

χ^2 GOODNESS OF FIT TESTS

Questions on probability have often talked about a **fair** die or **random** numbers or the distribution of colours for a particular flower.

Now, we wish to go behind the scenes as it were and ask the question: “Is the die really fair?”

We may use the chi-squares goodness of fit test to see how well a set of data fits the theoretical distribution. The distribution for a die is Uniform or

Rectangular, with the probability of each outcome being $\frac{1}{6}$. So if we throw our die 120 times and obtained the following results:

Number	1	2	3	4	5	6
Frequency (observed)	17	16	19	17	21	30

If the die were fair we should expect $\frac{1}{6} \times 120 = 20$ of each number.

Expected frequency	20	20	20	20	20	20
--------------------	----	----	----	----	----	----

Comparing the observed and expected frequencies suggest a die biased in favour of sixes.

But we need to do a statistical test to ascertain if these differences between observed and expected values are significant. We need a “thermometer”.

Let us first of all state a Null Hypothesis. This can be done in a sentence indicating that we believe there is nothing significant in our results.

H₀: A Rectangular Distribution is a suitable model. And we need an alternate hypothesis which suggests abnormality in the die.

H₁: A rectangular distribution is not a suitable model.

We need a **statistic** to measure the degree of discrepancy between the observed and expected values.

If we look at the differences (observed – expected) and sum these, the answer would be zero.

If we **square** these before adding, we would not get zero but we would get **identical answers** for both sets of data below:

Observed frequency	17	16	19	17	21	30
Expected frequency	20	20	20	20	20	20

$$\sum(O - E) = 0, \quad \sum(O - E)^2 = 9+16+1+3+1+100 = \mathbf{130}$$

Observed frequency	197	196	199	197	201	210
Expected frequency	200	200	200	200	200	200

$$\sum(O - E) = 0, \quad \sum(O - E)^2 = 9+16+1+3+1+100 = \mathbf{130}$$

But we don’t want identical answers since clearly; the second set of figures seems to represent a more normal die in terms of percentage difference between observed and expected values particularly on the number of sixes.

Expressing each $(O - E)^2$ in proportion to its Expected value before adding will mean that the identical answers will no longer appear.

The measure of discrepancy will be different. I.e. the calculated χ^2 statistic, $\chi^2_{(calc)} = \sum \frac{(O-E)^2}{E} = \frac{3^2}{20} + \frac{4^2}{20} + \frac{1^2}{20} + \dots + \frac{10^2}{20}$

Which is equal to **6.8** for the first set of data and equal to a much smaller amount of **0.68** for the second set of data.

We need to decide if either of these values are **significant** if we are to make a statement about the fairness of the die.

A chi-squared table of values will help us find the critical value for this particular problem which had **six** "expected" cells to be filled.

If the **expected** values are largely different from the **observed** values, this will be reflected in a large calculated value and if it is larger than the critical value found from tables, we will conclude that the null hypothesis is false and there **is** an element of bias in the die.

Or in other words: **A rectangular distribution is not a suitable model.**

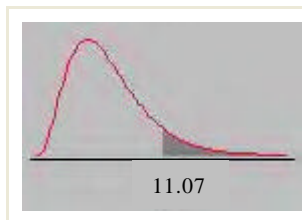
To find this critical value look to the tables. The values are entered and we look at the 5% value but must establish the row to look along.

The rows give the degrees of freedom: the degree of play we have when calculating the expected values.



Out of the six cells to be filled with expected values, only five have to be worked out and the last entered by subtraction. So we use 5 degrees of freedom for this test.

$$\chi^2_{5(5\%)} = 11.07$$



If our calculated value lies in the shaded region which it has a 5% chance of doing, then we reject the null hypothesis that the rectangular distribution is a good fit.

Taking $\chi^2_{(calc)}$ as **6.8**:

Since $\chi^2_{(calc)} < 11.07$, we have no grounds to reject H_0 and conclude that

A rectangular distribution is a suitable model.

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52

Testing the fit to other distributions.

We can test if data could have come from a Binomial, Poisson or Normal Distribution and also if data conforms to any prior pattern which we believe to be true.

For example it is believed that three major political parties gain votes in the ratio A(5): B(4): C(1)3. 60 people were asked who they were going to vote for and the results were: 23 for A, 20 for B and 17 for C. Is this consistent with the ratios 5: 4: 3?

Ho: The data is consistent with the ratios 5:4:3.

H₁: The data is not consistent with the ratios 5:4:3. We use the ratios 5:4:3 to work out the expected values as 25 - 20 - 15, put them in a table and perform a χ^2 goodness of fit test. χ^2 (calc) = $\sum \frac{(O-E)^2}{E} = \frac{2^2}{25} + \frac{0^2}{20} + \frac{2^2}{15} = 0.43$

Observed values	23	20	17
Expected values	25	20	15

The value **0.43** is not greater than the critical value **5.99** for **two** degrees of freedom and so we don't reject Ho
The distribution with ratios 5:4:3 is a suitable model.

Fitting a Binomial distribution

A survey of 200 families with exactly 4 children gave the following results for the number of girls.

Number of girls	0	1	2	3	4
Frequency (observed)	15	68	69	38	10

Test the suggestion that the distribution of girls in a family of four children is Binomial with n = 4 and p = 0.5.

Ho: The data comes from the Binomial distribution B(4,0.5).

H₁: The data does not come from a Binomial distribution. We use the probability density function to get the expected values:

X:	0	1	2	3	4	χ^2 (calc) = 10.84, $\chi^2_{4(5\%)} = 9.49$
P:	0.0625	0.25	0.375	0.25	0.0625	
200 x P	12.5	50	75	50	12.5	

Reject Ho. Based on the evidence, the data does not come from a Binomial distribution.

Fitting a Poisson distribution

A survey of 200 families with exactly 4 children gave the following results for the number of girls.

Number of girls	0	1	2	3	4
Frequency (observed)	15	68	69	38	10

Test the suggestion that the distribution of girls in a family of four children is Binomial with $n = 4$ and $p = 0.5$.

H₀: The data comes from the Binomial distribution **B(4,0.5)**.

H₁: The data does not come from a Binomial distribution. We use the probability density function to get the expected values:

X:	0	1	2	3	4
P:	0.0625	0.25	0.375	0.25	0.0625
200 x P	12.5	50	75	50	12.5

Observed frequency	15	68	69	38	10
Expected frequency	12.5	50	75	50	12.5

$$\chi^2_{(calc)} = 10.84, \chi^2_{4(5\%)} = 9.49$$

Reject H₀. Based on the evidence, the data does not come from a Binomial distribution.

Fitting a Normal distribution

The times taken for a group of children to complete a mathematical puzzle is shown in the following table:

Test that the data comes from a Normal Distribution. **H₀:** The data comes from a Normal distribution.

H₁: The data is not Normal

Times taken (secs)	0 - 5	6 - 10	11 - 15	16 - 20	21 - 25	26 - 30	Over 31
Number of children	13	21	42	63	38	16	7

This is a continuous distribution with class boundaries $0 < t < 5.5$, $5.5 < t < 10.5$ etc but to work out the expected probabilities, and then the expected frequencies, we need the mean and standard deviation of the distribution.

We use the data to give an estimate of the mean as 17.18 and a variance of 51.48. (using mid points etc.)

We then use standardized normal tables to work out the probabilities for each class. For instance $P(5.5 < t < 10.5) = P\left(\frac{5.5 - 17.18}{\sqrt{51.48}} < z < \frac{10.5 - 17.18}{\sqrt{51.48}}\right)$

$= P(-1.72 < z < -0.93) = 0.124$ from tables and an expected frequency for the class $5.5 < t < 10.5$ will be $0.124 \times 200 = 24.8$. The rest of the expected frequencies are

Expected frequency	10.22	24.8	45.6	54.2	39.8	18.2	6.3
--------------------	-------	-------------	------	------	------	------	-----

$$\chi^2_{(calc)} = 3.18, \chi^2_{4(5\%)} = 9.49$$

Since we have estimated the mean **and** variance we have 2 less degrees of freedom. **Df = 7 - 1 - 2 = 4**

We have no evidence to reject H₀. The Normal distribution is a good model for the data.