Lectures 20 and 21: Hypothesis Testing and Difference of Means Tests
November 15, 17-2000

What do social scientists do?

1) Primary goal is to test hypotheses, that is, we are interested in describing the social world and making inferences, finding relationships, differences, etc. in that social world.
2) Often, we strive to make comparisons—does one group differ from another? If so, how? These questions motivate our analysis of the social world.
   a) Are first time offenders more frequent drug users that repeat offenders?
   b) Do Republicans and Democrats have significantly different perspectives on welfare policy?
   c) What % of women favor Gore compared to men who favored Gore?

**All of these questions demand hypotheses and they ask us to test those hypotheses with statistics.

What types of hypotheses do we have?

1) Let's say we have two populations, 1 and 2, and we are interested in seeing whether or not they have different mean values for a particular variable.
   → Republicans and Democrats and we're interested in their opinions on welfare reform. The scale runs from 0 (greatly disfavor) to 6 (greatly favor).
   a) If we could see the mean of the populations, μ_1 and μ_2, we might find that the Democrats had a $\mu_1 = 2$, and the Republicans had a $\mu_2 = 4$.
   b) Let's say that we want to test whether or not these population mean values (μ_1 and μ_2) are significantly different from one another. That is, we want to see if Republicans' and Democrats' views on welfare reform are significantly different from another.
      → When we say significant, we are saying significant <u>statistically</u>.

2) How do we do it? We rely on the **Null and the Research Hypotheses**
   a) Recall that the null expresses that there is no difference between two things, while the research, or alternative hypothesis, suggests that there is a difference between two things. Further, these hypotheses can be specified to be **directional**, meaning that we can note *how* we expect the two things to be different from one another (greater than or less than).
   b) **Null Hypothesis**: there is no difference between Republicans and Democrats in their opinions on welfare reform. We denote this:

$$\mu_1 = \mu_2$$

   → Keep in mind here that we are talking about **populations**. This does not necessarily imply that sample means are different. Often, though, we will use the μ_1 and μ_2 convention to denote sample means (as in difference of means tests). We may, if we are interested in doing so, find out that our sample means are different even if our population means are not. But this is really not of significant consequence right now.

→ What do we do with the null? We either **accept the null that there is no difference between Republicans and Democrats**, or we **reject the null and say that there is a statistical difference between Republicans and Democrats**.

c) **Research or Alternative Hypothesis**: there is a statistical difference between Republicans and Democrats on the issue of welfare reform. We denote this:

$$\mu_1 \neq \mu_2$$

Sampling and Difference of Means Testing

1) Let's say that we have samples of populations. When we draw samples, we will sometimes find that the sample means are equal, while other times they are not equal. Most of the time, they are not equal. But if we were to take an infinite sample of the difference in means (that is, the mean of one sample minus the mean of another sample), we would produce **a normal distribution with a mean of zero**.

→ This suggests that the average value for the difference in means between any two samples chosen out of a population is 0.

→ Basically, by calculating the difference between means, we end up producing a **sampling distribution of differences between means**, with a sampling distribution that is approximately normal and whose mean is zero.

Testing Hypotheses with Differences of Means

1) This ends up almost like assessing confidence intervals, only the way we label our graph is slightly different. The important points on the graph, instead of being labeled σ, 2σ, etc. are labeled $\sigma_{x\ bar1\ -\ xbar2,}$ where this represents the standard deviation of the differences between x-bar$_1$ and x-bar$_2$.
2) To illustrate how this distribution becomes relevant when testing the differences in means, let's use our Republican and Democrat example:

   a) Draw a random sample of 50 Republicans and another sample consisting of 50 Democrats and ask each sample their opinions on welfare reform. Let's say that the Republican mean is 5 and the Democratic mean is 2.
   b) If we assume that in infinite populations, the difference between means for these population should average out to zero.
   c) A quick look tells us that the difference between our sample means is 3 (5 – 2).

→ This gives us a raw score of 3, but we know that a raw score doesn't tell us much about where the score sits percentile wise, etc. For this, we rely on transforming the raw score into a z-score or a t-score.

- → When we have two sample means, we use modified versions of the z-score, where we subtract one sample mean from another and divide that value by the standard deviation of the differences between means (see Levin and Fox or in-class notes for the equation).
- → If we were to calculate the z-score for our raw score of three, we would simply divide 3 (5-3) by 2 (the standard deviation of the differences between means, assuming that it's the value we're given). This gives us a z-score of 1.33.
- → Thus, a differences of 3 between the two samples falls 1.33 standard deviations of the differences between means above the mean differences of means (whew!).

d) What is the probability that a difference of means of 3 or more can happen strictly on the basis of sampling error (that is, by chance)? Table A in Appendix C of Levin and Fox suggests that 40.82% of the distribution occurs between μ and a z-score of 1.33. This also means that z cuts out 9.18% of the distribution in each tail (draw this out of you're not sure).

e) But do we know whether one mean is significantly different from another? Not quite yet…

Statistical Significance

1) When we do research, we are typically looking at levels of statistical significance, and this is denoted by α. Recall that α is equal (remembering our examples with t-values in confidence interval stuff) to 1 – the desired level of confidence. So if we want to be 95% confident (or better) that Democrats and Republicans have different opinions on welfare reform, to find the appropriate level of statistical significance, we just take $1 - 0.95 = 0.05 = \alpha$. Alpha represents the area to the right or to the left of the critical t-value we find in our Appendix tables (more on that later).
   - → The α value represents the probability that the difference in means we observe will occur by chance. So if $\alpha = 0.05$, we would say that the probability that we observe our difference in means occurring by chance is 0.05.
   - → Remember that the 0.05 level can have to appearances—it can be evenly divided between two tails (so that the area in each tail of the distribution is 0.025), or it can all be in one tail of the distribution (so that the area in that tail is 0.05).
   - → What marks the beginning of the tail of the distribution? The **critical value for t** marks the beginning of the tail of the distribution. How do we find that critical value for t? First, we examine our **degrees of freedom,** then we find the desired level of statistical significance (perhaps we're interested in $\alpha = 0.05$), then we look at Table B (for two-tailed tests of significance) or Table C (one-tailed tests) in Levin and Fox to find the associated **critical value for t**.
   - → For an $\alpha = 0.05$, let's say we obtain a t-value of 1.96 in a two-tailed test of significance. This means that the probability of being in the right tail is 0.025 and the probability of being in the left tail of the distribution is 0.025. The point that marks whether you're in the right or left tail of the distribution is our **critical value for t**.

How do we do a difference in means test? There are just a few easy steps…

1) Write our your null hypothesis and the alternative hypothesis. If the null and alternative do not specify a direction (that is, there are no > or < signs present), then we use a **two-tailed test of significance** and rely on Table B in Levin and Fox. If the null and the alternative specify a direction (that is, there are > or < signs present), then we use a **one-tailed test of significance** and rely on Table C in Levin and Fox.
2) Figure out the sample means (x-bar) for the two samples under consideration.
3) Figure out the sample standard deviations for the two samples under consideration.
4) Calculate the standard error of the difference between means (that really long, messy equation).
5) Calculate our **computed t-statistic**, which is just x-bar_1 minus x-bar_2, divided by the standard error of the difference between means.
6) Use Table B or Table C in Appendix C of Levin and Fox (depending on one-tailed or two-tailed test of significance) to figure out the **critical value for t**. We need three pieces of information to figure out the appropriate t-value.
   a) The degrees of freedom
   b) The alpha level (significance level) desired.
   c) The appropriate "tail test," i.e. one- or two-tailed.
7) Compare the **computed t-statistic and the critical value for t**.
   a) In a two-tailed test, we can reject the null hypothesis that there is no difference between the means if…
      1) The computed t-value is greater than the positive critical value for t **or**
      2) The computed t-value is less than the negative critical value for t.
   b) In a one-tailed test, we deal with two possibilities. If the alternative hypothesis is that $\mu_1 > \mu_2$, then
      1) We reject the null hypothesis if our calculated t-value is larger than our critical t-value.
      2) We accept the null hypothesis if our calculated t-value is smaller than our critical t-value.
   c) If the alternative hypothesis in a one-tailed test is $\mu_1 < \mu_2$, then
      1) We reject the null hypothesis if our calculated t-value is smaller than our negative t-value (- t).
      2) We accept the null hypothesis if our calculated t-value is larger than our negative t-value (-t).