# Tests for Goodness of Fit

Dr R Sinuvasan

## Computational steps for testing the significance of goodness of fit:

**Step 1 : Framing of hypothesis**

**Null hypothesis** $H_0$ : The goodness of fit is appropriate for the given data set

**Alternative hypothesis** $H_1$ : The goodness of fit is not appropriate for the given data set

**Step 2 : Data**

Calculate the expected frequencies ($E_i$) using appropriate theoretical distribution such as Binomial or Poisson.

**Step 3 :** Select the desired level of significance $\alpha$

**Step 4 : Test statistic**

The test statistic is

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

where $k =$ number of classes

$O_i$ and $E_i$ are respectively the observed and expected frequency of $i^{th}$ class such that $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$ .

If any of $E_i$ is found less than 5, the corresponding class frequency may be pooled with preceding or succeeding classes such that $E_i$'s of all classes are greater than or equal to 5. It may be noted that the value of $k$ may be determined after pooling the classes.

The approximate sampling distribution of the test statistic under $H_0$ is the chi-square distribution with $k$-1-$s$ $d.f$ , $s$ being the number of parametres to be estimated.

**Step 5 : Calculation**

Calculate the value of chi-square as

$$\chi_0^{\ 2} = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

The above steps in calculating the chi-square can be summarized in the form of the table as follows:

**Step 6 : Critical value**

The critical value is obtained from the table of $\chi^2$ for a given level of significance $\alpha$.

**Step 7 : Decision**

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value, at the desired level of significance.

**Example**

Five coins are tossed 640 times and the following results were obtained.

| Number of heads | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Frequency | 19 | 99 | 197 | 198 | 105 | 22 |

Fit binomial distribution to the above data.

***Solution:***

**Step 1** : **Null hypothesis** $H_0$: Fitting of binomial distribution is appropriate for the given data.

**Alternative hypothesis** $H_1$: Fitting of binomial distribution is not appropriate to the given data.

**Step 2** : **Data**

Compute the expected frequencies:

$n$ = number of coins tossed at a time = 5

Let $X$ denote the number of heads (success) in $n$ tosses

$N$ = number of times experiment is repeated = 640

To find mean of the distribution

| $x$ | $f$ | $fx$ |
|---|---|---|
| 0 | 19 | 0 |
| 1 | 99 | 99 |
| 2 | 197 | 394 |
| 3 | 198 | 594 |
| 4 | 105 | 420 |
| 5 | 22 | 110 |
| Total | 640 | 1617 |

Mean : $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{1617}{640} = 2.526$

The probability mass function of binomial distribution is :

$p(x) = {}^nC_x \, p^x \, q^{n-x}, x = 0,1,\ldots, n$

Mean of the binomial distribution is $\bar{x} = np$ .

Hence, $$\hat{p} = \frac{\bar{x}}{n} = \frac{2.526}{5} \approx 0.5$$

For $x = 0$, the equation becomes

$P(X = 0) = P(0) = 5c_0\,(0.5)^5 = 0.03125$

The expected frequency at $x = N\,P(x)$

The expected frequency at $x = 0$ : $N \times P(0)$

$= 640 \times 0.03125 = 20$

We use recurrence formula to find the other expected frequencies.

The expected frequency at $x+1$ is

$$\frac{n-x}{x+1}\left(\frac{p}{q}\right) \times \text{Expected frequency at } x$$

| $x$ | $\frac{n-x}{x+1}$ | $\frac{p}{q}$ | $\frac{n-x}{x+1}\left(\frac{p}{q}\right)$ | Expected frequency at $x = N\,P(x)$ |
|---|---|---|---|---|
| 0 | 5 | 1 | 5 | 20 |
| 1 | 2 | 1 | 2 | 100 |
| 2 | 1 | 1 | 1 | 200 |
| 3 | 0.5 | 1 | 0.5 | 200 |
| 4 | 0.2 | 1 | 0.2 | 100 |
| 5 | 0 | 1 | 0 | 20 |

**Table of expected frequencies:**

| Number of heads | 0 | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| Expected frequencies | 20 | 100 | 200 | 200 | 100 | 20 | 640 |

**Step 3 : Level of significance**

$\alpha = 5\%$

**Step 4 : Test statistic**

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

| Observed frequency ($O_i$) | Expected frequency ($E_i$) | $O_i - E_i$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|
| 19 | 20 | –1 | 1 | 0.050 |
| 99 | 100 | –1 | 1 | 0.010 |
| 197 | 200 | –3 | 9 | 0.045 |
| 198 | 200 | –2 | 4 | 0.020 |
| 105 | 100 | 5 | 25 | 0.250 |
| 22 | 20 | 2 | 4 | 0.200 |
| | | | Total | 0.575 |

$$\chi_0^{\,2} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 0.575$$

**Step 6 : Critical value**

Degrees of freedom = $k - 1 - s = 6 - 1 - 1 = 4$

Critical value for *d.f* 4 at 5% level of significance is 9.488 *i.e.*, $\chi^2_{4,0.05} = 9.488$

**Step 7 : Decision**

As the calculated $\chi_0^{\,2}$(=0.575) is less than the critical value $\chi^2_{4,0.05} = 9.488$, we do not reject the null hypothesis. Hence, the fitting of binomial distribution is appropriate.

**Example**

A packet consists of 100 ball pens. The distribution of the number of defective ball pens in each packet is given below:

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $f$ | 61 | 14 | 10 | 7 | 5 | 3 |

Examine whether Poisson distribution is appropriate for the above data at 5% level of significance.

***Solution:***

**Step 1** : **Null hypothesis** $H_0$: Fitting of Poisson distribution is appropriate for the given data.

**Alternative hypothesis** $H_1$: Fitting of Poisson distribution is not appropriate for the given data.

**Step 2** : **Data**

The expected frequencies are computed as under:

To find the mean of the distribution.

| $x$ | $f$ | $fx$ |
|---|---|---|
| 0 | 61 | 0 |
| 1 | 14 | 14 |
| 2 | 10 | 20 |
| 3 | 7 | 21 |
| 4 | 5 | 20 |
| 5 | 3 | 15 |
| Total | 100 | 90 |

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{90}{100} = 0.9$$

Probability mass function of Poisson distribution is:

$$p(x) = \frac{e^{-m} m^x}{x!}; \; x = 0, 1, \ldots$$

In the case of Poisson distribution mean $(m) = \bar{x} = 0.9$.

At $x = 0$, equation becomes

$$p(0) = \frac{e^{-m}\ m^{0}}{0!} = e^{m} = e^{0.9} = 0.4066 \cdot$$

The expected frequency at $x$ is $N\,P(x)$

Therefore, The expected frequency at 0 is

$$\begin{aligned} &N \times P(0) \\ &= 100 \times 0.4066 \\ &= 40.66 \end{aligned}$$

We use recurrence formula to find the other expected frequencies.

The expected frequency at $x+1$ is

$$\frac{m}{x+1} \times \text{Expected frequency at } x$$

| $x$ | $\frac{m}{x+1}$ | Expected frequency at $x = N\,P(x)$ |
|---|---|---|
| 0 | 0.9 | 40.66 |
| 1 | $\frac{0.9}{2}$ | 36.594 |
| 2 | $\frac{0.9}{3}$ | 16.4673 |
| 3 | $\frac{0.9}{4}$ | 4.94019 |
| 4 | $\frac{0.9}{5}$ | 1.1115 |
| 5 | $\frac{0.9}{6}$ | 0.20007 |

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Expected frequency | 41 | 37 | 16 | 5 | 1 | 0 |

**Step 3 : Level of significance**

$\alpha = 5\%$

**Step 4 : Test statistic**

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

**Step 5 : Calculation**

Test statistic is computed as under:

| Observed frequency ($O_i$) | Expected frequency ($E_i$) | $O_i - E_i$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|
| 61 | 41 | 20 | 400 | 9.756 |
| 14 | 37 | -23 | 529 | 14.297 |
| 10 | 16 | -6 | 36 | 2.250 |
| 7 | 5 | | | |
| 5 } 15 | 1 } 6 | 9 | 81 | 13.5 |
| 3 | 0 | | | |
| | | | Total | 51.253 |

$$\chi_0^{\ 2} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
$$=51.253$$

**Step 6 : Critical value**

Degrees of freedom = $(k - 1 - s) = 4 - 1 - 1 = 2$

Critical value for 2 *d.f* at 5% level of significance is 5.991 *i.e.*, $\chi^2_{2,0.05} = 5.991$

**Step 7 : Decision**

The calculated $\chi_0^{\ 2}$(=51.253) is greater than the critical value (5.991) at 5% level of significance. Hence, we reject $H_0$. i.e., fitting of Poisson distribution is not appropriate for the given data.

## Example

A sample 800 students appeared for a competitive examination. It was found that 320 students have failed, 270 have secured a third grade, 190 have secured a second grade and the remaining students qualified in first grade. The general opinion that the above grades are in the ratio 4:3:2:1 respectively. Test the hypothesis the general opinion about the grades is appropriate at 5% level of significance.

**Step 1** : **Null hypothesis** $H_0$: The result in four grades follows the ratio 4:3:2:1

**Alternative hypothesis** $H_1$: The result in four grades does not follows the ratio 4:3:2:1

**Step 2** : **Data**

Compute expected frequencies:

Under the assumption on $H_0$, the expected frequencies of the four grades are:

$$\frac{4}{10} \times 800 = 320\,; \frac{3}{10} \times 800 = 240; \frac{2}{10} \times 800 = 160; \frac{1}{10} \times 800 = 80$$

**Step 3 : Test statistic**

The test statistic is computed using the following table.

| Observed frequency ($O_i$) | Expected frequency ($E_i$) | $O_i - E_i$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|
| 320 | 320 | 0 | 0 | 0 |
| 270 | 240 | 30 | 900 | 3.75 |
| 190 | 160 | 30 | 900 | 5.625 |
| 20 | 80 | -60 | 3600 | 45 |
| | | | Total | 54.375 |

The test statistic is calculated as

$$\chi_0^{\,2} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 54.375$$

**Step 4 : Critical value**

The critical value of $\chi^2$ for 3 d.f. at 5% level of significance is 7.81 *i.e.*, $\chi^2_{3,0.05} = 7.81$

**Step 5 : Decision**

As the calculated value of $\chi_0^{\,2}$ (=54.375) is greater than the critical value $\chi^2_{3,0.05} = 7.81$, reject $H_0$. Hence, the results of the four grades do not follow the ratio 4:3:2:1.

**Example**

The following table shows the distribution of digits in numbers chosen at random from a telephone directory.

| Digit | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 1026 | 1107 | 997 | 966 | 1075 | 933 | 1107 | 972 | 964 | 853 |

Test whether the occurence of the digits in the directory are equal at 5% level of significance.

**Step 1** : **Null hypothesis** $H_0$: The occurrence of the digits are equal in the directory.

**Alternative hypothesis** $H_1$: The occurrence of the digits are not equal in the directory.

**Step 2** : **Data**

The expected frequency for each digit $= \dfrac{10000}{10} = 1000$

**Step 3** : **Level of significance** $\alpha = 5\%$

**Step 3 : Test statistic**

The test statistic is computed using the following table.

| Observed frequency ($O_i$) | Expected frequency ($E_i$) | $O_i - E_i$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|
| 1026 | 1000 | 26 | 676 | 0.676 |
| 1107 | 1000 | 107 | 11449 | 11.449 |
| 997 | 1000 | 3 | 9 | 0.009 |
| 966 | 1000 | 34 | 1156 | 1.156 |
| 1075 | 1000 | 75 | 5625 | 5.625 |
| 933 | 1000 | 67 | 4489 | 4.489 |
| 1107 | 1000 | 107 | 11449 | 11.449 |
| 972 | 1000 | 28 | 784 | 0.784 |
| 964 | 1000 | 36 | 1296 | 1.296 |
| 853 | 1000 | 147 | 21609 | 21.609 |
| | | | Total | 58.542 |

The test statistic is calculated as

$$\chi_0^{\ 2} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 58.542$$

**Step 4 : Critical value**

Critical value for 9 df at 5% level of significance is 16.919 i.e., $\chi^2_{9,0.05} = 16.919$

**Step 5 : Decision**

Since the calculated $\chi_0^2$ (58.542) is greater than the critical value $\chi^2_{9,0.05} = 16.919$, reject $H_0$. Hence, the digits are not uniformly distributed in the directory.