

Biomedical Engineering Degree

5. CATEGORICAL DATA

Felipe Alonso Atienza

✉felipe.alonso@urjc.es

🐦@FelipeURJC

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos

References

- 1 R. Bernard. *Fundamentals of Biostatistics*. Ed.: Thompson. Chapter 10
- 2 D. Díez, M Cetinkaya-Rundel and CD Barr. *OpenIntro Statistics*. Chapter 6.
- 3 J. Oakley. *MAS113 Introduction to Probability and Statistics (Part 2): Data Science*. Chapters 8, 10.

What's categorical data?

- The variable under study is not continuous but is instead **may be divided into groups**, so called categories.
 - ▶ Blood type: A, B, AB, O
 - ▶ Sex: M/F
 - ▶ Age group: 18-24, 25-30, 31-35, etc.
 - ▶ Educational level: primary school, high school, college, etc.
- They are normally represented in a **two-way table** that counts the number of observations that fall into each group for two variables

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

- Do not confuse categorical data (hair color) with ordinal data (days of the week).

Outline

- 1 Single proportion inference
- 2 Difference of two proportions
- 3 Chi-square goodness-of-fit test
- 4 Testing for independence in two-way tables

Sampling distribution of \hat{p}

- Recall that we estimate a population proportion p as the sample proportion

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

where x is the total number of successes and n is the sample size.

The sampling distribution for \hat{p} based on a sample of size n from a population with a true proportion p is nearly normal

$$\hat{p} \sim \mathcal{N} \left(p, \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{SE}^a} \right)$$

- 1 The sample's observations are independent, e.g. are from a simple random sample.
- 2 $np(1-p) \geq 5$

^aIf p is unknown (most cases), we use \hat{p} in the calculation of the standard error

Confidence interval for a proportion

- When \hat{p} can be modeled using a normal distribution, the **confidence interval** for p takes the form

$$\hat{p} \pm z_{1-\alpha/2} \times \text{SE} = \hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Example

We are given that $n = 670$, $\hat{p} = 0.85$. Which of the below is the correct calculation of the 95 % confidence interval?

- (a) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}}$
- (b) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$
- (c) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$
- (d) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

Choosing a sample size

Example

We are given that $n = 670$, $\hat{p} = 0.85$. How big a sample is required to ensure the margin of error is smaller than 0.01 using a 95 % confidence level?

Choosing a sample size

Example

We are given that $n = 670$, $\hat{p} = 0.85$. How big a sample is required to ensure the margin of error is smaller than 0.01 using a 95 % confidence level?

$$1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \leq 0.01$$

Choosing a sample size

Example

We are given that $n = 670$, $\hat{p} = 0.85$. How big a sample is required to ensure the margin of error is smaller than 0.01 using a 95 % confidence level?

$$1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \leq 0.01$$
$$1.96^2 \times \frac{0.85 \times 0.15}{n} \leq 0.01^2$$

Choosing a sample size

Example

We are given that $n = 670$, $\hat{p} = 0.85$. How big a sample is required to ensure the margin of error is smaller than 0.01 using a 95 % confidence level?

$$\begin{aligned} 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} &\leq 0.01 \\ 1.96^2 \times \frac{0.85 \times 0.15}{n} &\leq 0.01^2 \\ \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} &\leq n \end{aligned}$$

Choosing a sample size

Example

We are given that $n = 670$, $\hat{p} = 0.85$. How big a sample is required to ensure the margin of error is smaller than 0.01 using a 95 % confidence level?

$$\begin{aligned}1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} &\leq 0.01 \\1.96^2 \times \frac{0.85 \times 0.15}{n} &\leq 0.01^2 \\ \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} &\leq n \\ n &\geq 4898.04\end{aligned}$$

We need at least 4899 participants

Choosing a sample size

Example

A university newspaper is conducting a survey to determine what fraction of students support a \$200 per year increase in fees to pay for a new football stadium. How big a sample is required to ensure the margin of error is smaller than 0.04 using a 95 % confidence level?

Choosing a sample size

Example

A university newspaper is conducting a survey to determine what fraction of students support a \$200 per year increase in fees to pay for a new football stadium. How big a sample is required to ensure the margin of error is smaller than 0.04 using a 95 % confidence level?

- Use $\hat{p} = 0.5$ the most conservative estimate (worst case scenario), yielding the highest possible sample size

$$\begin{aligned} 1.96 \times \sqrt{\frac{0.5 \times (1 - 0.5)}{n}} &< 0.04 \\ 1.96^2 \times \frac{0.5 \times 0.5}{n} &< 0.04^2 \\ \frac{1.96^2 \times 0.5 \times 0.5}{0.04^2} &< n \\ 600.25 &< n \end{aligned}$$

We need 601 participants or more

Hypothesis testing for proportions

- To test the hypothesis $H_0 : p = p_0$ vs $H_1 : p \neq p_0$ with a significance level of α

Test ($p \neq p_0$)

Compute

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim \mathcal{N}(0, 1)$$

- if $|z| > z_{1-\alpha/2}$, then we reject H_0

Confidence interval

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{p_0(1 - p_0)/n}$$

- if 100%(1 - α) CI **does not contain** p_0 , then we reject H_0

Hypothesis testing for proportions

Example

Suppose that 8 % of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

- (a) A random sample of 125 college students where 12 % are vegetarians would be considered unusual.
- (b) A random sample of 250 college students where 12 % are vegetarians would be considered unusual.

Outline

- 1 Single proportion inference
- 2 Difference of two proportions
- 3 Chi-square goodness-of-fit test
- 4 Testing for independence in two-way tables

Difference of two proportions

Example

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 (anticoagulant) 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

Point estimation of the difference of two proportions

- We estimate the **difference** between two population proportion $p_1 - p_2$ using the sample proportions

$$\hat{p}_1 - \hat{p}_2$$

based on sample sizes n_1, n_2


Sampling distribution of the difference of two proportions

- The sampling distribution for $\hat{p}_1 - \hat{p}_2$ is nearly normal

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N} \left(p_1 - p_2, \underbrace{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}_{SE^1} \right)$$

when

- 1 **Independence:** within groups and between groups (satisfied if the data come from two independent random samples or if the data come from a randomized experiment)
- 2 **Success-failure:** At least 10 observed successes and 10 observed failures in the two groups

¹If p_1, p_2 are unknown (most cases), we use \hat{p}_1, \hat{p}_2 in the calculation of the standard error 

Confidence interval for $p_1 - p_2$

- When $\hat{p}_1 - \hat{p}_2$ can be modeled using a normal distribution, the **confidence interval** for $p_1 - p_2$ takes the form

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Example

Calculate a 90 % confidence interval of the difference for the survival rates in the CPR study.

Example solution

- We first calculate the sample proportion difference

$$\hat{p}_1 - \hat{p}_2 = \frac{14}{40} - \frac{11}{50} = 0.35 - 0.22 = 0.13$$

- Then we calculate the standard error

$$SE \approx \sqrt{\frac{0.35(1 - 0.35)}{40} + \frac{0.22(1 - 0.22)}{50}} = 0.095$$

- For a 90 % confidence interval we use $z_{1-\alpha/2} = 1.65$, therefore

$$CI_{90\%} = 0.13 \pm 1.65 \times 0.095 \rightarrow (-0.027, 0.287)$$

Hypothesis testing for the difference of two proportions

- We would like to test the hypothesis

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_1 : p_1 \neq p_2$$

with a significance level of α

- Or equivalently,

$$H_0 : p_1 - p_2 = 0 \quad \text{vs} \quad H_1 : p_1 - p_2 \neq 0$$

with a significance level of α

Hypothesis testing for the difference of two proportions

- So we define the statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}} \sim \mathcal{N}(0, 1)$$

where \hat{p}_{pooled} is the **expected number of successes and failures across the entire study**, which is calculated as

$$\hat{p}_{\text{pooled}} = \frac{\# \text{successes}_1 + \# \text{successes}_2}{n_1 + n_2} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Example

Calculate \hat{p}_{pooled} in the CPR study.

Example solution

	Survived	Died	Total	\hat{p}
Control	11	39	50	0.220
Treatment	14	26	40	0.350
Total	25	65	90	0.278

- In this case

$$\hat{p}_{\text{pooled}} = \frac{11 + 14}{50 + 40} = \frac{25}{90} = 0.278$$

Hypothesis testing for the difference of two proportions

- To test the hypothesis

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_1 : p_1 \neq p_2$$

with a significance level of α

Test

Compute the statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}} \sim \mathcal{N}(0, 1)$$

- if $|z| > z_{1-\alpha/2}$, then we reject H_0
- if $|z| \leq z_{1-\alpha/2}$, then we fail to reject H_0

Hypothesis testing for the difference of two proportions

Example

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 (anticoagulant) 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

Is the blood thinner useful for a 5 % significance level?

Example solution

- We know $\hat{p}_1 = 0.35$, $\hat{p}_2 = 0.22$, $\hat{p}_{\text{pooled}} = 0.278$, $n_1 = 40$, $n_2 = 50$, so

$$z = \frac{0.35 - 0.22}{\sqrt{\frac{0.278(1-0.278)}{40} + \frac{0.278(1-0.278)}{50}}} = 1.367$$

- For a $\alpha = 0.05$ significance level, we have $z_{1-\alpha/2} = 1.96$
- Therefore

We fail to reject the H_0 at a significance level of 5 %

Outline

- 1 Single proportion inference
- 2 Difference of two proportions
- 3 Chi-square goodness-of-fit test
- 4 Testing for independence in two-way tables

Outline

- 1 Single proportion inference
- 2 Difference of two proportions
- 3 Chi-square goodness-of-fit test
- 4 Testing for independence in two-way tables