UMP OPEN COURSEWARE

Universiti Malaysia PAHANG
Engineering • Technology • Creativity

CHAPTER 6

GOODNESS OF FIT AND CONTINGENCY TABLE

Expected Outcomes

- ✓ Able to test the goodness of fit for categorical data.
- ✓ Able to test whether the categorical data fit to the certain distribution such as Binomial, Normal and Poisson.
- ✓ Able to use a contingency table to test for independence and homogeneity proportions.

PREPARED BY: DR SITI ZANARIAH SATARI & FARAHANIM MISNI

Contents

6.1 Goodness of Fit Test

6.1.1 Goodness of Fit Test for Categorical Data

6.1.2 Fitting of the Distribution

6.2 Contingency Table

6.2.1 Testing for Two Variables between Independence

6.2.2 Test of Homogeneity Proportions

6.1 GOODNESS OF FIT TEST

When to use Chi-Square Distribution?

1. Find confidence Interval for a variance or standard deviation
2. Test a hypothesis about a single variance or standard deviation
3. Tests concerning frequency distributions for categorical data **(Goodness of Fit)**
4. Tests concerning probability distributions **(Goodness of Fit)**
5. Test the Independence of two variables **(Contingency Table)**
6. Test the homogeneity of proportions **(Contingency Table)**

When to use Goodness of fit test?

1. To compare between observed and expected frequencies for categorical data.

Example: To meet customer demands, a manufacturer of running shoes may wish to see whether buyers show a preference for a specific style. If there were no preference, one would expect each style to be selected with equal frequency.

2. When you have some practical data and you want to know how well a particular statistical distribution (such as poisson, binomial or normal models) fit the data.

Example: A researcher wish to test whether the number of children in a family follows a Poisson distribution.

6.1.1 GOODNESS OF FIT TEST FOR CATEGORICAL DATA

Hypothesis Null and Alternative

H_0 : There is **no difference …** or **no change …** or **no preference …**

H_1 : There is a **difference …** or **change…**or **preference …**

Or

H_0 : State the claim of the categorical distribution

H_1 : The categorical distribution is not the same as stated in H0.

Example:

H_0: Buyers show no preference for a specific style.

H_1: Buyers show a preference for a specific style.

Assumptions/Conditions

1. The data are obtained from a random sample.
2. The variable under study is categorical data.
3. The expected frequency for each category must be **at least 5**. If the expected frequency is less than 5, combine the adjacent category.

The Test Statistics

$$\chi^2_{test} = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i} \sim \chi^2_{\alpha,v}$$

Where

O_i = observed frequency for the *i* category

E_i = expected frequency for the *i* category

k = the number of categories

degrees of freedom, $v = k - 1$

and

$$E_i = nP_i \text{ where } P_i \text{ is a probability for } i = 1, 2, \ldots, k$$

Procedures

1. State the hypothesis and identify the claim.
2. Compute the test statistics value. $\chi^2_{test} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
3. Find the critical value. The test is always right-tailed since O – E are square and always positive.
4. Make the decision – Reject Ho if $\chi^2_{test} > \chi^2_{\alpha,k-1}$.
5. Draw a conclusion to reject or accept the claim.

Why this test is called goodness of fit?

- If the graph between observed values and expected values is fitted, one can see whether the values are close together or far apart.

- When observed values and expected values are close together:
  - ✓ the chi-square test value will be small.
  - ✓ Decision must be not reject H_0 (accept H_0).
  - ✓ Hence there is a “good fit”.

- When observed values and expected values are far apart:
  - ✓ the chi-square test value will be large.
  - ✓ Decision must be reject H_0 (accept H_1).
  - ✓ Hence there is a “not a good fit”.

Example 1: GoF for Categorical Data

A market analyst whished to see whether consumers have any preference among five flavors of a new fruit soda. A sample of 100 people provided these data.

Cherry	Strawberry	Orange	Lime	Grape
32	28	16	14	10

Is there enough evidence to reject the claim that there is no preference in the selection of fruit soda flavors at 0.05 significance level?

Example 1: solution

H_0: There is no preference in the selection of fruit soda flavours (claim)
H_1: There is preference in the selection of fruit soda flavours

$$E_i = nP_i$$
$$= 100\left(\frac{1}{5}\right)$$
$$= 20$$

Frequency	Cherry	Strawberry	Orange	Lime	Grape
Observed (O_i)	32	28	16	14	10
Expected (E_i)	20	20	20	20	20

Example 1: solution

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

$$= \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(10-20)^2}{20}$$

$$= 18.0$$

$$\chi^2_{critical} = \chi^2_{\alpha,k-1}$$

$$= \chi^2_{0.05,4}$$

$$= 9.4877$$

Since $\left(\chi^2_{test} = 18.0\right) > \left(\chi^2_{0.05,4} = 9.4877\right)$, then we reject H_0.

At $\alpha = 0.05$, there is enough evidence to reject the claim that there is no preference in the selection of fruit soda flavours.

6.1.2 FITTING OF DISTRIBUTION

Hypothesis Null and Alternative

H_0: The population of a set of observed data comes from a specific distribution (Poisson/Binomial/Normal).

H_1: The population of a set of observed data does not comes from a specific distribution (Poisson/Binomial/Normal).

Example:

H_0: The number of children in a family follows a Poisson distribution

H_1: The number of children in a family does not follows a Poisson distribution

NOTES

1. The expected frequency for each category must be **at least 5**. If the expected frequency is less than 5, combine the adjacent category.

2. Reject H0 if $\chi^2_{test} > \chi^2_{\alpha,k-p-1}$ where *p* is the number of parameters in the hypothesized distribution estimated by sample statistics.

Procedures

1. State the hypothesis and identify the claim.

2. Compute the test value $\chi^2_{test} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. If the expected frequency is less than 5, it should be combined with the expected frequency in the adjacent class interval.

3. Find the critical value. The test is always right-tailed since *O* – *E* are square and always positive.

4. Make the decision – reject Ho if $\chi^2_{test} > \chi^2_{\alpha,k-p-1}$ where *p* is the number of parameters in the hypothesized distribution estimated by sample statistics.

5. Draw a conclusion to reject or accept the claim.

Example 2: GoF for Fitting Distribution

The number of defects in the printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of 60 printed boards has been collected and the following numbers of defects observed.

Number of defect	Observed frequency
0	32
1	15
2	9
3	4

Test the hypothesis that number of defects in the printed circuit boards is follows a Poisson distribution at $\alpha = 0.05$.

Example 2: solution

H_0 : The number of defects in printed circuit boards follows a Poisson distribution.
H_1 : The number of defects in printed circuit boards does not follow a Poisson distribution.

For Poisson distribution, find the average value, λ

$$\lambda = \frac{0(32)+1(15)+2(9)+3(4)}{60} = 0.75$$

We estimated the value of λ , thus *parameter, p* = 1.

No. of defects	i	O_i	$P_i = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x\,!}$	$E_i = nP_i$
0	1	32	$P_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0\,!} = 0.4724$	$E_1 = 60(0.4724) = 28.344$
1	2	15	$P_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.3543$	$E_2 = 60(0.3543) = 21.258$
2	3	9	$P_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.1329$	$E_3 = 60(0.1329) = 7.974$
3 (or more)	4	4	$P_4 = P(X \geq 3) = 1 - [P_1 + P_2 + P_3]$ $= 1 - [0.4724 + 0.3543 + 0.1329] = 0.0404$	$E_4 = 60(0.0404) = 2.424$

Example 2: solution

No. of defects	Observed frequencies (O_i)	Expected frequencies (E_i)
0	32	28.344
1	15	21.258
2	9	7.974
3 (or more)	4	2.424

$E_i < 5$. Combine the adjacent category and reconstruct the table

No. of defects	Observed frequencies (O_i)	Expected frequencies (E_i)
0	32	28.344
1	15	21.258
2 (or more)	13	10.398

Example 2: solution

No. of defects	Observed frequencies (O_i)	Expected frequencies (E_i)
0	32	28.344
1	15	21.258
2 (or more)	13	10.398

$$\chi^2_{test} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(32-28.344)^2}{28.344} + \frac{(15-21.258)^2}{21.258} + \frac{(13-10.398)^2}{10.398}$$
$$= 2.965$$

$$\chi^2_{critical} = \chi^2_{\alpha,k-p-1} = \chi^2_{0.05,3-1-1} = \chi^2_{0.05,1} = 3.8415$$

Since $\left(\chi^2_{test} = 2.965\right) < \left(\chi^2_{0.05,1} = 3.8415\right)$, then we do not reject H_0.

At $\alpha = 0.05$, there is sufficient evidence to conclude that the number of defects in printed circuit boards follows a Poisson distribution.

Example 3

A farmer kept a record of the number of heifer calves born to each of his cows during the first five years. The results are summarized below.

No of heifers	0	1	2	3	4	5
No of cows	4	19	41	52	26	8

Test at the 5% level of significance, whether these data adequate for binomial distribution or not with parameter $n = 5$ and $p = 0.5$.

The parameters $n = 5$ and $p = 0.5$ are given thus *parameter,* $p = 0$.

Example 3: solution

H_0 = The numbers of heifer calves born to each of his cows are adequate for binomial distribution.

H_1 = The numbers of heifer calves born to each of his cows are not adequate for binomial distribution.

Probability, $P_i = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$	**Expected frequencies,** $E_i = nP_i$
$P_1 = P(X = 0) = \binom{5}{0} 0.5^0 (0.5)^5 = 0.0313$	$E_1 = 150(0.0313) = 4.695$
$P_2 = P(X = 1) = \binom{5}{1} 0.5^1 (0.5)^4 = 0.1563$	$E_2 = 150(0.1563) = 23.445$
$P_3 = P(X = 2) = \binom{5}{2} 0.5^2 (0.5)^3 = 0.3125$	$E_3 = 150(0.3125) = 46.875$
$P_4 = P(X = 3) =$	$E_4 =$
$P_5 = P(X = 4) =$	$E_5 =$
$P_6 = P(X = 5) =$	$E_6 =$

Example 3: solution

Observed frequencies (O_i)		Expected frequencies (E_i)	
4		4.695	
19		23.445	
41	41	46.875	46.875
52	52	46.875	46.875
26		23.445	
8		4.695	

$\chi^2_{test} =$

$\chi^2_{0.05,k-p-1} =$

Decision:

Example 4

The sugar concentrations in apple juice measured at 20° C were reported in article of Food Testing & Analysis for 50 readings in the frequency distribution table below.

Class interval (sugar concentration)	1.0-1.2	1.3-1.5	1.6-1.8	1.9-2.1
Observed frequency	10	15	15	10

At the 2.5% level of significance, is there any evidence to support the assumption that the sugar concentration is normally distributed when $\mu = 1.5$ and $\sigma = 0.5$?

The parameters $\mu = 1.5$ and $\sigma = 0.5$ are given thus *parameter*, $p = 0$.

Example 4: solution

H_0 : The sugar concentration in clear apple juice is normally distributed.

H_1 : The sugar concentration in clear apple juice is not normally distributed.

$$\begin{aligned} P(0.95 < X < 1.25) &= P\left(\frac{0.95-1.5}{0.5} < Z < \frac{1.25-1.5}{0.5}\right) \\ &= P(-1.1 < Z < -0.5) \\ &= 0.1728 \end{aligned}$$

$$\begin{aligned} P(1.25 < X < 1.55) &= P\left(\frac{1.25-1.5}{0.5} < Z < \frac{1.55-1.5}{0.5}\right) \\ &= P(-0.5 < Z < 0.1) \\ &= \end{aligned}$$

$$\begin{aligned} P(1.55 < X < 1.85) &= P\left(\frac{1.55-1.5}{0.5} < Z < \frac{1.85-1.5}{0.5}\right) \\ &= P(0.1 < Z < 0.7) \\ &= \end{aligned}$$

$$\begin{aligned} P(1.85 < X < 2.15) &= P\left(\frac{1.85-1.5}{0.5} < Z < \frac{2.15-1.5}{0.5}\right) \\ &= P(0.7 < Z < 1.3) \\ &= \end{aligned}$$

Example 4: solution

Class interval	Observed frequency	Class boundaries	Expected frequency
1.0 – 1.2	10	0.95 – 1.25	$50(0.1728) = 8.64$
1.3 – 1.5	15	1.25 – 1.55	$50(0.2313) = 11.565$
1.6 – 1.8	15	1.55 – 1.85	$50(0.2182) = 10.91$
1.9 – 2.1	10	1.85 – 2.15	$50(0.1452) = 7.26$

Since $(\chi^2_{test} = 3.8017) < (\chi^2_{0.025,3} = 9.3484)$, then we do not reject H_0

At $\alpha = 0.025$, there is enough evidence to conclude that the sugar concentration in apple juice is normally distributed.

6.2 CONTINGENCY TABLE

- ✓ The contingency table is called an *r* x *c* contingency table (*r* categories for the row variable and *c* categories for the column variable).
- ✓ We are interested to find out whether the row variable is independent of the column variable.

		Column variable , j	
Row variable i	O_{11}	O_{12}	$n_{1.}$
	O_{21}	O_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

The Test Statistics

$$\chi^2_{test} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_v$$

where

O_{ij} = the observed frequency in cell (i , j)

E_{ij} = the expected frequency in cell (i , j)

i = level on the first classification method (row variable)

j = level on the second classification method (column variable)

degree of freedom, $v = (r-1)\times(c-1)$

The Expected Frequency

		Column variable, *j*	
Row variable, *i*	O_{11}	O_{12}	$n_{1.}$
	O_{21}	O_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

$$E_{ij} = \frac{n_{i.} \text{ x } n_{.j}}{n_{..}}$$

6.2.1 THE CHI-SQUARE INDEPENDENCE TEST

To test the independence of two variables

Hypothesis Null and Alternative

H_0 : The row and column variables are independent/not related with each other

(x has no relationship with y)

H_1 : The row and column variables are dependent/ related with each other

(x has relationship with y)

Procedures

1. State the hypothesis and identify the claim.
2. Compute the test value $\chi^2{}_{test} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$.
3. Find the critical value $\chi^2_{\alpha,(r-1)(c-1)}$.
4. Make the decision – reject Ho $\chi^2_{test} > \chi^2_{\alpha,(r-1)(c-1)}$.
5. Draw a conclusion to reject or accept the claim.

Example 5: Chi-Square Independence Test

The data below shows the number of insomnia patient according to their smoking habit in Malaysia.

	Habit	
	Smoking	Not smoking
Insomnia	20	40
Not insomnia	10	80

At $\alpha = 0.01$, Can we say that insomnia is independent with smoking habit?

Example 5: solution

H_0: Insomnia is independent of smoking habit (claim)

H_1: Insomnia is dependent of smoking habit

	Habit		
	Smoking	Not smoking	$n_{i.}$
Insomnia	20	40	$n_{1.} = \mathbf{60}$
Not insomnia	10	80	$n_{2.} = \mathbf{90}$
$n_{.j}$	$n_{.1} = \mathbf{30}$	$n_{.2} = \mathbf{120}$	$n_{..} = 150$

Example 5: solution

O_{ij}	$E_{ij}=\dfrac{n_{i.}\times n_{.j}}{n_{..}}$	$\dfrac{(O_{ij}-E_{ij})^2}{E_{ij}}$
$O_{11}=20$	$E_{11}=\dfrac{60\times 30}{150}=12$	$\dfrac{(20-12)^2}{12}=5.3333$
$O_{12}=40$	$E_{12}=\dfrac{60\times 120}{150}=48$	$\dfrac{(40-48)^2}{48}=1.3333$
$O_{21}=10$	$E_{21}=\dfrac{90\times 30}{150}=18$	$\dfrac{(10-18)^2}{18}=3.5556$
$O_{22}=80$	$E_{22}=\dfrac{90\times 120}{150}=72$	$\dfrac{(80-72)^2}{72}=0.8889$
		$\chi^2_{test}=\sum_{i=1}^{r}\sum_{j=1}^{c}\dfrac{(O_{ij}-E_{ij})^2}{E_{ij}} = 11.1111$

$$\chi^2_{critical} = \chi^2_{0.01,(2-1)(2-1)} = \chi^2_{0.01,1} = 6.6349$$

Since $\left(\chi^2_{test}=11.1111\right)>\left(\chi^2_{0.01,1}=6.6349\right)$, then we reject H_0.

At $\alpha=0.01$, there is sufficient evidence to conclude that insomnia is not independent (or dependent) of smoking habit.

6.2.2 TEST FOR HOMOGENEITY OF PROPORTIONS

- ✓ Concerns the homogeneity or **similarity of two or more population proportions** with regard to the distribution of a certain characteristic.
- ✓ Considers the similarity of two or more population proportions.
- ✓ The procedure is similar to the procedure used to make a test of independence discussed.

Hypothesis Null and Alternative

H0 : $\pi_1 = \pi_2 = ... = \pi_n$

H1 : $\pi_i \neq \pi_j$ for at least $i \neq j$

OR

H0 : All proportions are the same

H1 : At least one proportion is different from the others

Example 6: Homogeneity Test for Proportions

A researcher selected a sample of 50 seniors from each of three area secondary schools and asked each students, " Do you come to school on your own or sent by your parents?". The data are shown in the table.

	SCHOOL 1	SCHOOL 2	SCHOOL 3
Yes	18	22	16
No	32	28	34

At $\alpha = 0.05$, test the claim that the proportion of students who come to school on their own or sent by their parents is the same for all schools.

Example 6: solution

H_0: All proportions are the same
H_1: At least one proportion is different from the others.

OR

$H_0: \pi_1 = \pi_2 = \pi_3$
$H_1: \pi_i \neq \pi_j$ for at least one $i \neq j$

	School 1	School 2	School 3	$n_{i.}$
Yes	18	22	16	$n_{1.} = \mathbf{56}$
No	32	28	34	$n_{2.} = \mathbf{94}$
$n_{.j}$	$n_{.1} = \mathbf{50}$	$n_{.2} = \mathbf{50}$	$n_{.3} = \mathbf{50}$	$n_{..} = 150$

Example 6: solution

O_{ij}	$E_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
$O_{11} = 18$	$E_{11} = \frac{56 \times 50}{150} = 18.6667$	$\frac{(18 - 18.6667)^2}{18.6667} = 0.0238$
$O_{12} = 22$	$E_{12} = \frac{56 \times 50}{150} = 18.6667$	$\frac{(22 - 18.6667)^2}{18.6667} = 0.5952$
$O_{13} = 16$	$E_{13} = \frac{56 \times 50}{150} = 18.6667$	$\frac{(16 - 18.6667)^2}{18.6667} = 0.3810$
$O_{21} = 32$	$E_{21} = \frac{94 \times 50}{150} = 31.3333$	$\frac{(32 - 31.3333)^2}{31.3333} = 0.0142$
$O_{22} = 28$	$E_{22} = \frac{94 \times 50}{150} = 31.3333$	$\frac{(28 - 31.3333)^2}{31.3333} = 0.3546$
$O_{23} = 34$	$E_{23} = \frac{94 \times 50}{150} = 31.3333$	$\frac{(34 - 31.3333)^2}{31.3333} = 0.2270$

Since $\left(\chi^2_{test} = 1.5958\right) < \left(\chi^2_{0.05,2} = 5.9915\right)$, then do not reject H_0.

$$\chi^2_{test} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad 1.5958$$

At $\alpha = 0.05$, there is sufficient evidence to conclude that the proportions of student come to school on their own or sent by their parents is the same for all schools

REFERENCES

1. Montgomery D. C. & Runger G. C. 2011. *Applied Statistics and Probability for Engineers*. 5th Edition. New York: John Wiley & Sons, Inc.
2. Walpole R.E., Myers R.H., Myers S.L. & Ye K. 2011. *Probability and Statistics for Engineers and Scientists*. 9th Edition. New Jersey: Prentice Hall.
3. Navidi W. 2011. *Statistics for Engineers and Scientists.* 3rd Edition. New York: McGraw-Hill.
4. Bluman A.G. 2009. *Elementary Statistics: A Step by Step Approach.* 7th Edition. New York: McGraw–Hill.
5. Triola, M.F. 2006. *Elementary Statistics.*10th Edition. UK: Pearson Education.
6. Weiss, N.A. 2002. *Introductory Statistics*. 6th Edition. United States: Addison-Wesley.
7. Sanders D.H. & Smidth R.K. 2000. *Statistics: A First Course*. 6th Edition. New York: McGraw-Hill.
8. Satari S. Z. et al. Applied Statistics Module New Version. 2015. Penerbit UMP. Internal used.

THE END. Thank You