MATH 243, LECTURE 21

1. Recap of confidence intervals for difference of means of two populations

In order to use t-statistics to study two-sample problems, we need the following conditions to be satisfied, just as for single-sample problems.

- An independent SRS from each populations. For example if we were trying to sample men and women, it won't work to take a random sample of men, and then to take their wives or girlfriends as the other sample.
- Both populations need to be normally distributed, or...
- If distributions aren't close to normal but no outliers and no strong skewedness, need sample sizes over 15.
- Generally sample sizes greater than 40 are OK even with strongly skewed distributions or outliers.

If these conditions hold then we can follow these steps to compare these different populations, finding a confidence interval for the difference of means $\mu_1 - \mu_2$.

- Compute the standard error of the two samples, $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- Use the $t(k)$ distribution where k is one less than the smaller of the two sample sizes.
- Find t^* as in the one-sample case so that C% of the area is between $-t^*$ and t^*. We can look this up in Table C.
- With C% confidence, we can say the true difference of means is between $(\overline{x_1} - \overline{x_2}) - t^*SE$ and $(\overline{x_1} - \overline{x_2}) + t^*SE$.

Example 1. *Mean body temperatures: In one study, 65 men and 65 women have their temperature taken (in similar conditions). The male mean is 98.105 with a standard deviation of 0.699. The female mean is 98.394, with a standard deviation of 0.743. Give a 95% confidence interval for the difference between these means and test the hypothesis that women have higher temperatures than men at the 0.05 level. What if the data were drawn from samples of only 20 men and 22 women?*

2. Recap of hypothesis testing for means of two populations

To test the null hypothesis, $H_0 : \mu_1 = \mu_2$, we calculate

$$t_0 = \frac{\overline{x_1} - \overline{x_2}}{SE}$$

Then $P(t \geq |t_0|) < \alpha$ supports the alternative hypothesis $H_a : \mu_1 \neq \mu_2$ at level α.

To summarize, if we use the standard error SE in the places the single-sample standard error was used, we may use the same methods as we have been using to understand the true difference of means from the observed difference of means.

Example 2. *Physicians wish to measure the effectiveness of leech therapy on arthritis pain (I'm not making this up! See*
`http://www.annals.org/cgi/content/full/139/9/724`*)*
They took a sample of 51 patients. They assigned a group of 24 patients to receive "leech therapy," and a control group of 27 patients to receive a conventional pain-relief therapy (diclofenac gel).

One week after the leech treatment (it was one treatment lasting a little over an hour involving 4 to 6 leeches), the leech group had a mean pain index of 19.3 with a standard deviation of 12.2.

After one week of the other treatment*, the control group had a mean pain index of 42.4 with standard deviation of 19.7.*

Test the null hypothesis that the effect of treatment by leeches is the same as the effect of conventional treatments.

3. Sampling to determine proportion of a population having some property

So far we have focused on using the mean (and deviation) of a sample to extrapolate some information about the mean of an entire population. These techniques are applicable in many settings, but (in case you didn't notice) they still don't let us how the confidence intervals of opinion polls work, for example. Fortunately, it doesn't take much reworking of our tools so far to address this setting.

So for example, we might be interested in the question: what proportion of the population is left-handed?

- Take sample from class. Not truly random, but probably random enough for a question like this. Let $\hat{p}$ be the proportion of left-handed people.
- How well does this approximate the proportion p in the general population?

Left-handedness is a categorical variable, with only two values (YES and NO). So it *can't* be normally distributed. What can a histogram look like?

But consider samples of size 10, for example. If we look at what *proportion* of a sample is left-handed, we have 11 possible values: $0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1$.

As the size of the sample increases, the number of possible values for the proportion of left-handed people also increases. We have the following variant of the central limit theorem:

Theorem 3. *Let X be some random variable of a large population which has values YES and NO. Take SRS of size n from our population, and let $\hat{p}$ be the proportion of the sample which is "YES."*

- *For large n, the sampling distribution of $\hat{p}$ is approximately normal; $N(p, \sigma)$ where*
- *p is the proportion of the entire population which is "YES" and*
- $$\sigma = \sqrt{\frac{p(1-p)}{n}}.$$

Example 4. *Suppose that two-thirds of college students have cheated on an exam. What is the probability that in a random sample (taken discretely) of 20 students, 15 or more would have cheated? What is the probability that 10 or more have cheated?*

Getting back to statistical inference, we would like to do inference aimed at estimating p from $\hat{p}$. The normalized z-statistic, which is behind the scenes of both confidence intervals and hypothesis testing, would be $z = \frac{\hat{p}-p}{\sigma}$ where $\sigma = \sqrt{p(1-p)/n}$ as in the theorem. If n is large, then $\hat{p}$ was approximately normal. Thus z will be approximately *standard* normal.

In practice, we won't know p. We use $\hat{p}$ in place of p to get the standard error in place of the standard deviation. So we set $s = \sqrt{\hat{p}(1-\hat{p})/n}$, and then $z = \frac{\hat{p}-p}{s}$. To get a confidence interval with certainty $C\%$, we choose z^* a critical value for C, and then with confidence $C\%$ we know p is between $\hat{p} - z^* \times s$ and $\hat{p} + z^* \times s$.

Example 5. *Use an in-class survey to estimate the percentage of left-handers with 90 and 95 percent confidence.*

To do inference, we need to know we are reasonably close to a normal distribution. Here are some conditions:

- Our sample is a SRS.
- The population is at least 10 times the sample size.
- The sample size is "large enough." (At least 15 successes and 15 failures.)

Unfortunately, even for relatively large n, this can be not so close to Normal. Fix (recommended to always use): the "Plus four" confidence interval. Let $\overline{p} = \frac{\text{successes}+2}{n+4}$.

Then the C% confidence interval is between $\overline{p} - z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$ and $\overline{p} + z^* \sqrt{\frac{\overline{p}(1-\overline{p})}{n+4}}$.

Example 6. *Redo our estimate for left-handers using the "plus four" confidence interval.*

Example 7. *Establish some confidence intervals (both the usual and plus four) for polls found at:* `http://www.usatoday.com/news/polls/tables/live/2005-02-28-poll.htm`

Example 8. *Find some polls on the web which publish their sample size and margin of error, and determine with what certainty the number being measured is within that margin or error.*