

Categorical Data Analysis

1. Inferences about a Population Proportion (Large Sample)

For p ($np \geq 5, n(1 - p) \geq 5$)

Estimation

Parameter	Point Estimate	Confidence Interval
p	$\hat{p} = \frac{x}{n}$	$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}$

Testing

	Case (a)	Case (b)	Case (c)
Step 1	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p \leq p_0$ $H_1 : p > p_0$	$H_0 : p \geq p_0$ $H_1 : p < p_0$
Step 2	$\alpha = ?$	$\alpha = ?$	$\alpha = ?$
Step 3	$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$ <u>Rejection region</u> $ z \geq z_{\alpha/2}$	$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$ <u>Rejection region</u> $z \geq z_{\alpha}$	$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$ <u>Rejection region</u> $z \leq -z_{\alpha}$
Step 4	$z = ?$ Decision	Substitute \hat{p} and n $z = ?$ Decision	$z = ?$ Decision
Step 5	$p = 2 \times \text{area}$	$p = \text{area}$	$p = \text{area}$

Sample size determination

- i. Sample size needed to attain maximum error of estimate E :

$$n = p(1 - p) \left(\frac{z_{\alpha/2}}{E} \right)^2$$

- ii. If the value of p is unknown,

$$n = \frac{1}{4} \left(\frac{z_{\alpha/2}}{E} \right)^2$$

2. Difference Between Population Proportions (Large Samples)

Estimation

Parameter	Point Estimate	Confidence Interval
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2 = \frac{x}{n_1} - \frac{y}{n_2}$	$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Testing

Under H_0 (assuming $p_1 = p_2 = \pi$), $\hat{p} = \frac{x+y}{n_1+n_2} = \frac{n_1}{n_1+n_2}\hat{p}_1 + \frac{n_2}{n_1+n_2}\hat{p}_2$.

	Case (a)	Case (b)	Case (c)
Step 1	$H_0 : p_1 - p_2 = 0$ $H_1 : p_1 - p_2 \neq 0$	$H_0 : p_1 - p_2 = 0$ $H_1 : p_1 - p_2 > 0$	$H_0 : p_1 - p_2 = 0$ $H_1 : p_1 - p_2 < 0$
Step 2	$\alpha = ?$	$\alpha = ?$	$\alpha = ?$
Step 3	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ <u>Rejection region</u> $ z \geq z_{\alpha/2}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ <u>Rejection region</u> $z \geq z_{\alpha}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ <u>Rejection region</u> $z \leq -z_{\alpha}$
Step 4	Substitute $\hat{p}_1, \hat{p}_2, \hat{p}, n_1$ and n_2		
	$z = ?$ Decision	$z = ?$ Decision	$z = ?$ Decision
Step 5	$p = 2 \times \text{area}$	$p = \text{area}$	$p = \text{area}$

3. Chi-square Goodness-of-fit Test

Type of the problem: The data belong to k categories and the frequency of each category has been observed. The goal is to find out whether a simple model (specifying the probability of each category) fits the data.

Step 1: $H_0: p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$ vs. $H_1: \text{not } H_0$.

Step 2: $\alpha = ?$

Step 3:

$$\chi^2 = \sum_{\text{cells}} \frac{(O - E)^2}{E}$$

where O : observed frequency and E : expected frequency

Rejection region: $\chi^2 \geq \chi_{\alpha, k-1}^2$

Step 4: Decision

4. Contingency Table

Type of the problem: The data can be tabulated according to 2 different variables. The goal is to find out whether there is any relationship between the two variables (or if one variable is dependent on the other).

Step 1: H_0 : Two populations are independent, H_1 : Two populations are not independent.

Step 2: $\alpha = ?$

For a 2×2 table: Data are given as

	a	b	Row total
	c	d	n_1
			n_2
Column total	m_1	m_2	N

Step 3:

$$\chi^2 = \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2}$$

Rejection region: $\chi^2 \geq \chi_{\alpha, 1}^2$

Step 4: Decision

For a table bigger than 2×2 : r rows and c columns

	Column 1		Column 2		...	Column c		Total
	O	(E)	O	(E)		O	(E)	
Row 1	O_{11}	$(m_1 n_1 / N)$	O_{12}	$(m_2 n_1 / N)$		O_{1c}	$(m_c n_1 / N)$	n_1
Row 2	O_{21}	$(m_1 n_2 / N)$	O_{22}	$(m_2 n_2 / N)$		O_{2c}	$(m_c n_2 / N)$	n_2
...								
Row r	O_{r1}	$(m_1 n_r / N)$	O_{r2}	$(m_2 n_r / N)$		O_{rc}	$(m_c n_r / N)$	n_r
Total		m_1		m_2	...		m_r	N

Step 3:

$$\chi^2 = \sum_{\text{cells}} \frac{(O - E)^2}{E}$$

where O : observed frequency and E : expected frequency

Rejection region: $\chi^2 \geq \chi_{\alpha, (r-1)(c-1)}^2$

Step 4: Decision