Analyzing Nominal Variables: Comparison of Means and Crosstabs

1) Last time, you learned how to calculate **SAMPLE STATISTICS** for central tendency and for dispersion:

2)
$$\bar{x} = (\sum_i^n x_i)/n$$
$$s^2 = (\sum_i^n (x_i - \bar{x})^2)/(n-1)$$

   a) With this information we can say, for example, what is the probability that we would draw an INDIVIDUAL from a known population who exceeds the population average by a specific amount. This is shown in your book p. 326.

2) What we want to now is make an **inferential leap** to think about samples of samples. We want to be able to say: what is the probability that our **sample mean** will exceed the population mean by a specific amount.

   a) A statistical proof, not shown here, tells us that we can now calculate something called the **standard error** for the mean (the calculation of s is shown above):

3) $$se = \sqrt{\frac{\sigma^2}{N}} = \sqrt{\frac{s^2}{N}} = \frac{s}{\sqrt{n}}$$

   b) This is intuitively correct: any "sample of samples" or set of averages should vary less than the individual cases in a single sample. This is shown in your book pg. 329-331.

   c) For a proportion, the formula looks like this:

5) $$se = \sqrt{\frac{P(1-P)}{N}}$$

      i) which is what most newspapers use when they say that a candidate is leading another candidate by "55% plus or minus" some amount.

3) To use the standard error, we calculate something called the **Z score.** In general, a Z or "t" statistic looks like this:

4) $$Z = \frac{estimate - nullhypothesis}{se}$$

   a) This statistic is normally distributed, and can be looked up in your book (pg. 465)

   b) In most cases, our "null hypothesis" is "no effect" or "no difference", so for example, Women do not view George Bush any differently from Men; Blacks turn out at the same rate as Whites, etc.

      i) DECOMPOSE the women's ranking of Bush, and calculate whether the difference is significant. Can only evaluate α by comparing it to ε

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

      ii) NULL: Women's ranking of Bush = Overall Ranking of Bush ($\alpha=0$)
      iii) HYP: Women's ranking of Bush is significantly different from overall ranking ($\alpha\neq0$)

      i) Verbally, this says: Any individual's score (Y) can be decomposed into the grand mean, the systematic effect of a categorizing variable, and then idiosyncratic error. To use our example: an individual woman's rankikng of Bush is a function of the average ranking of Bush,) the effects of "woman-ness" on ranking of Bush, and c) unique and idiosyncratic features.

ii) The difference of means gives us a **T-TEST** statistic. This allows you to say:

(1) I can state with (1-p value)% confidence that this group is (higher/lower) than the average.

CROSSTABULAR ANALYSIS: A way for you to analyze two nominal variables

c) What does it mean if two variables are INDEPENDENT?
$P(X \cap Y) = P(X) + P(Y)$

In tabular form:

	Male	**Female**	
Democrat	**8**	**8**	**8**
Republican	**4**	**4**	**8**
	12	**12**	

Are gender and party independent? See calculations in class. If they are NOT independent, than there is some sort of systematic effect of Gender on Party (or Party on Gender, but that is a silly hypothesis).

Essentially, we ask "how much do the percentages in the internal cells deviate from what we would expect given the marginals"? (Note like t-test: does categorizing help us discriminate among groups?)

The statistical technique for testing for independence in a crosstab is again the **"Chi-square" or χ^2**. This statistic has a probability value associated with it, just like the t-value that we saw in the difference of means. There are other measures reported in your SDA output; these are interpreted similar to the correlation coefficient, a topic for next week.

4) **ANOVA:** another way for us to calculate differences among groups. For the difference of means to work, your DEPENDENT VARIABLE must be continuous or interval; but your INDEPENDENT variable can be nominal. Difference of means and ANOVA are, for all practical purposes, identical.

b) These statistics are based on the concept of **TOTAL VARIANCE**, just like the difference of means.

c) ANOVA gives you a **Chi-square** can also be looked up in your book (pg. 458).

i) This Chi-square and associated p-value tells us:

ii) *"The probability that there is a statistically significant relationship between X and Y is (1-p value)" or*

"I can say with (1-p value)% confidence that there is a statistically discernable relationship between X and Y."

IMPORTANT CONCEPT: Difference of means, ANOVA ("Analysis of Variance) and Regression are all just variations on a theme: given 100% variance in the DEPENDENT VARIABLE, are the effects of a categorizing variable "significant"? Does the effect of the INDEPENDENT VARIABLE differ significantly from zero.

For ANOVA and difference of mean, we compare differences of means across groups; for regression, we compare the slopes of lines. See decomposition example above or 367-8 in your book.

Next Time: REGRESSION ANALYSIS