10th Class **6/21/10**

"What used to be called prejudice is now called a null hypothesis."--A.W.F. Edwards

"79.48% of all statistics are made up on the spot."--John Paulos

[show psychic cartoon]

[hand out two overhead handouts from last time that I forgot to give out]

review formulas from last class: five are of general use: single mean, difference in means between independent samples, difference in means between matched samples, single proportion, difference between proportions

1) 95% confidence interval for the population mean when σ is unknown:

$$\mu = \overline{X} \pm t_{.025}\frac{s}{\sqrt{n}} \quad ; \text{d.f.} = n-1$$

problem 8-8:

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

and sample size n = 5, so d.f. = 4

$$\text{so } \mu = 44.2 \pm 2.78\,\frac{35.6}{\sqrt{5}} = 44.2 \pm 44.3$$

and total = $50(44.2 \pm 44.3) = 2210 \pm 2215$, so about 0 to 4400, which includes 3620

2) 95% confidence interval for the difference between population means, independent samples, if the population variances are unknown but assumed to be equal:

$$(\mu_1 - \mu_2) = (\overline{X_1} - \overline{X_2}) \pm t_{.025}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{where } s_p^2 = \frac{\sum_{n_1} (X_1 - \overline{X_1})^2 + \sum_{n_2} (X_2 - \overline{X_2})^2}{(n_1 - 1) + (n_2 - 1)} \text{ ; d.f.} = (n_1 - 1) + (n_2 - 1)$$

problem 8-12:

$$(\mu_1 - \mu_2) = (\overline{X_1} - \overline{X_2}) \pm t_{.025}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{where } s_p^2 = \frac{\sum (X_1 - \overline{X_1})^2 + \sum (X_2 - \overline{X_2})^2}{(n_1 - 1) + (n_2 - 1)}$$

and d.f. for use in the t-table $= (n_1 - 1) + (n_2 - 1) = 28$

$$\text{so } (\mu_1 - \mu_2) = (11.0 - 16.0) \pm 2.05 \sqrt{\frac{40 + 786}{4 + 24}} \sqrt{\frac{1}{5} + \frac{1}{25}}$$

$$= -5.0 \pm 5.5, \text{ or } -10.5 \text{ to } 0.5 \text{ for the difference}$$

so it is still possible that women could earn more than men, as part of the confidence interval is positive!

3) 95% confidence interval for the mean population difference, dependent/matched/paired samples:

$$\Delta = \overline{D} \pm t_{.025} \frac{s_D}{\sqrt{n}}$$

$$\text{where } \overline{D} = \frac{\sum D}{n} = \frac{\sum (X_1 - X_2)}{n}; \text{ n = number of paired measurements;}$$

$$s_D^2 = \frac{\sum_{n} (D - \overline{D})^2}{n - 1} \text{ ; d.f.} = n - 1$$

note that the mean of the differences Δ equals the difference of the means $(\mu_1 - \mu_2)$

compare this formula for calculating the confidence interval for differences in matched samples:

$$(\mu_1 - \mu_2) = (\overline{X_1} - \overline{X_2}) \pm t_{.025}\frac{s_D}{\sqrt{n}} \; ; \; \text{d.f.} = n - 1$$

$$\text{where } s_D^2 = \frac{\sum_n (D - \overline{D})^2}{n - 1}$$

to the formula for calculating the confidence interval for differences between independent samples:

$$(\mu_1 - \mu_2) = (\overline{X_1} - \overline{X_2}) \pm t_{.025}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{where } s_p^2 = \frac{\sum_{n_1} (X_1 - \overline{X_1})^2 + \sum_{n_2} (X_2 - \overline{X_2})^2}{(n_1 - 1) + (n_2 - 1)} ; \; \text{d.f.} = n_1 + n_2 - 2$$

[do a couple of problems comparing unmatched and matched samples from overhead]

Clarify section on the advantage of matched samples:

in comparing a matched sample of size n to two independent samples, each of size n, while the t-value is lower in the second equation due to the greater degrees of freedom (2n - 2 or 2(n - 1) vs. n - 1), this effect diminishes rapidly as n increases, and the sum of squared deviations for the second equation will generally be much larger than for the first:

$$\sqrt{\frac{\sum_n (D - \overline{D})^2}{n - 1}}\sqrt{\frac{1}{n}} \; < \text{or} > \; \sqrt{\frac{\sum_n (X_1 - \overline{X_1})^2 + \sum_n (X_2 - \overline{X_2})^2}{(n - 1) + (n - 1)}} \; \sqrt{\frac{1}{n} + \frac{1}{n}} \quad ?$$

rearranging slightly:

$$\sqrt{\frac{\sum_n (D - \overline{D})^2}{n-1}}\sqrt{\frac{1}{n}} \; < \text{or} > \; \sqrt{\frac{\sum_n (X_1 - \overline{X_1})^2 + \sum_n (X_2 - \overline{X_2})^2}{2(n-1)}} \; \sqrt{\frac{2}{n}} \quad ?$$

and combining radicals:

$$\sqrt{\frac{\sum_n (D - \overline{D})^2}{(n-1)(n)}} \; < \text{or} > \; \sqrt{\frac{\sum_n (X_1 - \overline{X_1})^2 + \sum_n (X_2 - \overline{X_2})^2}{(n-1)(n)}} \quad ?$$

now, dropping the equal denominators and squaring both sides:

$$\sum_n (D - \overline{D})^2 \; < \text{or} > \; \sum_n (X_1 - \overline{X_1})^2 + \sum_n (X_2 - \overline{X_2})^2 \quad ?$$

note that the left hand side can be rewritten:

$$\sum_n [\, (X_1 - X_2) - (\overline{X_1} - \overline{X_2}) \,]^2$$

and then reorganized:

$$\sum_n [\, (X_1 - \overline{X_1}) - (X_2 - \overline{X_2}) \,]^2$$

and then multiplied out:

$$\sum_n [\, (X_1 - \overline{X_1})^2 + (X_2 - \overline{X_2})^2 - 2(X_1 - \overline{X_1})(X_2 - \overline{X_2}) \,]$$

and the summation sign distributed; now compare to r.h.s.:

$$\sum_n (X_1 - \overline{X_1})^2 + \sum_n (X_2 - \overline{X_2})^2 - 2\sum_n (X_1 - \overline{X_1})(X_2 - \overline{X_2})$$

$$< > \; \sum_n (X_1 - \overline{X_1})^2 + \sum_n (X_2 - \overline{X_2})^2$$

since the first two terms can be cancelled from both sides:

$$-2\sum_{n}(X_1 - \overline{X_1})(X_2 - \overline{X_2}) <> 0 \ ?$$

note that the l.h.s. has the covariance formula:

$$-2\,\mathrm{Cov}(X_1,X_2) <> 0\ ?$$

so the l.h.s. is smaller, so long as $\mathrm{Cov}(X_1,X_2) > 0$

this is a reasonable assumption to make; if $\mathrm{Cov}(X_1,X_2) < 0$ then using matched pairs would be counterproductive and it would be in fact better to use independent samples

example: pairing neighboring plots of land (split-plot design): if one plot is very fertile, it would be natural to expect the neighboring plot to also be very fertile, so they would both vary positively from the mean fertility level both before and after a tested fertilizer treatment. This is essentially the assumption underlying the idea of matched samples, i.e., that $\mathrm{Cov}(X_1,X_2) > 0$

example: checking nutrition effects on IQ: If a child has IQ above the mean before the experiment, would still expect him/her to have IQ above the mean after the experiment; few people would be affected so much by an experiment that they would switch sides of the mean after the experiment

4) 95% confidence interval for a proportion, for large n:

$$\pi = P \pm 1.96\sqrt{\frac{P\,(1 - P)}{n}}$$

if n is small, would use the graphical/geometrical method shown in Figure 8-4 (do a demo of this method)

5) 95% confidence interval for the difference between two population proportions, for large n_1 and n_2, assuming independent samples:

$$(\pi_1 - \pi_2) = (P_1 - P_2) \pm 1.96\sqrt{\frac{P_1\,(1 - P_1)}{n_1} + \frac{P_2\,(1 - P_2)}{n_2}}$$

confidence intervals are not always symmetric
--and not always a simple formula
classical method can give values outside the range (e.g., for π, <0 or >1--see the graph for another way to adjust for the range and allow for asymmetry)

thus may prefer the bootstrap method of calculating confidence intervals in complex situations so as to avoid this problem

--replicate a random sample many times (e.g., a million times); or think of sampling with replacement over and over again

--do a Monte Carlo study to estimate the mean (e.g.) of this "population"

--draw multiple samples from the "population" and calculate correspondingly many $\overline{X}$ s; this is the "bootstrap" sampling distribution

--the central 95% of this distribution provides the confidence interval for μ.

--useful for complicated parameters and funky nonnormal distributions

[do it in class: bring out coin purse, draw five coins out of it, pass sample around in a container for people to draw from and write down their drawing]

On to Chapter 9

Relationship between confidence intervals and hypothesis testing

Once a confidence interval has been calculated, it can be used to test <u>any</u> hypothesis.

example:

According to the *Annual Report of the Dean of Admission* (September 2001), there are 725 people in Wesleyan's Class of 2005; 374 are women, 102 have an alumni affiliation (i.e., a relative who went to Wesleyan), and 49 are both female and have an alumni affiliation.

b. If a randomly-selected student is a women, what is the probability she has an alumni affiliation? If a randomly-selected student is a man, what is the probability he has an alumni affiliation?

49/374 = 13.1%; 53/351 = 15.1%

c. What can we infer from the two numbers in (b) regarding the relationship between gender and alumni affiliation?

now can test for significant difference between the two percentages:

(.151-.131) ± 1.96*sqrt[.151*.849/351 + .131*.869/374]

.02 ± .05

so not significantly different at a high level of confidence

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: $\mu > 0$

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

[next class, finish Ch. 9, start on Ch. 11]