Goodness of fit and contingency tables

Method for testing goodness of fit:

1. Determine which distribution would conceptually be most appropriate
2. Set significance level
3. Estimate parameters (if necessary) from observed data
4. Form hypotheses H_0 and H_1
5. Calculate expected frequencies
6. Combine expected frequencies so that none are < 5
7. Find degrees of freedom
8. Calculate critical value of χ^2 from the table
9. Calculate $\sum \frac{(O_i - E_i)^2}{E_i}$
10. See if the value is significant and draw conclusion

X^2 is distributed with a chi squared distribution χ^2_ν
Where $\nu =$ degrees of freedom

$$\text{The number of degrees of freedom} = \text{Number of classes (after combining)} - 1$$

0.1 Testing a Binomial distribution as a model

The data in the table is thought to be modelled by a binomial $B(10, 0.2)$. Use the table for the binomial cumulative distribution function to find expected values, and conduct a test to see if this is a good model. Use a 5% significance level.

x	0	1	2	3	4	5	6	7	8
Freq of x	12	28	28	17	7	4	2	2	0

Define Hypotheses

$$H_0 : \text{A B(10,0.2) distribution is suitable for the results}$$

$$H_1 : \text{The distribution is not suitable for the results}$$

Calculate the sum of frequencies

$$N = 100$$

Complete the table of probabilities and expected frequencies, expected frequency=probability×N

x	0	1	2	3	4	5	6	7	8
$p(x)$	0.1074	0.2684	0.3020	0.2013	0.0881	0.0264	0.0055	0.0008	0.0001
Expected freq	10.74	26.84	30.20	20.13	8.81	2.64	0.55	0.08	0.01

As expected frequencies need to be greater than or equal to five, combine all probabilities greater than or equal to four

x	0	1	2	3	$\geqslant 4$
O_i	12	28	28	17	15
E_i	10.74	26.84	30.20	20.13	12.09
$\frac{(O_i - E_i)^2}{E_i}$	0.1478	0.0501	0.1603	0.4867	0.7004

Find the value of ν

$$\nu = 5 - 1 = 4$$

Find the value of X^2

$$X^2 = 0.1478 + 0.0501 + 0.1603 + 0.4867 + 0.7004 = 1.5453$$

Compare the value of X^2 to the value on the tables corresponding to the 5% significance level and $\nu = 4$

$$9.488 > 1.5453$$

Write conclusion

Not in critical region so insufficient evidence to reject H_0, binomial is a possible model

0.2 What to do when p is not given

A study of the number of girls in families with 5 children was done on 100 such families. The results are summarised in the following table.

Num girls(r)	0	1	2	3	4	5
Frequency(f)	13	18	38	20	10	1

Test, at the 5% significance level, whether or not a binomial distribution is a good model.

State hypotheses

H_0: The binomial distribution is a good model
H_1: The binomial distribution is not a suitable model

Calculate the mean

$$\overline{x} = \frac{0\times 13+1\times 18+2\times 38+3\times 20+4\times 10+5\times 1}{100} = 1.99$$

Divide the mean by n, the number of children in the families, 5, to find p.

$$p = \frac{1.99}{5} = 0.398$$

Using the values of n and p, find the probability the value is a certain number, multiply by 100 to find the expected value.

r	0	1	2	3	4	5
p(r)	0.079	0.261	0.3456	0.229	0.0755	0.0009
E_i	7.91	26.14	34.56	22.85	7.55	0.09

Use the values in the two above tables to find the values of O_i, E_i and X^2, combine expected values when under 5

r	0	1	2	3	>3	Total
O_i	13	18	38	20	11	
E_i	7.91	26.14	34.56	22.85	8.54	
$\frac{(O_i-E_i)^2}{E_i}$	3.28	2.53	0.34	0.36	0.71	7.22

Calculate the degrees of freedom, subtracting one for a constant frequency sum and one for the estimated p

$$\nu = 5-1-1 = 3$$

Find the value of χ^2_3 at a 5% significance level

$$\chi^2_3 = 7.815$$

Compare the value of χ^2_3 to X^2 to determine the correct hypothesis

$$7.22 < 7.81$$

Not in critical region, so not significant, do not reject H_0, binomial is a suitable model

0.3 Testing a poisson distribution as a model

The numbers of telephone calls arriving at an exchange in six-minute periods were recorded over a period of 8 hours, with the following results

Num calls(r)	0	1	2	3	4	5	6	7	8
Freq(f)	8	19	26	13	7	5	1	1	0

Can these results be modelled by a Poisson distribution? Test at the 5% significance level

Calculate λ (the mean)

$$\lambda = \overline{x} = \frac{0 \times 8 + 1 \times 19 + 2 \times 26 + 3 \times 13 + 4 \times 7 + 5 \times 5 + 6 \times 1 + 7 \times 1 + 9 \times 0}{8 + 19 + 26 + 13 + 7 + 5 + 1 + 1 + 0} = 2.2$$

Use this value of λ to find $P(r)$ and $E(r)$ by multiplying by 80

r	$P(r)$	Expected freq of r
0	0.1108	8.864
1	0.2438	19.504
2	0.2681	21.448
3	0.1966	15.728
4	0.1082	8.656
5	0.0476	3.808
6	0.0174	1.392
$\geqslant 7$	0.0075	0.6

Use the two above tables to calculate O_i, E_i and X^2, combining classes where needed

r	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
0	8	8.864	0.0842
1	19	19.504	0.0130
2	26	21.448	0.9661
3	13	15.728	0.4732
4	7	8.656	0.3168
$\geqslant 5$	7	5.8	0.2483

Find the value of X^2 by finding the sum of $\frac{(O_i - E_i)^2}{E_i}$

$$X^2 = 0.0842 + 0.0130 + 0.9661 + 0.4732 + 0.3168 + 0.2483$$

Calculate the degrees of freedom, subtracting for constant probability and estimated λ

$$\nu = 6 - 1 - 1 = 4$$

Use the value for the degrees of freedom and significance level to find χ^2

$$\chi^2_4 = 9.488$$

Compare the value of χ^2_4 to the value of X^2 to determine the correct hypothesis

$$2.1016 < 9.488$$

Value not in critical region, non significant, accept H_0 insufficient evidence to reject H_0