Chi-Square "Goodness of Fit" Test

Ansar Shahzadi
School of Electrical Engineering & Computer Science
National University of Science and Technology(NUST)

When do we use chi square?

When the data we want to analyze is categorical, a chi-square test, denoted x^2, is usually the appropriate test to use.

The Chi-Square test is a statistical procedure used by researchers to examine the differences between categorical variables in the same population.

Chi-Square "Goodness of Fit" test

In Chi-Square goodness of fit test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution. Chi-Square goodness of fit test determines how well theoretical distribution (such as normal, binomial, or Poisson) fits the empirical distribution. In Chi-Square goodness of fit test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval.

Observed and Expected Value

- **Observed Value**

In probability and statistics, a realization, **observation**, or **observed value**, of a random variable **is** the **value** that **is** actually **observed** (what actually happened).

- **Expected Value**

An **expected value** is simply the number of successful outcomes **expected** in an experiment. The formula of expected value is Np.

Fitting of Probability Distribution

- Fitting of probability distribution to a series of observed data helps to predict the probability or to forecast the frequency of occurrence of the required variable in a certain desired interval.
- There are many probability distributions of which some can be fitted more closely to the observed frequency of the data than others, depending on the characteristics of the variables. Therefore one needs to select a distribution that suits the data well.

Fitting of Probability Distribution

Fitting of probability distribution has following steps

- Estimating the values of parameters, if the parameters are unknown
- Calculating the probabilities
  - for x=0,1, 2,…,n
  - for class intervals
- Then calculate expected frequency using following formula

Expected frequency= N.P(X=x), if p.d is discreate

Expected frequency= N.P(a≤X≤b), if p.d is continuous

Procedure for Chi-Square Goodness of Fit Test

- **State the Hypothesis:**

Null hypothesis: In Chi-Square goodness of fit test, the null hypothesis assumes that there is no significant difference between the observed and the expected value.

Alternative hypothesis: In Chi-Square goodness of fit test, the alternative hypothesis assumes that there is a significant difference between the observed and the expected value.

- **Level of Significance:**

Define the value of α. Where α is the probability of reject null hypothesis, When it is true.

Procedure for Chi-Square Goodness of Fit Test

Test Statistics

Find the Value of Chi-Square goodness of Fit test using the following formula.

$$\chi^2 = \sum \frac{\left(f_o - f_e\right)^2}{f_e}$$

Where,

f_o =observed frequency

f_e= expected frequency

Procedure for Chi-Square Goodness of Fit Test

- **Critical Value:**

The critical region is $\chi^2 \geq {\chi^2}_{\alpha,(df)}$

Where degree of freedom= k-1-number estimated of parameters

- **Computations:**

Compute the expected values and value of χ^2.

- **Conclusion:**

Reject H_0; if the calculated of χ^2 exceeds the ${\chi^2}_{\alpha,(df)}$

Accept H_0; otherwise

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Example#1

Suppose that 5 coins are tossed simultaneously 1000 times and the numbers of heads were observed is given below:

No of Heads	0	1	2	3	4	5
Frequency	38	144	342	287	164	25

Fit a binomial distribution test the goodness of fit.

Chi-Square Goodness of Fit Test For Binomial Distribution

- **Hypothesis**

H_0: The Population distribution is a binomial with n=5, but p is unknown

H_1: The Population distribution is not a binomial with n=5,

- **Level of Significance**

We choose the α=5%

- **Test Statistics**

We use the test statistics $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

Chi-Square Goodness of Fit Test For Binomial Distribution

Critical Value:

The critical region is $\chi^2 \geq \chi^2{}_{0.05,(4)}$=9.49

Where degree of freedom= 6-1-1

Computations:

For find the value of p, we first compute mean no of heads. Thus

Mean=$\frac{\sum fx}{\sum f}$= $\frac{2470}{1000}$

$\bar{x}$ =2.47

X	0	1	2	3	4	5
f	38	144	342	287	164	25
fx	0	144	684	861	656	125

Here we know that $\bar{x}$=np, so that $\hat{p}$=$\frac{\bar{x}}{n} = \frac{2.47}{5}$=0.494

Chi-Square Goodness of Fit Test For Binomial Distribution

Hence the expected frequencies are the terms in the binomial distribution is

$$f_e = 1000P(X=x) = 1000\binom{5}{x}\left(0.494\right)^{x} . \left(0.506\right)^{5-x}$$

Next we calculate the value of χ^2 as follows

Chi-Square Goodness of Fit Test For Binomial Distribution

Numbers of Heads	f_o	f_e	f_o-f_e	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
0	38	33.2	4.8	23.04	0.69
1	144	161.9	-17.9	320.41	1.98
2	342	316.2	25.8	665.64	2.15
3	287	308.7	-21.7	470.89	1.53
4	164	150.7	13.3	176.89	1.17
5	25	29.4	-4.4	19.36	0.66
Total	1000	1000			χ^2=8.18

Chi-Square Goodness of Fit Test For Binomial Distribution

- **Conculsion**

Since the ${\chi^2}_{cal}$=8.18 does not fall in the critical region, we are therefore unable to reject our null hypothesis. Here we conclude that number of heads is a binomial distribution.

Question#2

The following table shows a data set of the number of errors found in a total of n=1000 software products. Is it possible that the number of errors has a Poisson distribution with mean 3?

Number of errors	0	1	2	3	4	5	6	7	8
Frequency	35	165	235	294	165	70	23	11	2

Question#3

A random sample of 500 long distance telephone calls revealed the following distribution of call length (in minutes).

LENGTH (IN MINUTES)	FREQUENCY
0-under 5	48
5-under 10	84
10-under 15	164
15-under 20	126
20-under 25	50
25-under 30	28

- Compute the mean and standard deviation of this frequency distribution.
- At the 0.05 level of significance, does call length follow a normal distribution?