second term can be written as..

$= Var(V) + [Bias(V)]^2$ second term is bias formula
squared

So can now generalize our criterion for deciding on relative efficiency between two estimators for case of any two estimators, whether biased or unbiased:

$$\text{Efficiency of estimator V compared to estimator W} \equiv \frac{MSE(W)}{MSE(V)}$$

a more <u>accurate</u> estimator has lower MSE than comparison estimators; a more <u>precise</u> estimator has lower variance than comparison estimators

which is more important?
[robot arm example--know the amount of the bias and adjust for it]
[shooting at a ship example--small bias off of the sighter so still hit the ship]

What does it mean for an estimator to be <u>consistent</u>? Idea of the limit on the MSE as n approaches infinity. A consistent estimator has MSE approaching 0 as n approaches infinity. So one of the conditions that makes an estimator consistent is if its bias and variance both approach zero as n approaches infinity.

> If Bias(V) approaches 0 as n approaches infinity, V is called <u>asymptotically unbiased</u>.

So if V's variance also approaches 0, V will be consistent.

mention example 7-5: both MSD and sample variance are consistent estimators of the population variance.

read Ch. 8 for next class

Statistical principles that violate common sense

Questions to illustrate statistical principles

A. The policy of one school is to punish students for being late, while the corresponding policy in an otherwise identical school is to reward students for being on time. If effectiveness is measured by behavior on the day following punishment or reward, which policy will seem more effective?

B. Studies have shown that in the New York City subways crime rates fall in the years following increased police patrols. Does this pattern suggest that the increased patrols are the cause of the crime reductions?

C. A certain town is served by two hospitals. In the larger hospital, about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. Although the overall proportion of boys is around 50 percent, the actual proportion at either hospital may be greater or less than 50 percent on any given day. At the end of a year, which hospital will have the greater number of days on which more than 60 percent of the babies born were boys?

a. The larger hospital
b. The smaller hospital
c. Neither—the number of days will be about the same

Regression to the mean

In any series of random events clustering around an average, or mean, an extraordinary event is most likely to be followed by a an ordinary event. Although regression is usually discussed in narrowly statistical terms, it affects virtually every series of events that is to some degree random. And since there is almost nothing in life that is not at least partly a matter of chance, regression shows up in a wide variety of places. It helps explain why brilliant wives tend to have slightly duller husbands, great movies have disappointing sequels, disastrous presidents have better successors, and "rookies of the year" generally do worse their second year.

Regression to the mean can also explain why behavior modification techniques often appear to work conversely. As psychologists Daniel Kahneman and Amos Tversky write: "By regression alone, behavior is most likely to improve after punishment and to deteriorate after

reward. Consequently, the human condition is such that…one is most often rewarded for punishing others and most often punished for rewarding them."

People often detect the effects of regression to the mean but invent other causes for it. Tversky comments: "Listen to the commentators at the Winter Olympics. If a ski jumper has done well on his last jump, they say, 'He's under immense pressure, so he's unlikely to do as well this time.' If he did poorly, they say, 'He's very loose and can only improve.'" People feel intuitively that an output should be representative of the input.

You can profit from understanding regression to the mean. Economists Richard Thaler and Werner De Bondt, using stock market data for the past 50 years, show that an investor who bought only stocks that had declined the most in value over the previous five years would earn about 30 percent above the market average in the next five years, even though some of these acquisitions went out of business.

The effects of sample size

Imagine an urn filled with white balls and black balls. You know that two-thirds of the balls are one color and one-third are the other, but you don't know which color predominates. One blindfolded person plunges a hand into the urn and comes up with three black balls and one white ball. Another uses both hands and comes up with 14 black balls and ten white balls. Both samples suggest that black balls are more numerous. But which sample provides the more convincing evidence?

Many people find the first sample more compelling. After all, it has black in the majority by a three-to-one margin, while the second sample is only a little more than half black. However, the odds that the second sample accurately indicates the majority color in the urn are 16 to 1, but are only 4 to 1 for the first sample. This is because the first sample is smaller and therefore less reliable. Remember that it is fairly common to flip a coin four times and get heads three-quarters of the time. But the chance of obtaining a proportion so out of line with the real odds of 50-50 is exceedingly small after 1,000 flips.

Though people are aware of this fact in an abstract way, they often ignore it. Squash players may think it makes no difference whether they play a game to 9 or to 15 points; actually the shorter game gives the weaker player a much better chance of winning, because he/she has to win that many fewer lucky points. Similarly, only about one person in five realizes in the hospital problem that the institution with the smaller number of babies per day will have many more days of 60 percent boys (55 such days a year on average at the small hospital, 27 days at the large hospital).