CHI-SQUARED TEST

1

Team Members:

RE6121011 Jen-Lung Hsu H24094043 Cuei-Ting Pan

R26121065 Shi-Cian Wu N86124185 Kuei-Tien Lin

CONTENTS

- Assumptions
- Three primary applications and purpose
    - Goodness-of-fit test
    - Test for homogeneity
    - Test for independence
- Practical Applications
- R

ASSUMPTIONS

- **Categorical Data:** The variables being compared are categorical in nature. Each observation should fall into one and only one category for each variable.
- **Random Sampling:** The data used to construct the contingency table should come from a random sample or an appropriately designed experiment.
- **Independence of Observations:** The observations in the contingency table are assumed to be independent. This means that the presence or absence of an event in one category does not affect the presence or absence of an event in another category.
- **Expected Frequencies:** The expected frequency for each cell in the contingency table should be greater than or equal to 5. This assumption ensures that the chi-squared distribution approximation is valid.
- **Large Sample:** The chi-squared test relies on asymptotic theory, meaning it is most accurate and reliable when sample sizes are large.

THREE PRIMARY APPLICATIONS

- **Goodness-of-Fit Test**
- **Test for homogeneity**
- **Test for independence**

GOODNESS-OF-FIT TEST

- Purpose:

  determine if the observed data fits a specified theoretical model or expected pattern.

- Data:

  One population, a categorical variable with r levels

	類別	1	2	$\cdots$	r	總和
樣本觀察次數 $\longrightarrow$	O_i	O_1	O_2	$\cdots$	O_r	n
H_0 爲眞下之理論機率 $\longrightarrow$	p_i	p_1^*	p_2^*	$\cdots$	p_r^*	1
H_0 爲眞下之期望次數 $\longrightarrow$	E_i	$E_1 = np_1^*$	$E_2 = np_2^*$	$\cdots$	$E_r = np_r^*$	n

GOODNESS-OF-FIT TEST

GOODNESS-OF-FIT TEST

- H0: the observed data follows a specific distribution or pattern.
- H1: the observed data differs significantly from the expected distribution.

- Test statistic: $\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{H0} \chi^2(r-1)$

- Rejection region: $RR = \{\, \chi^2 \geq \chi^2_\alpha(r-1) \,\}$

TEST FOR HOMOGENEITY

- Purpose:

  determine whether two or more populations or groups have the same distribution

- Data:

  k populations, a categorical variable with r levels

		k 個母體				
		1	2	……	k	O_{ij} / E_{ij}
r 個類別	1	O_{11} / E_{11}	O_{12} / E_{12}	……	O_{1k} / E_{1k}	R_1
	2	O_{21} / E_{21}	O_{22} / E_{22}	……	O_{2k} / E_{2k}	R_2
	⋮	⋮	⋮	⋱	⋮	⋮
	r	O_{r1} / E_{r1}	O_{r2} / E_{r2}	……	O_{rk} / E_{rk}	R_r
		$C_1 = n_1$	$C_2 = n_2$	……	$C_k = n_k$	n

TEST FOR HOMOGENEITY

TEST FOR HOMOGENEITY

- H0: the populations have the same distribution for the categorical variable.
- H1:there are significant differences.

- Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow{H0} \chi^2((r-1)(k-1))$

- Rejection region: $RR = \{ \chi^2 \geq \chi^2_\alpha((r-1)(k-1)) \}$

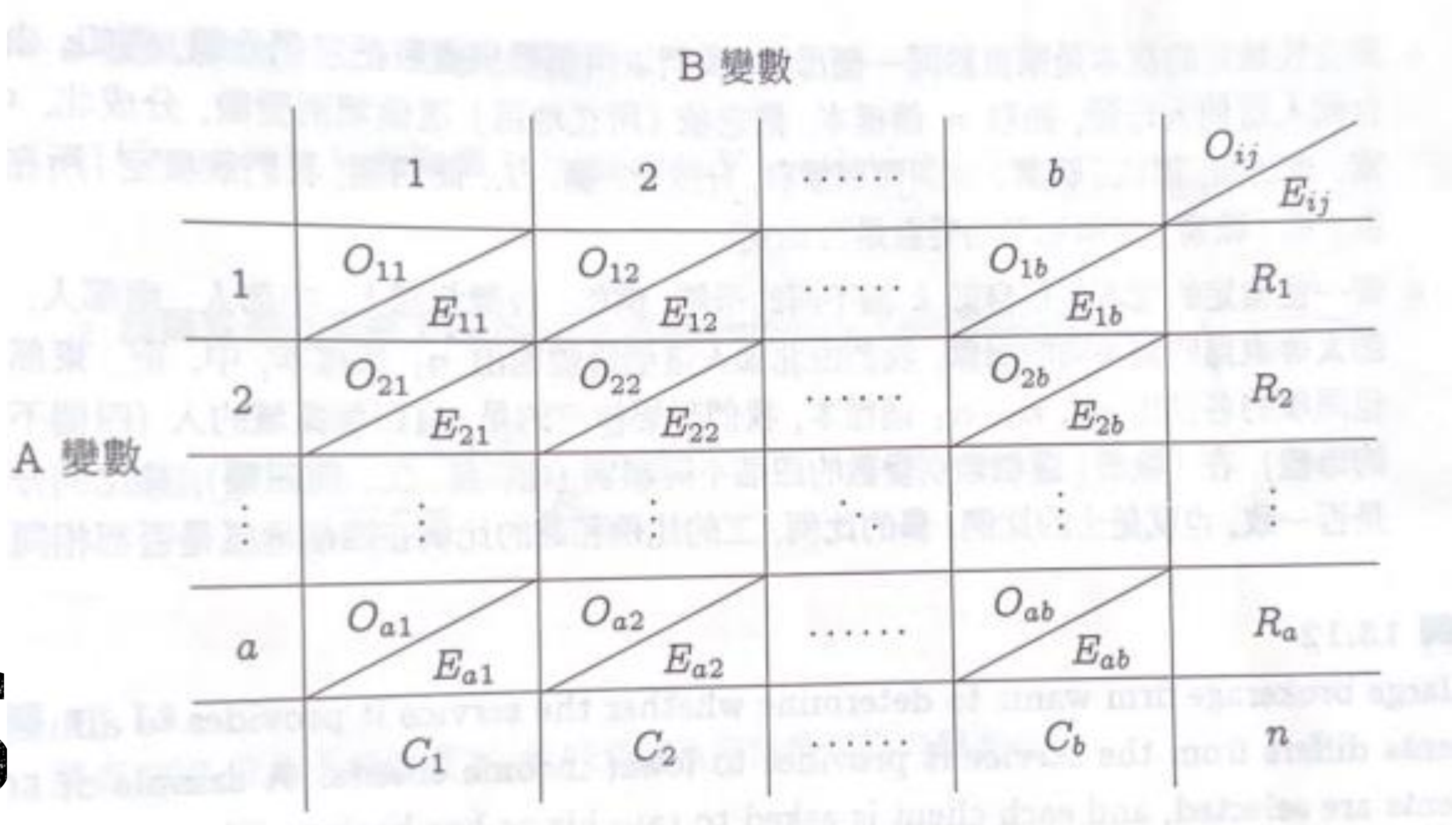
TEST FOR INDEPENDENCE

- Purpose:

  assess whether there is a significant association or relationship between two categorical variables.

- Data:

  one populations, two categorical variables with a levels, b levels



B 變數

A 變數	1	2	……	b	O_{ij} / E_{ij}
1	O_{11} / E_{11}	O_{12} / E_{12}	……	O_{1b} / E_{1b}	R_1
2	O_{21} / E_{21}	O_{22} / E_{22}	……	O_{2b} / E_{2b}	R_2
⋮	⋮	⋮	⋱	⋮	⋮
a	O_{a1} / E_{a1}	O_{a2} / E_{a2}	……	O_{ab} / E_{ab}	R_a
	C_1	C_2	……	C_b	n

TES FOR INDEPENDENC E

TEST FOR INDEPENDENCE

- H0: The two categorical variables are independent.
- H1: The two categorical variables are dependent.

- Test statistic: $\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow{H0} \chi^2((a-1)(b-1))$

- Rejection region: $RR = \{\chi^2 \geq \chi^2_\alpha((a-1)(b-1))\}$

PRACTICAL APPLICATIONS

GOODNESS-OF-FIT TEST

- Draw a die for 150 times and the outcome is:

Number	1	2	3	4	5	6
Times(O_i)	30	28	42	20	15	15

- H0:the die is fair(equal probability for each number)

  H1:the die is not fair

- The expected value for each number is $150 \times \frac{1}{6} = 25$

Number	1	2	3	4	5	6
Times(E_i)	25	25	25	25	25	25

GOODNESS-OF-FIT TEST

- Test Statistics

$$Q = \sum_{i=1}^{6} \frac{(Oi-Ei)^2}{Ei} = 21.92 \xrightarrow{H0} \chi^2(5)$$

- Since the p-value =0.0005423<0.05 , reject H0

  → The die is not fair.

- Code

```
> Oi<- c(30,28,42,20,15,15)
> chisq.test(Oi,p=rep(1/6,6))

        Chi-squared test for given probabilities

data:  Oi
X-squared = 21.92, df = 5, p-value = 0.0005423
```

TEST FOR HOMOGENEITY

- Conducting a study on alcohol poisoning among workers from various industries. (850 respondents)

\> data

	Alcoholism	no Alcoholism
Worker	67	233
civilservant	51	199
educators	32	268

Calculate the proportions

\> prop.table(data, margin = 1)

	Alcoholism	no Alcoholism
Worker	0.2233333	0.7766667
civilservant	0.2040000	0.7960000
educators	0.1066667	0.8933333

- H0: The proportions of alcohol poisoning among workers in the three industries are the same.
- H1: The proportions of alcohol poisoning among workers in the three industries are not the same.

TEST FOR HOMOGENEITY

- Test Statistics

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim X^2((r-1)(c-1))$$

其中：期望次數 $E_{ij} = \frac{(\text{第 i 列合計})\times(\text{第 j 行合計})}{\text{總樣本大小}}$。

- Rejection region

C=$\{X^2 > X^2_{0.05}(2) = 5.99\}$

> chisq.test(data)

```
        Pearson's Chi-squared test

data:  compare
X-squared = 15.896, df = 2, p-value = 0.0003534
```

- Since χ^2 = 15.896 > 5.99 and P-value = 0.0003534 < 0.05 , reject H0.
- This suggests that the proportions of alcohol poisoning are not equal across the three industries.

TEST FOR INDEPENDENCE

- We have a list of movie genres; this is our first variable. Our second variable is whether or not the patrons of those genres bought snacks at the theater. (600)
  - actual data

Type of Movie	Action	Comedy	Family	Horror
Snacks	50	125	90	45
No Snacks	75	175	30	10

  - expected data

Type of Movie	Action	Comedy	Family	Horror
Snacks	65	155	62	28
No Snacks	60	145	58	27

- H0: Movie Type and Snack purchases are independent

  H1: Movie Type and Snack purchases are not independent

TEST FOR INDEPENDENCE

- Test Statistics

$$\chi^2 = \Sigma\Sigma \frac{(Oij-Eij)^2}{Eij^2} = 65.03 > 7.815$$

- df=(r−1)×(c−1)，df=(4−1)×(2−1)=3
- P-value < 0.0001 <0.05
- Rejection region

Since χ^2 = 65.03 > 7.815 and P-value <0.0001< 0.05 , reject H0.

- The results that we collected from our movie goers would be extremely unlikely if there were truly no relationship between types of movies and snack purchases.

THANK YOU FOR LISTENING.