11th Class **6/22/10**

"It is a part of probability that many improbable things will happen. "--Agathon (445 - 400 BC)

[show results from bootstrap experiment in last class]

review briefly from last class
[A confidence interval can be interpreted as the set of plausible, or acceptable hypotheses. Any hypothesis lying outside the confidence interval may be judged implausible, and is therefore rejected. If a 95% confidence interval is being used, it is natural to speak of the hypothesis as being tested at a 95% confidence level. However, we traditionally speak of testing at an error level of 5%, and say that a plausible hypothesis is statistically significant at the 5% significance level.

distinguish statistical significance (at a stated significance level) from significance/importance
note problem of large samples--even very small differences may be statistically significant if sample size is large enough
also, difference can be significant, but doesn't prove any particular explanation causes the difference (consider case of men's and women's salaries)

null and alternative hypotheses. Often, the alternative hypothesis is the one we want to prove and null is everything else.

alternative hypotheses can be one-sided or two sided

e.g., two-sided alternative hypothesis

H_0: $\mu = 0$

H_A: $\mu \neq 0$

one-sided alternative hypothesis

H_0: $\mu \leq 0$ (can just write $\mu = 0$, but this is what we are testing)

H_A: $\mu > 0$

[handout of example of ground beef percent fat testing]

traditional approach: either reject the null hypothesis or can't reject it, at the given significance level, based on the sample value (i.e., does the null hypothesis fall within the confidence interval surrounding the sample value)

modern approach: calculate the probability, or p-value, that the null hypothesis is true, given the sample value

p-value ≡ Pr(the sample value would be as large as the value actually observed if H_0 is true)

in practice, we do this by calculating a t-statistic:

$$t = \frac{\text{estimate} - \text{null hypothesis}}{\text{SE}}$$

If $H_0 = 0$ (very common), then this reduces to:

$$t = \frac{\text{estimate}}{\text{SE}}$$

(again, use t-table because SE is really estimated SE, using the sample variance instead of the population variance)

we have already gotten the material to calculate this if we have calculated a confidence interval:

$$\text{parameter} = \text{estimate} \pm t_{.025}\text{SE}$$

pull out the estimate and SE, take their ratio to get a t-value, and look up the corresponding probability or p-value in the t-table in the row corresponding to the sample's degrees of freedom (or use the standard normal table if n is sufficiently large)

if testing some null hypothesis ≠ 0, pull out the estimate and SE, subtract the null hypothesis from the estimate and then take the ratio and look it up in the table.

classical hypothesis testing: reject H_0 if its p-value ≤ α, the predetermined error level; generally just report whether H_A is significant at the α% significance level, rather than reporting the p-value

modern hypothesis testing: just report the p-value and let people decide for themselves whether or not to reject H_0]

use the standard normal table if n is sufficiently large [show example using polling data from right before 2004 presidential election: Bush polled 49%, actually got 51.5% of the popular vote. SE = SQRT(.49*.51/1200) = .01443, t = (.49-.515)/.01443, This gives a t-statistic of about 1.73, which leads to a p-value using the standard normal table (assumed sample size of about 1200) of .041; two-sided p-value would be .082, so above the .05 error level but not that great]
[ask what null I am testing--that the poll was drawn from the true voting population]

Type I vs. Type II error

α = probability of making a Type I error (rejecting H_0 when H_0 is true);
value of α is called the <u>level of the test</u>

β = probability of making a Type II error (not rejecting H_0 when H_0 is false)

$1 - \alpha$ = probability of making the correct decision to accept H_0 when H_0 is true;
value of $1 - \alpha$ is called the <u>confidence level of the test</u>

$1 - \beta$ = probability of making the correct decision to reject H_0 when H_0 is false;
value of $1- \beta$ is called the <u>power of the test</u>

dilemma: holding n constant, the higher the confidence level of the test, the lower the power of the test, and vice-versa, so if α is decreased, β is raised, and vice-versa
[show diagram like Figure 9-6]

examples:
[overhead on decision theory]

[handout on biometrics and associated type I and type II error]

mention discussion of other types of error besides I and II [wikipedia page]

modify idea that Hs must cover the whole number line: composite hypothesis can be broken into parts, particularly if have strong prior about the alternative value

holding n and α constant, we can calculate β for different alternative hypotheses (for something like H_A: $\mu > 0$ is a composite hypothesis of all possible μs that exceed 0), where β drops as H_A is farther away from H_0 [as in Figure 9-10]

however, increasing sample size n can allow for both the confidence level and power of the test to be increased (or one to be increased without lowering the other)

discuss the operating characteristics curve (OCC) and graph one

The U.S. Justice system as a paradigm of hypothesis testing

1) A person is presumed innocent until proven guilty. So the null hypothesis is that the defendant is innocent, and the alternative hypothesis is the plaintiff's or prosecutor's challenging claim that the person is guilty. Precisely one of these hypotheses is correct.
2) The burden of proof rests on the accuser. So just as in hypothesis testing, evidence must be collected by the accuser.
3) Two verdicts are possible. The defendant can be found guilty (rejecting the null hypothesis) or not guilty (failing to reject the null hypothesis). In the latter case, the defendant is not proclaimed innocent; the evidence given is simply deemed insufficient to convict.
4) Two errors are possible. An innocent person can be convicted (a type I error, rejecting the null hypothesis when it is true) or a guilty person can be set free (a type II error, not rejecting the null hypothesis when it is false).
5) The accuser must show their case beyond a reasonable doubt. In our system, it is much worse if an innocent person is convicted, and it is less critical if a guilty person is set free. Thus we want a small value of α (the probability of a type I error), and we aren't as concerned with the value of β (the probability of a type II error).

note: of course it is always better to have more evidence in order to reduce both type I and type II error, but sometimes more evidence is not available (or is extremely costly to gain)

review loss function idea

give spam control example [from Macworld April 2003]

problem is to identify an incoming email message as spam or not spam
spam filter tries to identify its true nature
[ask what the null Hypothesis is in this case, and therefore what the Type I and Type II errors are]
different kinds of spam filters:
Points-based
Bayesian
it is possible to reduce both Type I and Type II error through improved technology

one-sided vs. two-sided tests: if sampling distribution is symmetric, the two-sided p-value is just twice the one-sided p-value (two-sided classical test: distribute error equally between tails; so an error level of $\alpha = 5\%$ puts 2.5% in each tail)

hypothesis testing vs. confidence interval calculations

two-sided classical tests and two-sided confidence intervals (the kind we have so far considered) are equivalent: classical tests and confidence intervals both check to see whether the magnitude of the difference
$|\overline{X} - \mu_0|$ exceeds 1.96 standard errors (for large n); they just have different reference points (classical test refers to μ_0; confidence interval to the observed $\overline{X}$)

can also have a one-sided confidence interval that will be equivalent to a one-sided test; again, all the 5% error is concentrated in one tail. E.g., for a single mean, the one-sided 95% confidence interval to see if the mean is above a certain figure is:

$$\mu > \overline{X} - t_{.05}\frac{s}{\sqrt{n}}$$

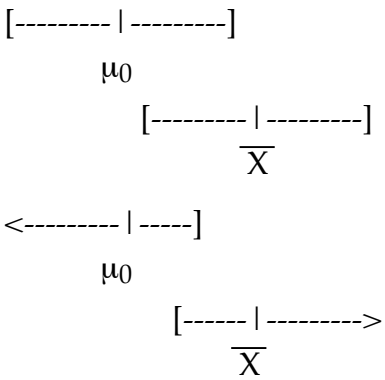
or, if we are testing to see if the mean is below a certain figure:

$$\mu < \overline{X} + t_{.05}\frac{s}{\sqrt{n}}$$

compare one-tailed to two-tailed testing; can make a difference in some cases:

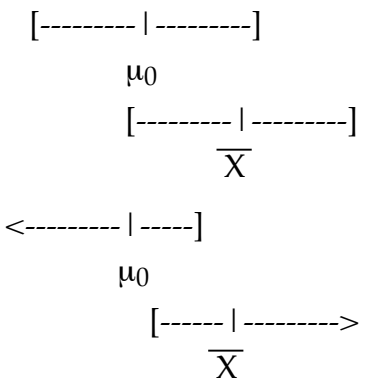
3 examples (for case where $\overline{X} > \mu_0$; μ_0 the boundary point for the one-sided test):

1) reject the null hypothesis under both two-tailed and one-tailed test



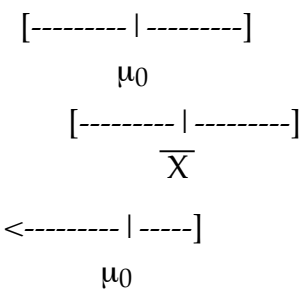
actual distance: | ---------- | | $\mu_0 - \overline{X}$ |

2) accept under the two-tailed test, reject under the one-tailed test



actual distance: | ------- | | $\mu_0 - \overline{X}$ |

3) accept the null hypothesis under both two-tailed and one-tailed test



[------|---------->
$\overline{X}$

actual distance: |-----| $|\mu_0 - \overline{X}|$

to summarize, there are four approaches to hypothesis testing: one-sided and two-sided confidence intervals and classical hypothesis tests

most standard form (and one preferred by statisticians): two-sided confidence interval

why?

i) easiest to understand

ii) gives the crucial point estimate as well as the sampling allowance surrounding it (point estimate often used in further calculations)

iii) it is the form related to more advanced techniques

p-value approach generally lends itself to one-sided hypothesis testing

we have now finished Part II of the course: Inference

start Ch. 11: first part of Part III of the course: relating two or more variables

fitting a line to a scatter plot (Y is <u>dependent variable</u>, X is <u>independent variable</u>; so we are making causality assumption)

[draw up picture]

fit by eye or use a criterion

commonly-used criterion (but not the only one): minimize the sum of squared deviations of Ys from the line

$$\hat{Y} = a + bX$$

$$d = Y - \hat{Y}$$

$$\min. d^2 = \sum (Y - \hat{Y})^2$$

solution to this (using calculus) gives optimal a (intercept) and b (slope):

where $b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}$

and $a = \overline{Y} - b\overline{X}$

note also that

$$b = \frac{Cov(XY)}{Var(X)}$$

(actually MSD in denominator, but close enough for large n)

this is known as ordinary least squares (OLS) estimates of a and b

next time: review this information and extend; go on to discuss regression further

Population of my coin purse:

15 quarters, 5 dimes, 3 nickels, 30 pennies:

$$\mu = \frac{\$4.70}{53} = 8.9 \text{ cents}$$

My sample:

1 quarter, 1 dime, 3 pennies:

$$\overline{X} = \frac{\$0.38}{5} = 7.6 \text{ cents}$$

$$s = \sqrt{\frac{(25-7.6)^2 + (10-7.6)^2 + 3*(1-7.6)^2}{4}}$$

$$= 10.5 \text{ cents}$$

Classical confidence interval recipe:

$$\mu = \overline{X} \pm t_{.025} * \frac{s}{\sqrt{n}}$$

$$\mu = 7.6 \pm 2.78 * \frac{10.5}{\sqrt{5}}$$

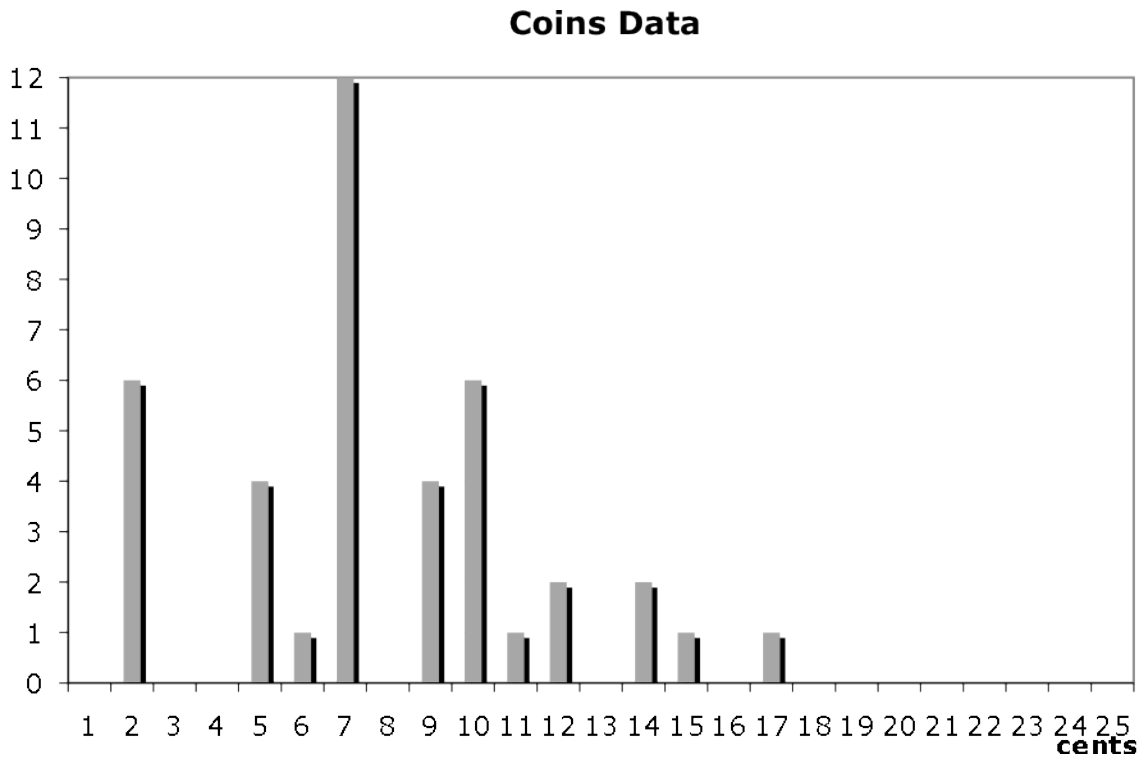
$$\mu = 7.6¢ \pm 13.1¢$$

$$-5.5¢ < \mu < 21.7¢$$

this has the problem of containing unrealistic values, namely values below 1 cent, which we know cannot be true in this case

Bootstrap confidence interval:
take 40 random samples with replacement from my sample of 1 quarter, 1 dime, 3 pennies:

all found combinations:	f	f/n	$\overline{X}$
1 dime, 4 pennies:	6	.150	2.8¢
1 quarter, 4 pennies:	4	.100	5.8¢
3 dimes, 2 pennies:	1	.025	6.4¢
1 quarter, 1 dime, 3 pennies:	12	.300	7.6¢
1 quarter, 2 dimes, 2 pennies:	4	.100	9.4¢
2 quarters, 3 pennies:	6	.150	10.6¢
1 quarter, 3 dimes, 1 penny:	1	.025	11.2¢
2 quarters, 1 dime, 2 pennies:	2	.050	12.4¢
2 quarters, 2 dimes, 1 penny:	2	.050	14.2¢
3 quarters, 2 pennies:	1	.025	15.4¢
3 quarters, 1 dime, 1 penny:	1	.025	17.2¢
	40	1.000	



throw out the two extreme observations to get a 95% confidence interval (not relevant on the high end here):

$$2.8¢ < \mu < 15.4¢$$

the bootstrap method generates a narrower 95% confidence interval than the classical method and does not contain unrealistic values (namely values below 1 cent or above 25 cents).