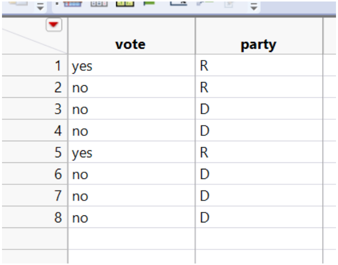
DSA 8010 - Inference on two proportion

Two binary variables

Two binary variables

Data collected on two binary variables will come as a spreadsheet with columns corresponding to the two binary outcomes or as a two-way table of counts.



	vote	party
1	yes	R
2	no	R
3	no	D
4	no	D
5	yes	R
6	no	D
7	no	D
8	no	D

		vote		
		Yes	No	Total
party	Democrat	0	5	5
	Republican	2	1	3
	Total	2	6	8

Two binary variables

Remember from Unit 1 that row or column proportions can be calculated to look for association between the "row" variable and "column" variable in a contingency table.

Table from the previous slide with row proportions calculated:

		vote		
		Yes	No	Total
party	Democrat	0	1	1
	Republican	0.667	0.333	1

In limited samples, how can we discern whether differences in proportions are due to change or to real association between the variables?

Two binary variables

One variable is a grouping variable that divides the observational units into two groups. The other is an outcome variable indicating whether some uncertain outcomes occurred on each observational unit.

It is conventional to use the grouping variable as the row variable in a two-way table, but this may not always be the case.

		vote		
		Yes	No	Total
party	Democrat	0	5	5
	Republican	2	1	3
	Total	2	6	8

Statistical model for binary data

Let $y_{i1}, \quad i = 1, \ldots, n_1$, be n_1 observed measurements of a binary variable from group 1.

Let $y_{i2}, \quad i = 1, \ldots, n_2$, be n_2 observed measurements of a binary variable from group 2.

Statistical model:
$y_{11}, \ldots, y_{n_1 1}$ are i.i.d. realizations from a Bernoulli(π_1) distribution.
$y_{12}, \ldots, y_{n_2 2}$ are i.i.d. realizations from a Bernoulli(π_2) distribution.

Confidence interval for $\pi_1 - \pi_2$

Inference on two proportions

The inferential goal is to compare the probabilities of success across the two groups. This is represented by the parameter $\pi_1 - \pi_2$ (or, equivalently, $\pi_2 - \pi_1$).

- If $\pi_1 - \pi_2$ is near zero, the probability of success is similar between the groups.
- If $\pi_1 - \pi_2$ is far from zero, the two groups differ in their probability of success. This indicates that there is association between the grouping variable and the outcome variable.

Sampling distribution of $\widehat{\pi}_1 - \widehat{\pi}_2$

For for large-sample interval for $\pi_1 - \pi_2$, the point estimate is the difference between the sample proportions.

$$\widehat{\pi}_1 - \widehat{\pi}_2 = \frac{\sum_{i=1}^{n_1} Y_{i1}}{n_1} - \frac{\sum_{i=2}^{n_1} Y_{i2}}{n_2}.$$

- The standard error of $\widehat{\pi}_1 - \widehat{\pi}_2$ is $\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$.
- If $n\pi$ and $n(1-\pi)$ are not too small, then the approximate sampling distribution of $\widehat{\pi}_1 - \widehat{\pi}_2$ is $N\left(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}\right)$.

Large-sample confidence interval for $\pi_1 - \pi_2$

A large-sample $(1-\alpha)\cdot 100\%$ confidence interval for $\pi_1 - \pi_2$ is

$$\widehat{\pi}_1 - \widehat{\pi}_2 \pm z^*_{\alpha/2}\sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}}$$

- $z^*_{\alpha/2}$ is the $(1-\alpha/2)\cdot 100$th percentile of the $N(0,1^2)$ distribution.
- Use this interval when the data have at least five successes and five failures in each group.

Example (speeding violations)

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year		
	Yes	No	Total
Uses cell phone while driving	25	280	305
Does not use cell phone while driving	45	405	450
Total	70	685	755

Make a 90% confidence interval for the difference between the proportion of drivers who use a cell phone who got a speeding violation and the proportion of drivers who do not use a cell phone who got a speeding violation.

Example (speeding violations)

	Speeding violation in the last year		
	Yes	No	Total
Uses cell phone while driving	25	280	305
Does not use cell phone while driving	45	405	450
Total	70	685	755

Make a 90% confidence interval for the difference between the proportion of drivers who use a cell phone who got a speeding violation and the proportion of drivers who do not use a cell phone who got a speeding violation.

Interpretation of confidence intervals

- A study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of females and males whose favorite color is black ($\pi_{female} - \pi_{male}$) was calculated to be (-0.06, -0.02).
  We can be 95% confident that the proportion of female undergraduates whose favorite color is black is between 0.02 and 0.06 less than the proporiton of male undergraduates whose favorite color is black.

Interpretation of confidence intervals

- A study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of females and males whose favorite color is black ($\pi_{female} - \pi_{male}$) was calculated to be (-0.06, -0.02).
  We can be 95% confident that the proportion of female undergraduates whose favorite color is black is between 0.02 and 0.06 less than the proporiton of male undergraduates whose favorite color is black.

  A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($\pi_{male} - \pi_{female}$) was calculated to be (0.02, 0.06).

Interpretation of confidence intervals

- A study asked 1,924 male and 3,666 female undergraduate college students their favorite color.
- The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers in several sectors. They found a 90% confidence interval for the difference between the proportions of truck drivers and pilots who sleep less than six hours per night on average to be (-0.04,0.09).
  With 90% confidence, we can say that the proportion of truck drivers who sleep less than 6 hrs/night is between -0.04 less and 0.09 greater than the proportion of pilots who sleep less than 6hrs/night.

Hypothesis test for $\pi_1 - \pi_2$

Hypothesis test for $\pi_1 - \pi_2$.

Hypotheses. Null hypothesis: $H_0 : \pi_1 - \pi_2 = D_0$;
Alternative hypothesis: $H_a : \pi_1 - \pi_2 \neq D_0$ (or $<, >$).
Most often $D_0 = 0$.

Test statistic.

$$z_0 = \frac{\widehat{\pi}_1 - \widehat{\pi}_2 - D_0}{\sqrt{\frac{\widehat{\pi}_{pool}(1-\widehat{\pi}_{pool})}{n_1} + \frac{\widehat{\pi}_{pool}(1-\widehat{\pi}_{pool})}{n_2}}},$$

where $\widehat{\pi}_{pool} =$
(no. successes across both groups)$/(n_1 + n_2)$.

P-value. Two-sided alternative: $2 * P(Z > |z_0|)$ where Z has a $\mathrm{N}(0, 1^2)$ distribution.

Decision. Reject H_0 if the p-value is less than α.

Example (blood thinners)

Consider an experiment for patients who underwent cardiopulmonary resuscitation (CPR) for a heart attack and were subsequently admitted to a hospital. These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive a blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Example (blood thinners)

Do the data provide strong evidence that using blood thinners has an effect on survival rates? Use $\alpha = 0.01$.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Example (blood thinners)

Do the data provide strong evidence that using blood thinners improve survival rates? Use $\alpha = 0.01$.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Example (blood thinners)

Do the data provide strong evidence that using blood thinners improve survival rates? Use $\alpha = 0.01$.

R performs the test automatically using the prop.test function. Use `correct=FALSE` to avoid the continuity correction and get the same answer as the large-sample formulas.

Model checking

- The test and confidence intervals rely on a Normal approximation and should be used only when there are at least 5 successes and failures per group.
- Check that the i.i.d. assumption for both groups.

Example (promotion video)

A political research group recruited 200 volunteers and asked if they plan to vote for Candidate A. The volunteers then watched a short video promoting Candidate A and two days later, they are asked again whether they plan to vote for A.

- Before the video, 36% of the volunteers planned to vote for A. After watching the video, 43% of volunteers reported intent to vote for A.
- The research group used the methods described in this lesson to find this 95% confidence interval for the difference in the before and after proportions of voters who intend to vote for A: (-0.16556709, 0.02556709).
- They report that they are 95% confident that the proportion of voters who will vote for A after seeing the video is between 0.1656 greater and 0.0256 less than the proportion who will vote for A without seeing the video.

What is wrong here?