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**[1] 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).

[1] Also known as *contingency* table

Outline

1. Single proportion inference
2. Difference of two proportions
3. Chi-square tests for contingency tables
   - Goodness-of-fit test
   - Independence
4. Fisher's exact test

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

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

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

[a] If 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?

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

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?

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

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

$\boxed{\text{We need } 601 \text{ participants or more}}$

Hypothesis testing for proportions

- To test the hypothesis $H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$ 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 $100\,\%(1-\alpha)$ 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

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

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(p_1 - p_2, \underbrace{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}_{\mathsf{SE}^2}\right)$$

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

[2] If 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

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

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

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

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 $\hat{p}_{\text{pooled}}$ 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 $\hat{p}_{\text{pooled}}$ 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 $|z| \leq z_{1-\alpha/2}$, the 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

$$\boxed{\text{We fail to reject the } H_0 \text{ at a significance level of } 5\,\%}$$

Outline

One-way vs Two-way tables

- A one-way table describes counts for each outcome in a single variable.
  - Test: **goodness-of-fit**. *"Does the data fit a particular distribution?"*, *"is there any inconsistency between the observed and the expected counts?"*

Outcome	Observed	Expected
1	53,222	52,612
2	52,118	52,612
3	52,465	52,612
4	52,338	52,612
5	52,244	52,612
6	53,285	52,612
Total	315,672	315,672

Zacariah Labby *experiment*, rolling 12 dice 26,306 times

One-way vs Two-way tables

- A two-way table describes counts for combinations of outcomes for two variables.
  - Test: **independence**
    - Row homogeneity: *"are proportions the same for every row at the different columns?"*

	Excellent	Very Good	Average	Poor	Terrible	Total
Restaurant A	146	70	33	24	25	298
Restaurant B	419	277	102	66	52	916

Ratings for two restaurants on Tripadvisor. A was ranked 187/1257, and B was ranked 116/1257

One-way vs Two-way tables

- A two-way table describes counts for combinations of outcomes for two variables.
  - Test: **independence**
    - Independence: *"are two measurements somehow related?"*

Smoking status	exercise: regular	exercise: some/none	Total
Never	87	102	189
Occasional	12	7	19
Regular	9	8	17
Heavy	7	4	11
Total	115	121	236

Smoking habits vs exercise level. Is smoking status independent of exercise level?

Goodness-of-fit test

H_0 : The observed counts follow the same distribution as the expected counts

H_1 : The observed counts **do not** follow the same distribution as the expected counts

- Quantify how different the observed counts are from the expected counts

Outcome	Observed	Expected
1	53,222	52,612
2	52,118	52,612
3	52,465	52,612
4	52,338	52,612
5	52,244	52,612
6	53,285	52,612
Total	315,672	315,672

Goodness-of-fit test: chi-square statistic

- Quantify how different the observed counts are from the expected counts we will use the **chi-square statistic**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

  where

  - O_i is the observed count in cell i
  - E_i is the expected count in cell i
  - k is the total number of cells

- Notice that this statistic can be written as $\chi^2 = \sum_{i=1}^{k} Z_i^2$, where

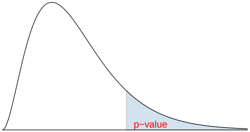
$$Z_i = \frac{O_i - E_i}{\sqrt{E_i}} = \frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

Calculating the chi-square statistic

Outcome	Observed	Expected	$\frac{(O_i - E_i)^2}{E_i}$
1	53,222	52,612	$\frac{(53{,}222 - 52{,}612)^2}{52{,}612} = 7.07$
2	52,118	52,612	$\frac{(52{,}118 - 52{,}612)^2}{52{,}612} = 4.64$
3	52,465	52,612	$\frac{(52{,}465 - 52{,}612)^2}{52{,}612} = 0.41$
4	52,338	52,612	$\frac{(52{,}338 - 52{,}612)^2}{52{,}612} = 1.43$
5	52,244	52,612	$\frac{(52{,}244 - 52{,}612)^2}{52{,}612} = 2.57$
6	53,285	52,612	$\frac{(53{,}285 - 52{,}612)^2}{52{,}612} = 8.61$
Total	315,672	315,672	24.73

Finding a p-value for a chi-square test

- We have calculated a test statistic of $\chi^2 = 24.67$
- We have df $= k - 1 = 6 - 1 = 5$ degrees of freedom
- The we compute the $p-$value as



$$p = P(\chi^2_5 > 24.67)$$
$$= \texttt{1-chi2(df=5).cdf(24.67)}$$
$$= 0.00016$$

- Therefore, at a significance level of 5 % ....

We reject H_0, the data provide convincing evidence that the dice are biased!

- It turns out that the 1-6 axis is consistently shorter than the other two (2-5 and 3-4), thereby supporting the hypothesis that the faces with one and six pips are larger than the other faces.

Conditions for the chi-square goodness-of-fit test

1. **Independence**: each case that contributes a count to the table must be independent of all the other cases in the table.

2. **Sample size:** each particular scenario (i.e. cell) must have at least 5 **expected** cases.

3. **df $>$ 1:** degrees of freedom must be greater than 1.

Failing to check conditions may unintentionally affect the test's error rates.

Independence test

	Excellent	Very Good	Average	Poor	Terrible	Total
Restaurant A	146	70	33	24	25	298
Restaurant B	419	277	102	66	52	916
Total	565	347	135	90	77	1214

Ratings for two restaurants on Tripadvisor. A was ranked 187/1257, and B was ranked 116/1257

H_0 : The probability of a particular rating is the same for either restaurant (ratings are independent of the restaurant)

H_1 : The probability of a particular rating **is not** the same for either restaurant (ratings depend on the restaurant)

- We might wonder whether the customer ratings are significantly different; if they are not, one could argue that the rankings are not meaningful.

Chi-square test of independence

- The test statistic is calculated as

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \sim \chi^2_{(R-1)\times(C-1)}$$

  where

  - $O_{i,j}$ is the observed count in row i, column j
  - $E_{i,j}$ is the expected count in row i, column j
  - R is the number of rows
  - C is the number of columns
  - df $= (R-1) \times (C-1)$

- The expected counts can be computed as

$$E_{i,j} = \frac{(\text{total in row } i) \times (\text{total in column } j)}{\text{table total}}$$

Expected counts

	Excellent	Very Good	Average	Poor	Terrible	Total
Restaurant A	146	70	33	24	25	298
Restaurant B	419	277	102	66	52	916
Total	565	347	135	90	77	1214

- The expected counts can be computed as

$$E_{i,j} = \frac{(\text{total in row } i) \times (\text{total in column } j)}{\text{table total}}$$

	Excellent	Very Good	Average	Poor	Terrible	Total
Restaurant A	$\frac{298\times565}{1214}$	$\frac{298\times347}{1214}$	$\frac{298\times135}{1214}$	$\frac{298\times90}{1214}$	$\frac{298\times77}{1214}$	298
Restaurant B	$\frac{916\times565}{1214}$	$\frac{916\times347}{1214}$	$\frac{916\times135}{1214}$	$\frac{916\times90}{1214}$	$\frac{916\times77}{1214}$	916
Total	565	347	135	90	77	1214

Calculating the test statistic

- Expected counts are shown in blue next to the observed counts

	Excellent	Very Good	Average	Poor	Terrible	Total
Restaurant A	$146 \mid 138.7$	$70 \mid 85.2$	$33 \mid 33.1$	$24 \mid 22.1$	$25 \mid 18.9$	298
Restaurant B	$419 \mid 426.3$	$277 \mid 261.8$	$102 \mid 101.9$	$66 \mid 67.9$	$52 \mid 58.1$	916
Total	565	347	135	90	77	1214

- We can now compute our observed test statistic:

$$\chi^2 = \frac{(146-138.7)^2}{138.7} + \frac{(70-85.2)^2}{85.2} + \ldots + \frac{(52-58.1)^2}{58.1} = 6.92$$

and the degrees of freedom

$$\text{df} = (R-1)\times(C-1) = (2-1)\times(5-1) = 4$$

Calculating the $p-$value

- Having $\chi^2 = 6.92$ and df $= 4$, then

$$p = P(\chi^2_4 > 6.92) = \texttt{1-chi2(df=4).cdf(6.92)} = 0.14$$

- Therefore, at a significance level of 5 %

$$\boxed{\text{We fail to reject } H_0}$$

- Thus, there is no evidence to say that a customer is more likely to rate one restaurant higher than the other. This would suggest that the difference in rankings between the two restaurants (116 and 187) is not particularly meaningful

Outline

Fisher's exact test

- **Nonparametric** test for testing independence
- Typically used only for 2×2 contingency table
- An alternative to Pearson's chi-squared test for independence
  - ☑ Fisher's exact test yields **exact p-values**
  - ☑ Typically used for tables with **small expected values**, where chi-squared test assumptions are not valid
  - ✗ Fisher's exact test may **not work for large sample sizes**, due to computational reasons

Fisher's exact test

	N	Y	total
A	18	2	20
B	11	9	20
total	29	11	40

- Assume the null hypothesis (independence) is true
- Constrain the marginal counts to be as observed
- What's the chance of getting this exact table?

- Fisher showed that the probability of obtaining this exact table was given by the *hypergeometric* distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\;(c+d)!\;(a+c)!\;(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

Fisher's exact test

	N	Y	total
A	18	2	20
B	11	9	20
total	29	11	40

- In Python, probability of obtaining this exact table can be calculated as

Option 1

```
from scipy.special import comb

a = 18; b = 2; c = 11; d = 9
n = a+b+c+d
p = comb(a+b,a) * comb(c+d,c) / comb(n,a+c)

>> 0.013804
```

Option 2

```
from scipy.stats import hypergeom

rv = hypergeom(M=a+b+c+d, n=a+c, N=a+b)
p = rv.pmf(18)

>> 0.013804
```

Fisher's exact test: $p-$value calculation

1. For all possible tables (with the observed marginal counts), calculate the relevant hypergeometric probability

20	0	$\rightarrow 0.00007$
9	11	
19	1	$\rightarrow 0.00160$
10	10	
18	**2**	$\rightarrow \mathbf{0.01380}$
11	**9**	
17	3	$\rightarrow 0.06212$
12	8	

16	4	$\rightarrow 0.16246$
13	7	
15	5	$\rightarrow 0.25994$
14	6	
14	6	$\rightarrow 0.25994$
15	5	
13	7	$\rightarrow 0.16246$
16	4	

12	8	$\rightarrow 0.06212$
17	3	
11	9	$\rightarrow 0.01380$
18	2	
10	10	$\rightarrow 0.00160$
19	1	
9	11	$\rightarrow 0.00007$
20	0	

Fisher's exact test: $p-$value calculation

2. Calculate the $p-$value as the sum of the probabilities for all tables having a probability equal to or smaller than that observed.

20 0 / 9 11	$\rightarrow$ 0.00007	
19 1 / 10 10	$\rightarrow$ 0.00160	
18 2 / 11 9	$\rightarrow$ **0.01380**	
17 3 / 12 8	$\rightarrow$ 0.06212	

16 4 / 13 7	$\rightarrow$ 0.16246	
15 5 / 14 6	$\rightarrow$ 0.25994	
14 6 / 15 5	$\rightarrow$ 0.25994	
13 7 / 16 4	$\rightarrow$ 0.16246	

12 8 / 17 3	$\rightarrow$ 0.06212	
11 9 / 18 2	$\rightarrow$ 0.01380	
10 10 / 19 1	$\rightarrow$ 0.00160	
9 11 / 20 0	$\rightarrow$ 0.00007	

$p =$ `2*(rv.pmf(20) + rv.pmf(19) +rv.pmf(18))` $= 0.03095$

Fisher's exact test: $p-$value calculation

	N	Y	total
A	18	2	20
B	11	9	20
total	29	11	40

- In Python, the $p-$value fo the Fisher's exact test can be directly calculated as

`_, p_value = fisher_exact( [ [18, 2] , [11,9] ] )` $= 0.03095$

Chi-square test vs Fisher's exact test

For the above table, calculate the chi-square test and compare its $p-$value with respect to the Fisher's exact test.

SOL: `chi2_contingency( [ [18, 2] , [11,9] ] )` $= 0.03361$