12.3: Inferences for Two Population Proportions

Often, instead of comparing a sample's proportion to a benchmark value, we want to compare the proportions of two different samples, drawn from two different populations.

If the sample sizes are sufficiently large, we can assume that the difference between the two sample proportions, $\hat{p}_1 - \hat{p}_2$, is normally distributed.

Suppose we have two samples, Sample A and Sample B, of sizes n_1 and n_2, respectively.
In Sample A, x_1 is the number of the n_1 observations that possess the characteristic of interest.
In Sample B, x_2 is the number of the n_2 observations that possess the characteristic of interest.

Then the sample proportions are $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$.

We'll follow this rule of thumb: For sample sizes to be sufficiently large to assume the difference in proportions is normally distributed, all of the quantities x_1, $n_1 - x_1$, x_2, $n_2 - x_2$ should be at least 5.
Another way to say this: All four cells (groups) in a contingency table must contain at least 5 data points.

Example 1: Suppose we are comparing the number of successful treatments for two groups. Treatment A is given to 20 patients, and is successful in 16 patients. Treatment B is given to 18 patients, and is successful in 12 patients. Can we assume a normal distribution for the difference between proportions?

	Successful	Not successful	Total
Treatment A	16	4	20
Treatment B	12	6	18
Total	28	10	

This is less than 5

No, we cannot assume the difference between proportions is normally distributed, because there are less than 5 people in the Not Successful, Treatment A category

Sampling distribution for the difference between two sample proportions:

Suppose independent samples of sizes n_1 and n_2 are taken from two populations, in which the proportions of observations possessing a characteristic are p_1 and p_2.

Then the mean and standard deviation of the difference between sample proportions are:

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$$

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \qquad \text{(standard error)}$$

Also, the shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, provided that the sample sizes are sufficiently large.
Rule of thumb: to assume the difference between sample proportion is normally distributed, all of x_1, $n_1 - x_1$, x_2, $n_2 - x_2$ should be at least 5.

However, we will almost never know the population proportions p_1 and p_2, so we cannot calculate the above mean and standard deviation.

Instead, we use statistics from the samples to calculate point estimates for $\mu_{\hat{p}_1-\hat{p}_2}$ and $\sigma_{\hat{p}_1-\hat{p}_2}$.

Point estimates for the mean and standard deviation of the difference between sample proportions:

$\hat{p}_1 - \hat{p}_2$ is an unbiased estimate for $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$.

To calculate a point estimate for $\sigma_{\hat{p}_1-\hat{p}_2}$, we first calculate the pooled sample proportion $\hat{p}_p$:

$$\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2},$$

where n_1 and n_2 are the sample sizes, and x_1 and x_2, respectively, are the numbers of observations possessing the characteristic of interest. Then the standard error is approximately

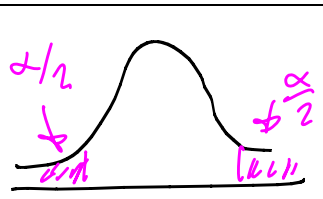
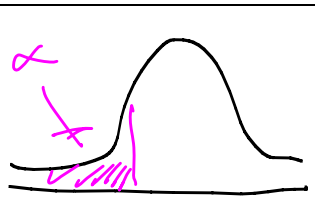
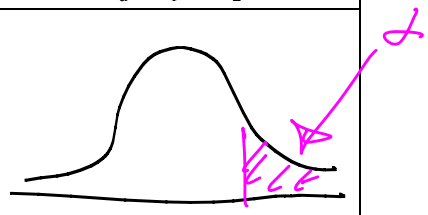
$$\sigma_{\hat{p}_1-\hat{p}_2} \approx \sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \quad = \sqrt{\frac{\hat{p}_p(1-\hat{p}_p)}{n_1} + \frac{\hat{p}_p(1-\hat{p}_p)}{n_2}}$$

Hypothesis Testing for Two Population Proportions:

Step 1: Determine the significance level α.

Step 2: Check that the assumptions are satisfied.
- Simple random sample
- Samples are independent (observations from one sample are not paired with observations from the other sample).
- x_1, $n_1 - x_1$, x_2, $n_2 - x_2$ are all at least 5.

Step 3: Determine the null and alternative hypotheses.

Two-Tailed Test (most common)	Left-Tailed Test (rare)	Right-Tailed Test (rare)
$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$	$H_0 : p_1 = p_2$ $H_a : p_1 < p_2$	$H_0 : p_1 = p_2$ $H_a : p_1 > p_2$
		
Rejection Region	Rejection Region	Rejection Region

Note: One tailed tests assume that the scenario not listed ($p_1 < p_2$ for a left-tailed test or $p_1 > p_2$ for a right-tailed test) is not possible or is of zero interest.

Step 4: Using your α level and hypotheses, sketch the rejection region.

Step 5: Use a normal curve table to determine the critical value for z associated with your rejection region.

Step 6: Compute the test statistic: $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.$$ $= \dfrac{\hat{p}_1 - \hat{p}_2}{\text{std error}}$

Step 7: Determine whether the value of z calculated from your sample (in Step 6) is in the rejection region.
- If z is in the rejection region, reject the null hypothesis.
- If z is not in the rejection region, do not reject the null hypothesis.

Step 8: State your conclusion and interpret the results of the hypothesis test.

Example 2: Suppose that a clinical trial for two different cancer drugs is conducted. For drug A, 637 of 2095 patients were cured. For drug B, 702 of 2119 patients were cured. Does this trial provide evidence that Drug B cures a higher percentage of patients than Drug A? Use a 5% level of significance.

	Cured	Not cured
Drug A	637	1458 (2095−637)
Drug B	702	2119−702 = 1417

one-tailed test

All cells ≥ 5, so the difference between proportions is normally distributed.

Drug A: Pop. 1
Drug B: Pop 2

H_0: $p_1 = p_2$

H_a: $p_1 < p_2$

$\alpha = 0.05$

(significance level of 5% corresponds with confidence level of 95%)

(Significance level is the percentage version of α)

$-z_{0.05} = -1.645$

From tables, critical value is $-z_{0.05} = -1.645$

Sample info

Drug A: $n_1 = 2095$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{637}{2095} \approx 0.304$

Drug B: $n_2 = 2119$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{702}{2119} \approx 0.331$

Calculate pooled sample proportion

$\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{637+702}{2095+2119}$

$= \frac{1339}{4214} \approx 0.31775$

$\hat{q}_p = 1 - \hat{p}_p = 1 - 0.31775$

$= 0.68225$

12.3.5

Standard error:

$$\sigma_{\hat{p}_1-\hat{p}_2} \approx \sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= \sqrt{\hat{p}_p\hat{q}_p}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{\hat{p}_p\hat{q}_p\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= \sqrt{0.31775 * 0.68225 * (1/2095 + 1/2119)}$$

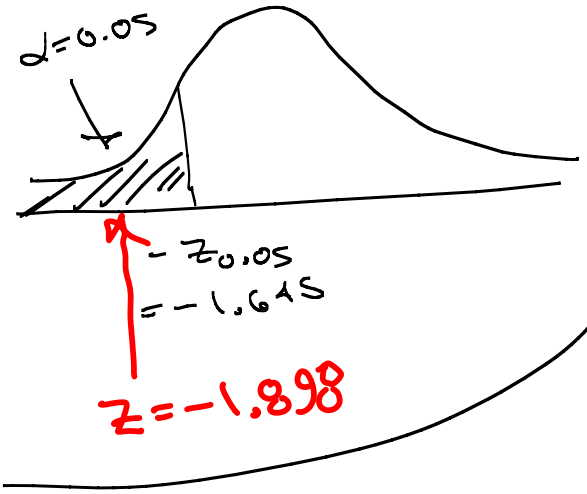
$$\approx \sqrt{2.05789 \times 10^{-4}}$$

$$\approx 0.014345$$

To get this, used exact values for $\hat{p}_1$ and $\hat{p}_2$

Calculate test statistic

$$z = \frac{\hat{p}_1-\hat{p}_2}{\text{std error}} = \frac{0.304-0.331}{0.014345} \approx -1.898$$



This falls in the rejection region, so we **Reject H_0**

If using rounded versions of $\hat{p}_1 = 0.304$ and $\hat{p}_2 = 0.331$, then $z = -1.882$ (still reject H_0)

This sample provides evidence that Drug B cures a higher proportion of patients than Drug A.