AP Statistics – Lesson Notes - Chapter 22: Comparing Two Proportions

Comparing Two Proportions is more common than One Proportion

Previous example: The DV counseling department records show that, in the past, 78% of DV seniors attend college in-state. We ask an SRS of 200 seniors this year and 168 of them say they are staying in-state. *Do we have reason to believe that more seniors are staying in-state this year?* Comparing 1 proportion to a value

Example: This year, 172 of a sample of 200 seniors stayed in-state. Last year, 114 of a sample of 150 seniors stayed in-state for college. *Do we have reason to believe that more seniors are staying in-state this year? Is this difference in proportions significant?*
Comparing 2 proportions to each other

Although it is possible we have a known or suspected population proportion to compare a sample to, it is more common that we have two sample proportions and want to compare them to each other. In experiments, we might need to compare the proportion of samples which show the effect of the experimental treatment between a treatment group and a control group.

Null and Alternative Hypotheses

Comparing 1 proportion to a value

$H_0: p = .78$
$H_A: p > .78 \ (or\ p < .78\ or\ p \neq .78)$

Comparing 2 proportions to each other

$H_0: p_1 = p_2$
$H_A: p_1 > p_2 \ (or\ p_1 < p_2\ or\ p_1 \neq p_2)$

or

$H_0: p_1 - p_2 = 0$
$H_A: p_1 - p_2 > 0 \ (or\ p_1 - p_2 < 0\ or\ p_1 - p_2 \neq 0)$

The hypotheses are now comparing two proportions instead of comparing one proportion to a value.

The statistic (estimate)

Comparing 1 proportion to a value

$$\hat{p} = \frac{x}{n}$$

Comparing 2 proportions to each other

(this year) (last year)

$$\hat{p}_1 = \frac{x_1}{n_1} \qquad \hat{p}_2 = \frac{x_2}{n_2}$$

$$\hat{D} = \widehat{difference} = \hat{p}_1 - \hat{p}_2$$

The statistic (estimate) is now the difference between the 2 proportions.

The standard error

Comparing 1 proportion to a value

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Comparing 2 proportions to each other

$$SE_{\hat{D}} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

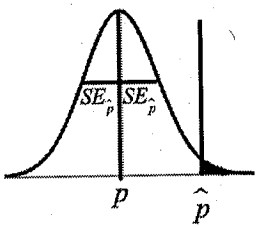
The standard error now must account for the variability of the difference. Similar to the speed dating example when there are two sources of variation (varying independently) the variances add:

$$\left(SE_{\hat{D}}\right)^2 = \left(SE_{\hat{p}_1}\right)^2 + \left(SE_{\hat{p}_2}\right)^2$$

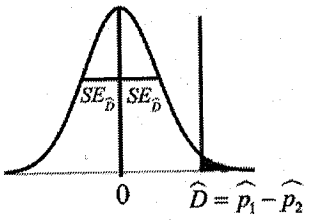
(A more detailed derivation is available on www.mrfelling.com below the filled-in notes)

The p-value

Comparing 1 proportion to a value



Comparing 2 proportions to each other



The p-value is always the probability of an estimate (statistic) being as far as observed in the sample (or farther) from the null hypothesis value as is observed in this sample. The null hypothesis for comparing two proportions is that there is no difference, so it is always 0.

We use normalcdf to compute the p-value.

The conclusion

Comparing 1 proportion to a value

With $\alpha = .05, p = .0203$ is low,
so we reject H_0
We have sufficient evidence to conclude that the percentage of students at DV staying in state for college this year is higher than last year.

Comparing 2 proportions to each other

With $\alpha = .05, p = .0095$ is low,
so we reject H_0
We have sufficient evidence to conclude that the percentage of students at DV staying in state for college this year is different than last year (this year is higher).

How much higher? We could provide a confidence interval...

Comparing 1 proportion to a value

$$CI = (statistic) \pm (margin\ of\ error)$$

$$CI = \hat{p} \pm (z^*)(SE_{\hat{p}})$$

$$CI = \hat{p} \pm (z^*)\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

Comparing 2 proportions to each other

$$CI = (statistic) \pm (margin\ of\ error)$$

$$CI = \hat{D} \pm (z^*)(SE_{\hat{D}})$$

$$CI = (\hat{p}_1 - \hat{p}_2) \pm (z^*)(SE_{\hat{p}_1 - \hat{p}_2})$$

The 'test statistic' vs. the 'statistic'

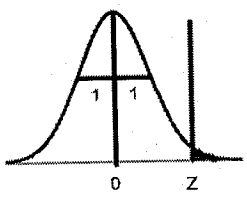
Whether we are talking about variation of a single proportion about an expected value, or variation of a difference of proportions about an expected difference of zero, we are always using a Normal distribution.

We could 'standardize' the distribution and the statistic. In both cases, the statistic would become a z-score. If we do this, this is referred to as computing the 'test statistic' and we would use a standardized Normal distribution to compute p-value:

Comparing 1 proportion to a value

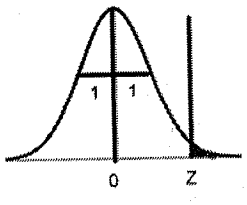
$Z =$

$$\boxed{Z = \frac{x - \mu}{\sigma}}$$



Comparing 2 proportions to each other

$Z =$



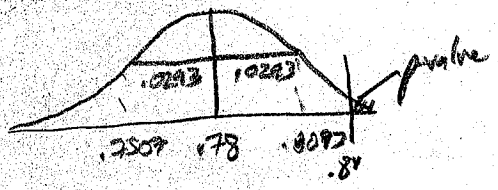
Previous example: The DV counseling department records show that, in the past, 78% of DV seniors attend college in-state. We ask an SRS of 200 seniors this year and 168 of them say they are staying in-state. *Do we have reason to believe that more seniors are staying in-state this year?*

Comparing 1 proportion to a value

$H_0: p = .78$
$H_A: p > .78$

$\hat{p} = \frac{168}{200} = .84$

$SE_{\hat{p}} = \sqrt{\frac{(.78)(.22)}{200}} = .0293$



p-value = normalcdf(.84,999,.78,.0293)
= .0203

With $\alpha = .05$, $p = .0203$ is low so we reject H_0.
We have sufficient evidence to conclude that the percentage of seniors staying in state this year is higher than the 78% historical value.

95% confidence interval ($z^* = 1.96$ for 95% CI)

$CI = \hat{p} \pm z^* SE_{\hat{p}}$
$= .84 \pm (1.96)(.0293)$
$= .84 \pm .0574$
$= (.783, .897)$

We are 95% confident that the percentage of all DV seniors staying in state is between 78.3% and 89.7%.

Example: This year, 172 of a sample of 200 seniors stayed in-state. Last year, 114 of a sample of 150 seniors stayed in-state for college. *Do we have reason to believe that more seniors are staying in-state this year? Is this difference in proportions significant?*

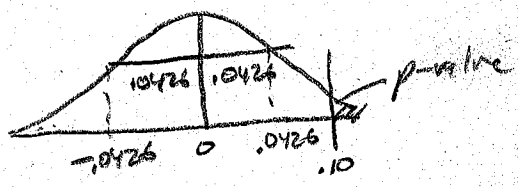
Comparing 2 proportions to each other

$H_0: p_T - p_L = 0$
$H_A: p_T - p_L > 0$

$\hat{p}_T = \frac{172}{200} = .86$ $\hat{p}_L = \frac{114}{150} = .76$

$\hat{D} = .86 - .76 = .10$

$SE_{\hat{D}} = \sqrt{\frac{(.86)(.14)}{200} + \frac{(.76)(.24)}{150}} = .0426$



p-value = normalcdf(.10,999,0,.0426)
= .0095

With $\alpha = .05$, $p = .0095$ is low so we reject H_0.
We have sufficient evidence to conclude that the percentage of seniors staying in state this year is higher than last year.

95% confidence interval ($z^* = 1.96$ for 95% CI)

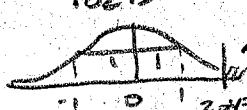
$CI = (\hat{p}_T - \hat{p}_L) \pm z^* SE_{\hat{D}}$
$= .10 \pm (1.96)(.0426)$
$= .10 \pm .0835$
$= (.017, .184)$

We are 95% confident that the percentage of all DV seniors staying in state is between 1.7% and 18.4% higher this year than last year.

---

test statistic:

$z = \frac{.84 - .78}{.0293} = 2.0478$



p value = normalcdf(2.0478,999,0,1)
= .0203

$z = \frac{.10 - 0}{.0426} = 2.3474$

pvalue = normalcdf(2.3474,999,0,
= .0095

Combining samples to obtain SE by pooling

In this last analysis, we could also have stated $H_0 : p_1 = p_2$ because we are assuming that the difference is zero and that the variation of the difference is Normal centered at zero.

When we did z-test for 1-proportion we used standard deviation rather than standard error because we had a known (or assumed) population proportion. Here, with just samples we use standard error, but the Null Hypothesis is stating that we are assuming the sample proportions are equal. Because of this assumption, when computing the standard error, we could treat both samples as a single, larger, sample instead of separate samples p_1. This is known as ***pooling***. This usually doesn't have a large effect on the results, but you could argue that by pooling the samples, our SE is likely to be a better estimate of the true standard deviation:

$$\hat{p}_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} \quad \left(where \quad x_1 = \hat{p}_1 n_1 \quad and \quad x_2 = \hat{p}_2 n_2\right)$$

$$SE_{diff,pooled} = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}}$$

Pooled SE is sometimes (but not always) small - smaller usually when sizes of the samples are very different.

Reworking the previous example, with pooling

Example: This year, 172 of a sample of 200 seniors stayed in-state. Last year, 114 of a sample of 150 seniors stayed in-state for college. *Do we have reason to believe that more seniors are staying in-state this year? Is this difference in proportions significant?*

$H_0 : p_1 - p_2 = 0$

$H_A : p_1 - p_2 > 0 \ \ (or \ p_1 - p_2 < 0 \ \ or \ p_1 - p_2 \neq 0)$

$$p_1 = \frac{x_1}{n_1} = .86 \qquad p_2 = \frac{x_2}{n_2} = .76$$

$$\hat{D} = .86 - .76 = .10$$

$$x_1 = \hat{p}_1 n_1 = (.86)(200) = 172 \qquad x_2 = \hat{p}_2 n_2 = (.76)(150) = 114$$

$$\hat{p}_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{172 + 114}{200 + 150} = .8171$$

$$SE_{diff,pooled} = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}} = \sqrt{\frac{(.8171)(.18286)}{200} + \frac{(.8171)(.18286)}{150}} = .0418$$

p-value = normalcdf(.10, 999, 0, .0418) = .0084

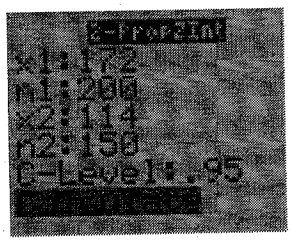
AP Reader says: "Always pool for 2-proportion hypothesis tests"

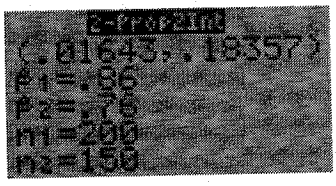
Calculator Functions

There are calculator functions for 2 proportion z-test and 2 proportion confidence interval:

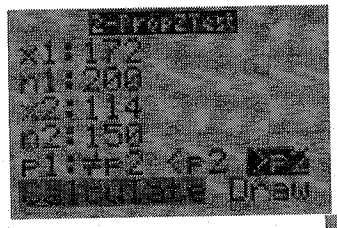
Example: Last year, 76% of a sample of 150 seniors stayed in-state for college. This year, 86% of a sample of 200 seniors stayed in-state. *Do we have reason to believe that more seniors are staying in-state this year? Is this difference in proportions significant?*

2-PropZInt

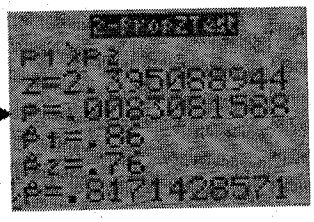


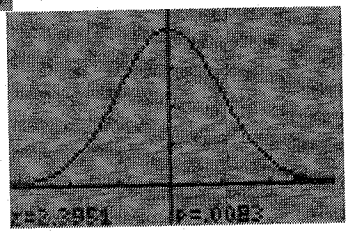


2-PropZTest



p-value →





(The calculator always performs a <u>pooled</u> calculation for 2 proportions)

Assumptions / Conditions *(Always check conditions first!)*

<u>SRS</u>	**<u>independent</u>**	**<u>n<10%population</u>**	**<u>success/fail</u>**
true random sample or believe sample represents population without bias	2 groups must vary independently of each other	Now $n_1+n_2<10\%$ of the total population (or must know that each sample has same probability of success, e.g. coin tosses)	$n_1\hat{p}_1 \geq 10$ $n_1\hat{q}_1 \geq 10$ $n_2\hat{p}_2 \geq 10$ $n_2\hat{q}_2 \geq 10$