**Math 114 ACTIVITY 11:** Inference on two population proportions

**Why**

Many questions that can be answered statistically involve comparisons between two populations, and the experimental or observational data will involve two samples. We have seen the methods for testing on means. Now we extend to comparisons for proportions, where we are interested in what proportion of a population has a certain characteristic or shows a certain response to a stimulus.

**LEARNING OBJECTIVES**

1. Be able to set up a test or estimate for a difference of proportions.
2. Be able to interpret the results of a test or estimate for a difference of proportions.
3. Understand and be able to carry out the mechanics of a two-sample test for the difference of proportions
4. Understand the similarities and differences in different inference situations

**CRITERIA**

1. Success in working as a team and in fulfilling the team roles.
2. Success in involving all members of the team in the conversation.
3. Success in completing the exercises.

**RESOURCES**

1. Your text, section 11.3 and Table VI p.A-13 [critical values for $t$]
2. The team role desk markers (handed out in class for use during the semester)
3. Your class notes and especially the handout "Hypothesis testing – generalities" with the six-step outline.
4. Your calculator
5. 40 minutes

**PLAN**

1. Select roles, if you have not already done so, and decide how you will carry out steps 2 and 3 (5 minutes)
2. Work through the exercises given here - be sure everyone understands all results & procedures(30 minutes)
3. Assess the team's work and roles performances and prepare the Reflector's and Recorder's reports including team grade (5 minutes).

**DISCUSSION**

We are considering inference on the difference in two populations on the proportion of members with some specified characteristic. In this situation we observe "yes" or "no" on each member of one sample from each of the populations and we wish to make statements about the difference in the proportion of "yes" individuals between the two populations. [Either estimating the difference – with a confidence interval – or deciding whether we have evidence of a difference – with a test]

We will approximate the difference [in *population* proportions] $p_1 - p_2$ with the difference in sample proportions $\hat{p}_1 - \hat{p}_2$. If sample sizes are large enough ($n_1 * p_1 * (1 - p_1) \geq 10, n_2 * p_2 * (1 - p_2) \geq 10$), then the distribution of values of $\hat{p}_1 - \hat{p}_2$ will be approximately normal, with standard deviation (if our samples are independent) $\sqrt{\frac{p_1(1-p_1}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.
Thus for our inference:
To check for large enough samples to use $Z$: we need to find $\hat{p}_1 = \frac{X_1}{n-1}$ and $\hat{p_2} = \frac{X_2}{n_2}$ and be sure that both $n_1 * \hat{p}_1 * (1 - \hat{p}_1)$ and $n - 2 * \hat{p_2} * (1 - \hat{p}_2)$ are at least 10.

A.) To estimate the difference $p_1 - p_2$ with confidence $1 - \alpha$ we will use the interval $\hat{p}_1 - \hat{p}_2 - Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ to $\hat{p}_1 - \hat{p}_2 + Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ [That is, the error allowance $E$ is $Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$] and we get results like "With 90% confidence we say the proportion of — in (the first population) is is between — and —- larger [or smaller, if we have negatives] than the proportion in (the second).

B.) Our tests are of the form:
$H_0 : p_1 - p_2 = 0$ (also written $p_1 = p_2$)
$H_1 : p_1 - p_2 \neq 0$ (may be "<" or ">" – usual three possibilities) (also written $p_1 \neq p_2$)
Our sample statistic is sample $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$; here $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$ is the combined ("pooled") estimate of the population proportion (they are the same, if $H_0$ is true)

NOTICE THAT in the test situation we combine the two proportions to give an estimate $\hat{p}$ of the "common proportion" because our null hypothesis says there is a common proportion, but for estimation of the difference we do not combine these—we use $\hat{p}_1$ and $\hat{p}_2$ individually in the formula for standard error (because we are *not* working form an assumption that $P_1$ and $p_2$ are the same).

We are limiting our tests to seeing whether the proportions are different (in the right direction)—we are not testing for a particular size difference.
Rejection criteria are exactly the same as for the single proportion $Z$-test:
For "≠" alternative, we reject $H_0$ if sample $Z < -Z_{\alpha/2}$ or if sample$Z > Z_{\alpha/2}$
For ">" alternative, we reject $H_0$ if sample $Z > Z_\alpha$
For "<" alternative, we reject $H_0$ if sample $Z < -Z_\alpha$
Finding a range for the $p$-value works exactly as it does for the test on one proportion.

**MODEL**

A large automobile insurance company is testing to determine whether there is a difference in claim rates between its single and married male policyholders. They have taken a sample of 400 single policyholders and found 56 filed a claim within the last 3 years and a sample of 900 married policyholders, of whom 90 filed a claim within the last three years.

a. Do these data show, at the .05 level, that there is a difference between the claim rates of single and married (male) policyholders?
b. What is our 95% confidence estimate of the difference in claim rates?

a. The first question is a "Do we have evidence of a difference?" question - a Test.

**I** Populations - single policyholders.
$p_1$ = proportion of single male policyholders who file a claim
$p_2$ = proportion of married male policyholders who file a claim Test is :
$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$

**II** $\hat{p}_1 = \frac{56}{400} = .14$, so $n_1 * \hat{p}_1 * (1 - \hat{p}_1) = 400 * .14 * .86 = 48.16 \geq 10$ and $\hat{p_2} = \frac{90}{900} = .1$ so $n_2 * \hat{p}_2 * (1 - \hat{p}_2) = 900 * .10 * .90 = 81 \geq 10$ so we can treat the test statistic as normally distributed. Test statistic is sample $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$.

**III** Reject $H_0$ if sample $Z > Z_{.025} = 1.960$ or if sample $Z < -Z_{.025} = -1.960$

**IV** $\hat{p} = \frac{56+90}{400+900} = .112$ so $Z = \dfrac{(.14 - .10) - 0}{\sqrt{\frac{.112*.888}{400} + \frac{.112*.888}{900}}} = 2.111$ which is greater than 1.960.

**V** Reject $H_0$ and support $H_1$ (Using the table, our $p$ value is between 2*.01 and 2*.02)

**VI** The sample does give evidence ($.02 < p < .04$) at the .05 level that there is a difference in claim rates (proportion of policyholders that file a claim) for single and married male policyholders. . [Note: Using Minitab gives p = .035]

b. The second question asks for an estimate of the difference in the proportions.
The estimate will be $\hat{p}_1 - \hat{p}_2 \pm E$ with $E = Z_{.025}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. Our error allowance is $E = 1.960 sqrt \dfrac{.14(.86)}{400} + \dfrac{.1(.9)}{900} = .039$ and $\hat{p}_1 - \hat{p}_2 = .14 - .10 = .04$.
With 95% confidence, we say the proportion of single male policyholders who file claims is between .001 and .079 greater than the proportion of married male policyholders who file claims.
(Note the statement for the confidence interval does indicate *which* proportion is larger—does *not* simply say "the difference is")

**EXERCISES**

Carry out (show all steps) for tests, and give the $p$-value. For estimates, be sure the "larger" or "smaller " (or "less" or "more") are given by your answer.

1. In a study that was described by a writer for the Washington Post (San Luis Obispo Tribune, February 2, 2000), 500 patients undergoing abdominal surgery were randomly assigned to breathe one of two oxygen mixtures during surgery and for 2 hr afterward. One group received a mixture containing 30% oxygen, a standard generally used in surgery. The other group was given 80% oxygen. Wound infections developed in 28 of the 250 patients Who received 30% oxygen and in 13 of the 250 patients receiVed 80% oxygen.

   (a) Is there sufficient evidence to conclude that the proportion of patients who develop wound infections is lower for the 80% oxygen treatment than for the 30% oxygen treatment? Test the relevant hypotheses using a significance level of .05

   (b) Give a 90% confidence estimate for the difference in the proportion of patients who develop wound infections for the two treatments. (be sure you indicate direction—which is higher/lower—you may have to allow for both possibilities).

2. In December 2001 the Department of Veterans Affairs announced that it Would begin paying benefits to soldiers suffering from Lou Gehrig's disease who had served in the first Gulf War (New York Times, December 11, 2001). This decision was based on an analysis in which the disease's incidence rate (the proportion developing the disease) for all of the approximately 700,000 soldiers sent to the Persian Gulf between August 1990 and July 1991 was compared to the incidence rate for all of the approximately 1.8 million other soldiers who were not in the Persian Gulf during this time period. Based on these data, explain why it is not appropriate to perform a formal inference procedure (such as the two-sample z test) and yet it is still reasonable to conclude that the incidence rate is higher for Gulf War veterans than for those who did not serve in the Gulf War.

3. Are college students who take a 3-credit freshman orientation course more or less likely to stay in college than those who do not take such a course? The article "A Longitudinal Study of the Retention and Academic Performance of Participants in Freshman Orientation Courses" (*Journal of College Student Development* [1994]: 444-449) reported that 50 of 94 randomly selected students who did not participate in an orientation course returned for a second year. Of 94 randomly selected students Who did take the orientation course, 56 returned for a second year.

   (a) Does this difference give evidence that a higher proportion of students who are in an orientation course will return for a second year?

   (b) Construct a 95% confidence interval for the difference in the proportion returning for students who do not take an orientation course and those who do. Give an interpretation of this interval. (This will be hard to phrase)

**READING ASSIGNMENT** (in preparation for next class)
Section 11.4 - Sorting out the methods

**SKILL EXERCISES:**(hand in - individually - at next class meeting) Sullivan **p.544** #14 - 19