AP Statistics: Chapter 12.2: Inference for TWO Proportions

In two-sample problems, we want to compare responses ot two independent samples. In chapter 11, we compare two means using a two-sample *t* procedures. In chapter 15, we will compare two standard deviations using an F statistic. In a **two-sample proportion problem** (this section) we want to compare two populations or the responses of two different treatments based on two independent samples.

We will now develop methods to compare the proportions of "successes" in two groups.

We will use subscripts to denote the information coming from each of the two populations.

Population	Population proportion	Sample size	Sample proportion
1	p_1	n_1	$\hat{p}_1$
2	p_2	n_2	$\hat{p}_2$

We typically compare populations by drawing inferences about the DIFFERENCE $p_1 - p_2$ between the population proportions. Of course, we're not going to know the true values of these, so we use the test statistic that estimates this difference, namely $\hat{p}_1 - \hat{p}_2$ (duh!)

Our new test statistic $\hat{p}_1 - \hat{p}_2$ has its own sampling distribution. Here's what you need to know (already know) about them in order to do inference correctly.

- The mean of the sampling distribution for $\hat{p}_1 - \hat{p}_2$ is $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$. $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$.
- The variance of $\hat{p}_1 - \hat{p}_2$ is the sum of the variances of $\hat{p}_1$ and $\hat{p}_2$. (Remember we can't add standard deviations). We get $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$
- When the sample size is large, the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

Each sample must be taken from independent random samples. Populations must still be at least $10n_1$ and $10n_2$. The only difference from single proportions is that $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, and $n_2\hat{q}_2$ must only be 5 or bigger (not 10 or bigger). Be sure to check them for both.

Calculating confidence intervals and doing significant tests will have the same feel as before, just with different equations. Here they are:

Confidence intervals for $\hat{p}_1 - \hat{p}_2$	Significance tests for $\hat{p}_1 - \hat{p}_2$
$\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$ Again, we're using our sample proportions $\hat{p}_1$ and $\hat{p}_2$ as approximations of the true population proportions p_1 and p_2 respectively.	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ Where $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ and is called the **pooled** proportion

Example 1: A study was conducted to determine the effect of preschool on later use of social services. It identified the proportion of two groups who needed social services later in life. The data is as follows:

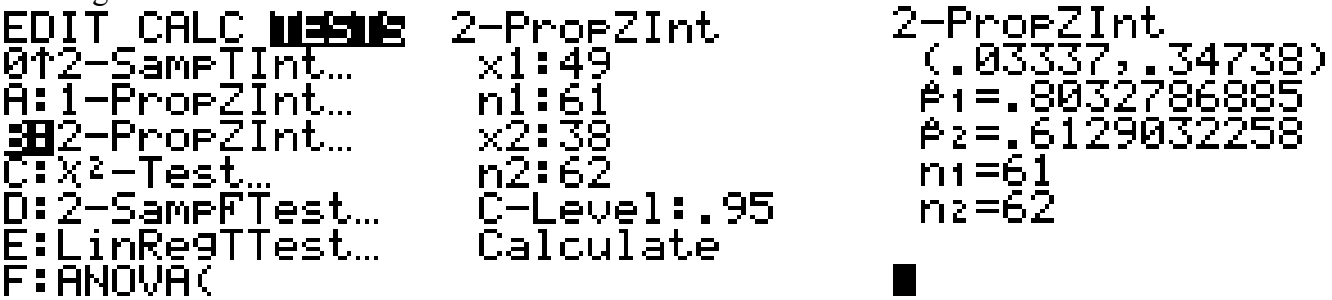
Population	Population Description	Sample Size	Number needing Service	Sample Proportion
1	Control	$n_1 = 61$	49	$\hat{p}_1 = 49/61 = 0.803$
2	Preschool	$n_2 = 62$	38	$\hat{p}_2 = 38/62 = 0.613$

Find a 95% confidence interval. First check the assumptions (It's a big drag, but you MUST show this step.)
Assumptions:

- Our distribution is approximately normal by the Central Limit Theorem because each sample size is larger than 30.
- We'll assume both samples were take from a random sample of the populations of people who attended preschool and those who didn't (the control).
- We'll also assume both populations of interest are at least 610 (for control) and 620 (for others).
- $n_1\hat{p}_1 = (61)(.803) = 49$, $n_1\hat{q}_1 = (61)(.197) = 12$, $n_2\hat{p}_2 = (62)(.613) = 38$, and $n_2\hat{q}_2 = (62)(.387) = 24$. All of these numbers are greater than 5, so our inference results will be sound.

Using the following equation for a 95% confidence interval with $z^* = 1.960$, $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$,

or using the Calculator:



We get our interval of $(0.033, 0.347)$.

Conclusion: I am 95% confident that the percent needing social services is between 3.3% and 34.7% lower among people who attended preschool.

Significance tests for $p_1 - p_2$

This is where things get a bit different. Try to follow the logic, and it will make sense.

Like before, we set up a hypothesis test. Our null hypothesis says EITHER that the difference of our two proportions is zero, but it is more common (and easier) to say the two proportions are the same. That is

$$H_0 : p_1 - p_2 = 0 \text{ or } H_0 : p_1 = p_2$$

Remember that for significance tests, we use p_1 and p_2, the true population proportions, and NOT $\hat{p}_1$ and $\hat{p}_2$. Significance tests make some claim about the **populations**, not the samples. **You will not know these values, so the hypothesis will be in terms of p_1 and p_2**

The alternative hypothesis states the kind of difference between the two population proportions we expect, or what we are testing for, namely

$$H_a : p_1 > p_2,\ H_a : p_1 < p_2, \text{ or } H_a : p_1 \neq p_2$$

In order to perform a significance test, we use a **pooled** sample proportion. Why pooled? Well, if our null hypothesis is true, then both samples come from a single population with a certain unknown proportion p. We act as if this is the case, so we combine the two samples and examine a "new" collective $\hat{p}$.

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

We use this pooled $\hat{p}$ in place of $\hat{p}_1$ and $\hat{p}_2$ in the formula for the standard error (SE). We use this to get a z statistic that has the standard normal distribution when H_0 is true. So here's the formula for the z test statistic when testing $H_0 : p_1 = p_2$.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Once we find this z test statistic (in fact, we most often find it with the calculator), we use it exactly as you'd expect. We must still check that $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, and $n_2\hat{q}_2$ be 5 or bigger.

Example 2: The Helsinki Heart Study wished to find out if a drug used to lower blood cholesterol would reduce heart attacks. They randomly assigned 2051 middle-aged men to a group that took gemfibrozil to reduce cholesterol and 2030 men to a placebo group. During the next 5 years, 56 men in the gemfibrozil group had heart attacks while 84 men in the placebo group did. Did the gemgibrozil help reduce heart attacks in those that took it?

Solution:

- State: "We will use a "Two-sample proportion z test."
- Calculate and define your proportions.

$$\hat{p}_1 = \frac{56}{2051} = 0.0273 \quad \text{(gemfibrozil group)}$$

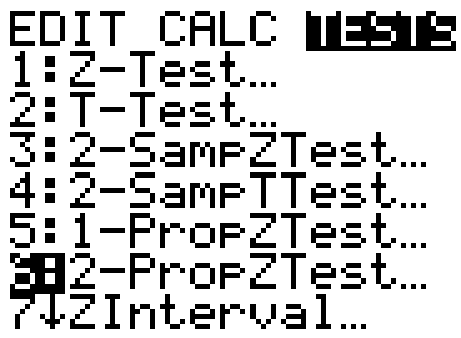
$$\hat{p}_2 = \frac{84}{2030} = 0.0414 \quad \text{(placebo group)}$$

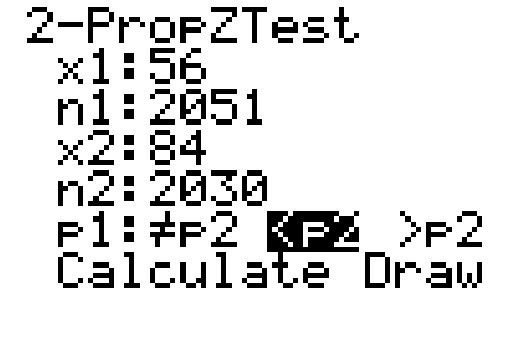
$$\hat{p} = \frac{56 + 84}{2051 + 2030} = \frac{140}{4081} = 0.0343$$

- Set up the null and alternative hypotheses:

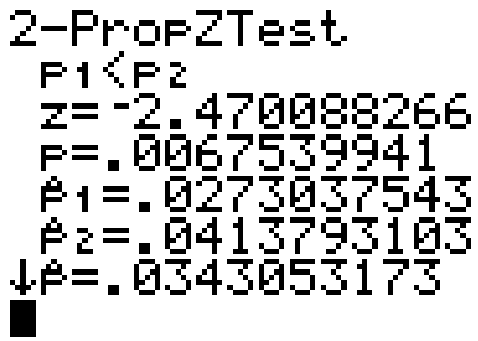
  $H_0 : p_1 = p_2$ (the two populations had heart attacks at same proportion)

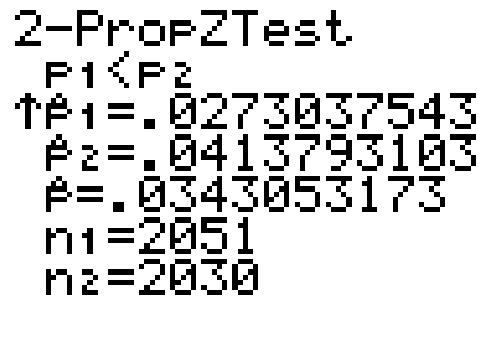
  $H_a : p_1 < p_2$ (the gemfibrozil group has smaller proportion of heart attacks).
- Check your $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, and $n_2\hat{q}_2$ and make sure they're 5 or bigger. Then state your other assumptions (normality, SRS, and random sample).
- Go to the calculator (STAT $\rightarrow$ TESTS)



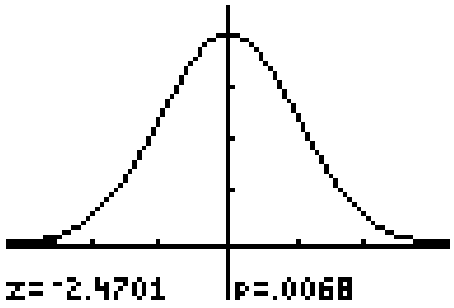


choosing "Calculate" give





or choosing "Draw" gives



- State your results:

$$z = -2.470 \text{ and } p = 0.0068$$

- Interpret your results and write your conclusion.
  Since our *p*-value of 0.0068 is less than 0.01, the results are statistically significant at the 1% ($\alpha = 0.01$) level. There is strong evidence that gemfibrozil reduced the rate of heart attacks. The large samples in the Helsinki Heart study helped the study get highly significant results.