

6A Goodness of Fit

Goodness of fit:

A dice is thrown 120 times. If the dice is unbiased, complete the expected frequencies:

Number, n	1	2	3	4	5	6
Observed frequency	23	15	25	18	21	18
Expected frequency						

Null hypothesis, H_0 :

Alternative hypothesis, H_1 :

O_i =

E_i =

The measure of goodness of fit is:

What happens to X^2 as the fit gets worse?

An alternative (easier) way of calculating X^2 :

1. Billy and Mel each have two 4-sided spinners numbered 1–4. They each carry out experiments, where they spin their spinners at the same time, and add the scores together. After each student has carried out 160 experiments, the frequency distributions are as follows:

Number, n	2	3	4	5	6	7	8
Observed by Billy (O_i)	12	15	22	41	33	21	16
Observed by Mel (O_i)	6	12	21	37	35	29	20
Expected (E_i)	10	20	30	40	30	20	10

Both Billy and Mel believe that their spinners are fair.

- a) State the null and alternative hypotheses for the experiment.

One of the students has a biased spinner.

- b) Calculate the goodness of fit for both students, and determine which of them is most likely to have the biased spinner.

n	2	3	4	5	6	7	8	Total
Billy $(O - E)^2/E$	0.4							
Mel $(O - E)^2/E$								

6B Degrees of Freedom and the Chi-Squared Family of Distributions

Degrees of freedom:

Constraint (restriction):

Cell:

Number of degrees of freedom =

A coin is flipped 100 times. How many degrees of freedom are there? Why?

A dice is rolled 120 times. How many degrees of freedom are there? Why?

χ^2 is approximated well by a χ^2 distribution as long as ...

If any of the expected values are less than 5:

Notation: χ^2_v means

Which member of the χ^2 family do you use to approximate χ^2 ?

1. In a sample of 100 households, the expected number of dogs is as follows:

Dogs	0	1	2	3	4	5	>5	Total
Expected	55	20	10	7	4	3	1	100

Select an appropriate chi-squared distribution to model the goodness of fit, χ^2 , for these data.

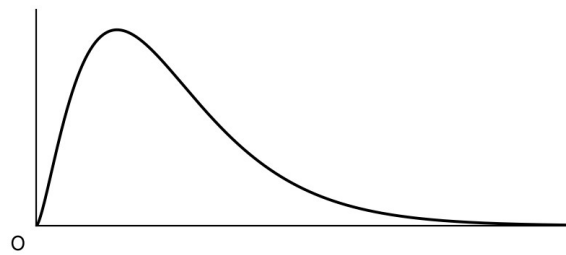
Dogs	0	1	2	3		Total
Expected						

Critical value notation: $\chi^2_{v(5\%)}$ or $\chi^2_{v(0.05)}$ means

Percentage points of the χ^2 distribution (the value exceeded with the probability shown):

v	0.995	0.990	0.975	0.950	0.100	0.050	0.025	0.010
1	0.000	0.000	0.001	0.004	2.705	3.841	5.024	6.635
2	0.010	0.020	0.051	0.103	4.605	5.991	7.378	9.210
3	0.072	0.115	0.216	0.352	6.251	7.815	9.348	11.345
4	0.207	0.297	0.484	0.711	7.779	9.488	11.143	13.277
5	0.412	0.554	0.831	1.145	9.236	11.070	12.832	15.086
6	0.676	0.872	1.237	1.635	10.645	12.592	14.449	16.812
7	0.989	1.239	1.690	2.167	12.017	14.067	16.013	18.475
8	1.344	1.646	2.180	2.733	13.362	15.507	17.535	20.090

2. With $v = 5$ find the value of χ^2 that is exceeded with 0.05 probability.



3.

- a) Find the following critical values: i $\chi^2_{3(95\%)}$ ii $\chi^2_{4(10\%)}$

- b) Find the smallest values of y such that: i $P(\chi^2_2 > y) = 0.95$ ii $P(\chi^2_4 > y) = 0.99$

6C Testing a Hypothesis

Significance level, α :

Critical value, $\chi^2_{v(\alpha)}$:

If X^2 exceeds the critical value ...

Watch out: a hypothesis test for goodness of fit is always

1. In an experiment where a dice is rolled 120 times, the frequency distribution is to be compared to a discrete uniform distribution as shown:

Number on dice, n	1	2	3	4	5	6	Total
Observed frequency, O_i	23	15	25	18	21	18	120
Expected frequency, E_i	20	20	20	20	20	20	120
$(O - E)^2/E$							

Test, at the 5% significance level, whether or not the observed frequencies could be modelled by a discrete uniform distribution.

2. Alan has two identical 4-sided spinners with the numbers 1–4 written on each of them. He carries out experiments, where he spins both of his spinners at the same time, and adds the scores together. After 160 experiments, the frequency distribution is as follows:

Number, n	2	3	4	5	6	7	8	Total
Observed by Alan (O_i)	14	11	26	33	42	18	16	160
Expected (E_i)	10	20	30	40	30	20	10	160
$(O - E)^2/E$								

The table also shows the expected distribution of the scores if both spinners are unbiased. Test, at the 2.5% significance level, whether the observed frequencies could be modelled by the expected distribution shown.

3. A school conducted a survey into the impact that a new exercise club was having on students. Prior to the new club starting, 60% of students said they had no regular exercise, 30% reported exercising once a week and 10% reported exercising more than once a week. After the new club started, they surveyed the 150 students to find out how often they exercised.

	No regular exercise	Once a week	More than once a week	Total
Frequency	73	57	20	150

Based on these data, is there evidence of a change in attitude to exercise following the introduction of the new club? Test the data at a 5% significance level.

	No regular exercise	Once a week	More than once a week	Total
Observed	73	57	20	150
Expected				
$(O - E)^2/E$				

6D Testing the Goodness of Fit with Discrete Data

The steps to test goodness of fit with discrete data:

1

2

3

4

5

6

7

8

9

10

11

Discrete uniform distribution: conditions

$P(X = x_r) =$ Number of degrees of freedom =

1. 100 digits between 0 and 9 are selected from a table with the frequencies as shown below.

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	11	8	8	7	8	9	12	9	13	15

Could the digits be from a random number table? Test at the 0.05 level.

[illegible]

Binomial distribution: conditions

$P(X = r) =$

Frequency, $f_r =$

Estimating p from the observed values: $p =$

If p is not estimated by calculation: $v =$

If p is estimated by calculation: $v =$

2. The data in the table are thought to be modelled by a binomial distribution $B(10, 0.2)$. 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
Frequency	12	28	28	17	7	4	2	2	0
P(X = x)									
Expected									

x	0	1	2	3	
O_i					
E_i					
(O - E)²/E					

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

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

It is suggested that the distribution may be modelled by a binomial distribution with $p = 0.5$.

a) Give reasons why this might be so.

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

r	0	1	2	3	4	5
O_i						
E_i						

r				
O_i				
E_i				
(O - E)²/E				

4. Look at the data in the previous question. By estimating a suitable value of the parameter, p , carry out a test, at the 5% significance level, to determine whether a binomial distribution is a suitable model for the data.

r	P(r)	E_i
0		
1		
2		
3		
4		
5		

r	O_i	E_i	O²/E
0			
1			
2			
3			
Total			

Poisson distribution: conditions

$P(X = r) =$ Frequency, $f_r =$

Grouping the tail: $P(X \geq n) =$

Estimating λ from the observed values: $\lambda =$

If λ is not estimated by calculation: $v =$

If λ is estimated by calculation: $v =$

5. 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.

Number of calls, r	0	1	2	3	4	5	6	7	8
Frequency, f _r	8	19	26	13	7	5	1	1	0

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

r	$P(X = r)$	Expected frequency
0		
1		
2		
3		
4		
5		
6		
7 or more		

r	O _i	E _i	$(O - E)^2/E$
0			
1			
2			
3			
4			

6E Using Contingency Tables

Contingency table:

The results of A level Maths passes at two schools:

Pass	A	B	C	Totals
School X	18	12	20	50
School Y	26	12	32	70
Totals	44	24	52	120

This is called a ... contingency table. Hypotheses:

If H_0 is true, $P(\text{A grade and school X}) =$

Expected frequency =

Expected	A	B	C	Totals
School X				50
School Y				70
Totals	44	24	52	120

Degrees of freedom for an $h \times k$ contingency table: $v =$

Watch out: if an expected frequency is less than 5 ...

1. Conduct a goodness-of-fit test, at the 5% significance level, on the data given above for the two schools X and Y.

	O_i	E_i	$(O - E)^2/E$
X, A	18		
X, B	12		
X, C	20		
Y, A	26		
Y, B	12		
Y, C	32		
Total			

2. During the trial of a new drug, 60 volunteers out of 200 were treated with the drug. Those who experienced relief of their symptoms and those who did not were recorded as in the table.

	Relief	No relief	Totals
Treated	10	50	60
Not treated	40	100	140
Totals	50	150	200

Use a suitable test to see if there is any association between treatment with the drug and relief of symptoms. Use a 5% significance level.

Expected	Relief	No relief
Treated		
Not treated		

	O_i	E_i	$(O - E)^2/E$
Total			

6F Apply Goodness-of-Fit Tests to Geometric Distributions

Geometric distribution: conditions

$P(X = r) =$ Frequency, $f_r =$

Grouping the tail (choosing the cut-off n):

Estimating p from the observed frequencies: $p =$

Degrees of freedom:

1. Sarah has a large DVD collection. Every week she picks DVDs off the shelf at random until she finds one that she would like to watch. Sarah thinks that there is about a 50% chance she will be in the mood to watch any particular DVD. Over the course of a year she records the number of DVDs she picks off the shelf before finding one she would like to watch. The results are recorded in the frequency table below.

Number of DVDs	1	2	3	4	Total
Observed frequency O_i	33	12	5	2	52

- a) Calculate the expected frequencies if the number of DVDs considered is modelled as a $\text{Geo}(0.5)$ random variable.

Number of DVDs	1	2	3		Total
Observed frequency O_i	33	12	5	2	52
Expected frequency E_i					
$(O - E)^2/E$					

Sarah wants to check if her guess that there is a 50% chance she'll watch any particular DVD is supported by the data.

b) Formulate the null and alternative hypotheses.

c) Is Sarah right in her assumption? Test at the 5% significance level.