5 GOODNESS OF FIT TESTS

Objectives

After studying this chapter you should

- be able to calculate expected frequencies for a variety of probability models;
- be able to use the χ^2 distribution to test if a set of observations fits an appropriate probability model.

5.0 Introduction

The chi-squared test is a particular useful technique for testing whether observed data are representative of a particular distribution. It is widely used in biology, geography and psychology.

Activity 1 How random are your numbers?

Can you make up your own table of random numbers? Write down 100 numbers 'at random' (taking values from 0 to 9). Do this without the use of a calculator, computer or printed random number tables. Draw up a frequency table to see how many times you wrote down each number. (These will be called your **observed** frequencies.)

If your random numbers really are random, roughly how many of each do you think there ought to be? (These are referred to as **expected** frequencies.)

What model are you using for this distribution of expected frequencies?

What assumptions must you make in order to use this model?

Do you think you were able to fulfil those assumptions when you wrote them down?

Can you think of a way to test whether your numbers have a similar frequency distribution to what we would expect for true random numbers?

For analysing data of the sort used in Activity 1 where you are comparing observed with expected values, a chart as shown opposite is a useful way of writing down the data.

	Frequency	
Number	Observed, O_i	Expected, E_i
1		
2		
3		
4		
.		
.		
.		

5.1 The chi-squared table

For your data in Activity 1, try looking at the differences $O_i - E_i$.

What happens if you total these?

Unfortunately the positive differences and negative differences always cancel each other out and you always have a zero total.

To overcome this problem the differences $O_i - E_i$ can be squared.

So $\sum(O_i - E_i)^2$ could form the basis of your 'difference measure'. In this particular example, however, each figure has an equal expected frequency, but this will not always be so (when you come to test other models in other situations). The importance assigned to a difference must be related to the size of the expected frequency. A difference of 10 must be more important if the expected frequency is 20 than if it is 100.

One way of allowing for this is to divide each squared difference by the expected frequency for that category.

Here is an example worked out for you:

Number	Observed frequency O_i	Expected frequency E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	11	10	1	1	0.1
1	12	10	2	4	0.4
2	8	10	–2	4	0.4
3	14	10	4	16	1.6
4	7	10	–3	9	0.9
5	9	10	–1	1	0.1
6	9	10	–1	1	0.1
7	8	10	–2	4	0.4
8	14	10	4	16	1.6
9	8	10	–2	4	0.4
					6.0

For this set of 100 numbers $\sum \frac{\left(O_i - E_i\right)^2}{E_i} = 6$

But what does this measure tell you?

How can you decide whether the observed frequencies are close to the expected frequencies or really quite different from them?

Firstly, consider what might happen if you tried to test some true random numbers from a random number table.

Would you actually get 10 for each number?

The example worked out here did in fact use 100 random numbers from a table and not a fictitious set made up by someone taking part in the experiment.

Each time you take a sample of random numbers you will get a slightly different distribution and it would certainly be surprising to find one with all the observed frequencies equal to 10. So, in fact, each different sample of 100 true random numbers will give a different value for

$$\sum \frac{\left(O_i - E_i\right)^2}{E_i}$$

The distribution of

$$\sum \frac{\left(O_i - E_i\right)^2}{E_i}$$

is approximately χ_v^2 where the parameter v is termed the **degrees of freedom**.

For any χ^2 **goodness of fit test**, the number of degrees of freedom shows the number of independent free choices which can be made in allocating values to the expected frequencies. In these examples, there are ten expected frequencies (one for each of the numbers 0 to 9). However, as the total frequency must equal 100, only nine of the expected frequencies can vary independently and the tenth one must take whatever value is required to fulfil the 'constraint'. To calculate the number of degrees of freedom

$$v = \text{ number of classes or groups } - \text{ number of constraints.}$$

Here there are ten classes and one constraint, so

$$\begin{aligned} v &= 10 - 1 \\ &= 9 \end{aligned}$$

Significance testing

A high value of χ^2 implies a poor fit between the observed and expected frequencies, so the upper tail of the distribution is used for most hypothesis testing in goodness of fit tests.

From χ^2 tables, only 5% of all samples of true random numbers will give a value of χ_9^2 greater than 16.919. Thus if the value of

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} < 16.919$$

it would support the view that the numbers are random. If not, there is evidence, at the 5% significance level, to suggest that the numbers are not truly random.

What do you conclude from the worked example above, where $\chi^2 = 6$?

The above procedure may be summarised usefully as a hypothesis test as follows:

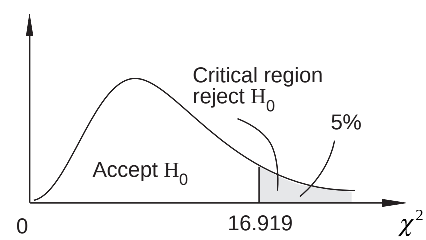
H_0: numbers are random

H_1: numbers are not random

Significance level, $\alpha = 0.05$

Degrees of freedom, $v = 10 - 1 = 9$

Critical region is $\chi^2 > 16.919$



Test statistic is $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 6$

This value does not lie in the critical region. There is no evidence, at the 5% significance level, to suggest that the numbers are not random.

Activity 2

What happens when you test your made up 'random' numbers? Is their distribution close to what you would expect for true random numbers?

Example

Nadir is testing an octahedral die to see if it is biased. The results are given in the table below.

Score	1	2	3	4	5	6	7	8
Frequency	7	10	11	9	12	10	14	7

Test the hypothesis that the die is fair.

Solution

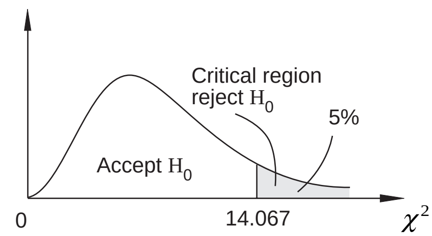
H_0: die is fair

H_1: die is not fair

Significance level, $\alpha = 0.05$

Degrees of freedom, $v = 8 - 1 = 7$

Critical region is $\chi^2 > 14.067$



As before, the expected frequencies are based on a uniform distribution which gives each E_i as

$$\frac{1}{8}(7+10+11+9+12+10+14+7) = 10$$

Hence

Score	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	7	10	–3	9	0.9
2	10	10	0	0	0
3	11	10	1	1	0.1
4	9	10	–1	1	0.1
5	12	10	2	4	0.4
6	10	10	0	0	0
7	14	10	4	16	1.6
8	7	10	–3	9	0.9
					4.0

Thus the test statistic is $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 4.0$

This value does not lie in the critical region. There is no evidence, at the 5% significance level, to suggest that the die is not fair.

Exercise 5A

1. Nicki made a tetrahedral die using card and then tested it to see whether it was fair. She got the following scores:

Score	1	2	3	4
Frequency	12	15	19	22

Does the die seem fair?

2. Joe has a die which has faces numbered from 1 to 6. He got the following scores:

Score	1	2	3	4	5	6
Frequency	17	20	29	20	18	16

He thinks that the die may be biased.
What do you think?

3. The table below shows the number of pupils absent on particular days in the week.

Day	M	Tu	W	Th	F
Number	125	88	85	94	108

Find the expected frequencies if it is assumed that the number of absentees is independent of the day of the week.

Test, at 5% level, whether the differences in observed and expected frequencies are significant.

4. Over a long period of time, a research team monitored the number of car accidents which occurred in a particular county. The following table summarises the data relating to the day of the week on which the accident occurred.

Day	M	Tu	W	Th	F	S	Su
Number of accidents	60	54	48	53	53	75	77

Investigate the hypothesis that these data are a random sample from a uniform distribution.
(AEB)

5. Entrance to, and exit from, a large departmental store is via one of four sets of doors. The number of customers entering or leaving the store is counted at each set of doors for a period of time with the following results.

Set of doors	North	South	East	West
Number of customers	327	402	351	380

It is claimed that the numbers of customers using each of the four sets of doors is the same. Investigate this claim.

6. The proportions of blood types O, A, B and AB in the general population of a particular country are known to be in the ratio 49:38:9:4, respectively. A research team, investigating a small isolated community in the country, obtained the following frequencies of blood type.

Blood type	O	A	B	AB
Frequency	87	59	20	4

Test the hypothesis that the proportions in this community do not differ significantly from those in the general population.

5.2 Discrete probability models

Four identical six-sided dice, each with faces marked 1 to 6, are rolled 200 times. At each rolling, a record is made of the number of dice whose score on the uppermost face is even. The results are as follows.

Number of even scores (x_i)	0	1	2	3	4
Frequency (f_i)	10	41	70	57	22

Why might a binomial model describe the distribution of X?

What values would you suggest for the two binomial parameters n and p?

Based upon your suggested values, how would you then obtain the expected frequencies for comparative purposes?

The number of computer malfunctions per day is recorded for 260 days with the following results.

Number of malfunctions (x_i)	0	1	2	3	4	5
Number of days (f_i)	77	90	55	30	5	3

Which probability model might be suitable for describing the number of malfunctions per day?

How would you estimate the value for the parameter of your model?

How might the suitability of your model be tested?

In each of the above examples, a comparison is required of observed frequencies $\left(O_i = f_i\right)$ and expected frequencies $\left(E_i\right)$ calculated from an assumed or hypothesised probability model.

This comparison is tested for significance again using

$$\chi^2 = \sum \frac{\left(O_i - E_i\right)^2}{E_i}$$

However, some complications are encountered when values for the parameters of the assumed distribution are unknown and/or when some expected frequencies are small. To demonstrate how these complications arise and show how they are overcome is best illustrated by considering each of the above two examples in turn.

Example (binomial)

Four identical six-sided dice, each with faces marked 1 to 6, are rolled 200 times. At each rolling, a record is made of the number of dice whose score on the uppermost face is even. The results are as follows.

Number of even scores (x_i)	0	1	2	3	4
Frequency (f_i)	10	41	70	57	22

Explain why a binomial model might describe the distribution of X, and test its goodness of fit.

Solution

A binomial distribution might be appropriate on the basis of:

(a) a fixed number (4) of repeated, independent (dice identical and scores independent) trials,

(b) each trial can result in success (even) or failure (odd),

(c) constant probability (dice identical) of success at each trial.

Assuming the dice are unbiased, then

$$P(\text{even score on a die}) = \frac{3}{6} = 0.5$$

Hence $\quad X \sim B(4, 0.5)$

Using the probability distribution of this binomial distribution (see Section 3.3) gives the following probability and hence expected frequency (200 $\times$ probability) for each value of X.

x_i	$O_i = f_i$	$P(X = x_i)$	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	10	0.0625	12.5	–2.5	6.25	0.500
1	41	0.2500	50.0	–9.0	81.00	1.620
2	70	0.3750	75.0	–5.0	25.00	0.333
3	57	0.2500	50.0	7.0	49.00	0.980
4	22	0.0625	12.5	9.5	90.25	7.220
		1.0000				10.653

H_0: number of even scores is ~ $B(4, 0.5)$

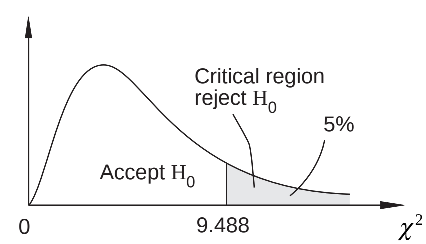
H_1: number of even scores is not ~ $B(4, 0.5)$

Significance level, $\alpha = 0.05$

Degrees of freedom, $v = 5 - 1 = 4$

(5 classes: 0, 1, 2, 3, 4; 1 constraint: $\sum E_i = \sum O_i$)

Critical region is $\quad \chi^2 > 9.488$



Test statistic is $\quad \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 10.653$

This value does lie in the critical region. There is evidence, at the 5% significance level, to suggest that the number of even scores is not distributed as $B(4, 0.5)$.

Suppose that it is now revealed that the dice are equally biased in some unknown way.

Can you now think of a way of estimating the probability of an even score on one of the dice?

The mean number of even scores per single rolling of the four dice is given by

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{(10 \times 0 + 41 \times 1 + 70 \times 2 + 57 \times 3 + 22 \times 4)}{200}$$

$$= \frac{440}{200} = 2.2$$

Thus

$$P(\text{even score on a single die}) = \hat{p} = \frac{2.2}{4} = 0.55$$

Again using the probability distribution of this binomial distribution $(n = 4,\ p = 0.55)$ gives the following probability and hence the expected frequency ($200 \times$ probability) for each value of X.

x_i	$O_i = f_i$	$P(X = x_i)$	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	10	0.0410	8.2	1.8	3.24	0.395
1	41	0.2005	40.1	0.9	0.81	0.020
2	70	0.3675	73.5	–3.5	12.25	0.167
3	57	0.2995	59.9	–2.9	8.41	0.140
4	22	0.0915	18.3	3.7	13.69	0.748
		1.0000				1.470

What is the value of $\sum E_i x_i$?

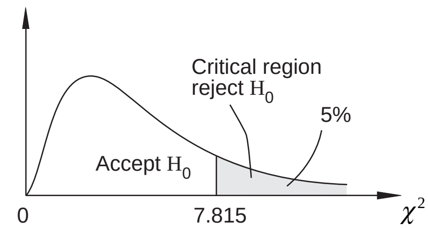
H_0: number of even scores is ~ $B(4, p)$

H_1: number of even scores is not ~ $B(4, p)$

Significance level, $\alpha = 0.05$

Degrees of freedom, $v = 5 - 2 = 3$

(5 classes: 0, 1, 2, 3, 4; 2 constraints: $\sum E_i = \sum O_i$ and $\sum E_i x_i = \sum O_i x_i$ from estimation of p)



Critical region is $\chi^2 > 7.815$

Test statistic is $$\chi^2 = \sum \frac{\left(O_i - E_i\right)^2}{E_i} = 1.470$$

This value does not lie in the critical region. There is no evidence, at the 5% significance level, to suggest that the number of even scores cannot be modelled by a binomial distribution with $n = 4$ and $p = 0.55$.

Example (Poisson)

The number of computer malfunctions per day is recorded for 260 days with the following results.

Number of malfunctions (x_i)	0	1	2	3	4	5
Number of days (f_i)	77	90	55	30	5	3

Test the goodness of fit of an appropriate probability model.

Solution

An appropriate probability model here could be a Poisson distribution on the basis of the following assumptions.

(a) Malfunctions occur independently.

(b) Simultaneous malfunctions are impossible.

(c) Malfunctions occur randomly in time.

(d) Malfunctions occur uniformly (mean number per time period proportional to the period length).

Can you devise an approximate, but quick, method for checking whether a Poisson distribution might be a suitable model?

A Poisson distribution has one parameter, λ, which is the mean (and also the variance). Thus the sample mean may be used to estimate λ.

Mean $\quad \bar{x} = \dfrac{\sum f_i x_i}{\sum f_i} = \dfrac{325}{260} = 1.25$

[Note that $\sum f_i x_i^2 = 735$, so variance, $s^2 = \dfrac{735}{260} - 1.25^2 = 1.26$]

Hence with $\lambda = 1.25$,

$$P(X = x) = \frac{e^{-1.25} 1.25^x}{x!}, \quad x = 0, 1, 2, 3, \ldots\ldots$$

which gives the following probabilities and expected frequencies (260 $\times$probability) for each value of X.

x_i	$P(X = x_i)$	E_i
0	0.2865	74.5
1	0.3581	93.1
2	0.2238	58.2
3	0.0933	24.2
4	0.0291	7.6
5	0.0073	1.9
≥ 6*	0.0019	0.5
	1.0000	260.0

*Since a Poisson distribution is valid for all positive values of X, this additional class is necessary, and $P(X \geq 6) = 1 - P(X \leq 5)$.

What is the value of $\sum E_i x_i$ if ≥ 6 is read as 6?

The test statistic

$$\sum \frac{(O_i - E_i)^2}{E_i}$$

is approximated by χ^2 providing none of the expected frequencies are less than 5. When expected frequencies fall below 5, then groups or classes must be combined.

In the current example, this rule necessitates combining 4, 5 and ≥ 6 malfunctions into ≥ 4 malfunctions, so giving the following table.

x_i	$O_i = f_i$	$P(X = x_i)$	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	77	0.2865	74.5	2.5	6.25	0.084
1	90	0.3581	93.1	–3.1	9.61	0.103
2	55	0.2238	58.2	–3.2	10.24	0.176
3	30	0.0933	24.2	5.8	33.64	1.390
≥ 4	8	0.0383	10.0	–2.0	4.00	0.400
		1.0000				2.153

H_0 : number of daily malfunctions is ~ Poisson

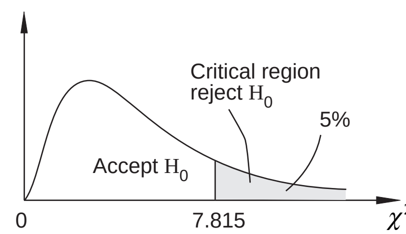
H_1 : number of daily malfunctions is not ~ Poisson

Significance level, $\alpha = 0.05$

Degrees of freedom, $v = 5 - 2 = 3$

(5 classes: 0, 1, 2, 3, ≥ 4; 2 constraints: $\sum E_i = \sum O_i$ and $\sum E_i x_i = \sum O_i x_i$ from estimation of λ)

Critical region is $\chi^2 > 7.815$



Test statistic is $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.153$

This value does not lie in the critical region. There is no evidence, at the 5% significance level, to suggest that the number of computer malfunctions per day does not have a Poisson distribution.

Activity 3 Shuffled cards

Obtain the use of a standard pack of 52 playing cards.

Shuffle the cards thoroughly, deal the top two cards, and record how many (0, 1 or 2) are diamonds.

Repeat this process at least 80 times.

Test the hypothesis that the number of diamonds per deal can be modelled by a distribution which is $B(2, 0.25)$.

Activity 4 Binomial quiz

[You may already have some data for this activity from Activity 5 in Chapter 5 of *Statistics*.]

Ask your fellow students, and anyone else who will participate, whether the following statements are 'true' or 'false'.

1. The *Portrait of a Lady* was written by Henry James.
2. *Psalms* is the 20th book of the *Old Testament*.
3. The equatorial diameter of Mercury is about 3032 miles.
4. Mankoya is a place in Zambia.
5. 'The Potato Eaters' is a painting by Cezanne.
6. The Battle of Sowton was fought in 1461.

Make a frequency table to show the number of correct answers out of six for those asked. You will need to ask about 200 people.

Calculate expected frequencies on the assumption that people are simply guessing the answers. Compare observed and expected frequencies using a χ^2 goodness of fit test.

Estimate the probability, $\hat{p}$, of a correct answer. Hence test the appropriateness of the distribution $B(6, \hat{p})$.

Activity 5 Simulated spelling errors

The random variable X denotes the number of spelling errors made by a typist per 1000 words typed.

Use a three-digit random number, y, to generate an 'observed' value, x, of X using the following rules.

$$000 \le y \le 135 \Rightarrow x = 0$$

$$136 \le y \le 406 \Rightarrow x = 1$$

$$407 \le y \le 677 \Rightarrow x = 2$$

$$678 \le y \le 857 \Rightarrow x = 3$$

$$858 \le y \le 947 \Rightarrow x = 4$$

$$948 \le y \le 999 \Rightarrow x \ge 5$$

Using the above rules, obtain at least 100 'observed' values of X.

Test the hypothesis that $X \sim Po(2)$ using the χ^2 goodness of fit test.

Activity 6 Customer arrivals

[You may already have some data for this activity from Activity 3 in Chapter 6 of *Statistics*.]

Record the number of arrivals of customers at a post office, bank or supermarket in 2-minute intervals until you have at least 120 results. The 2-minute interval may well be shortened in the case of a large and busy site.

At the same time ask a friend to record the times, in seconds, between successive arrivals of customers. (This information is required for Activity 8.) Alternatively, record, to the nearest second, the actual arrival time of each customer. Numbers of arrivals per interval and inter-arrival times can then be listed later.

Construct a frequency distribution of the numbers of arrivals per interval.

Use a χ^2 goodness of fit test to investigate whether a Poisson distribution provides a suitable model.

Exercise 5B

1. A farmer kept a record of the number of heifer calves born to each of his cows during the first five years of breeding of each cow. The results are summarised below.

Number of heifers	0	1	2	3	4	5
Number of cows	4	19	41	52	26	8

Test, at the 5% level of significance, whether or not the binomial distribution with parameters $n = 5$ and $p = 0.5$ is an adequate model for these data.

Explain briefly (without doing any further calculations) what changes you would make in your analysis if you were testing whether or not the binomial distribution with $n = 5$ and unspecified p fitted the data. (AEB)

2. The number of misprints on 200 randomly selected pages from the 1981 editions of the *Daily Planet*, a quality newspaper, were recorded. The table below summarises these results.

Number of misprints per page (x)	0	1	2	3	4	5	6	7	8	>8
Frequency (f)	5	12	31	40	38	29	22	14	5	4

Use a χ^2 distribution to test the claim that a suitable model for these data is Poisson with a mean of 4. (AEB)

3. Smallwoods Ltd run a weekly football competition. One part of this involves a fixed-odds contest where the entrant has to forecast correctly the result of each of five given matches. In the event of a fully correct forecast the entrant is paid out at odds of 100 to 1. During the last two years Miss Fortune has entered this fixed-odds contest 80 times. The table below summarises her results.

Number of matches correctly forecast per entry (x)	0	1	2	3	4	5
Number of entries with x correct forecasts (f)	8	19	25	22	5	1

(a) Find the frequencies of the number of matches correctly forecast per entry given by a binomial distribution having the same mean and total as the observed distribution.

(b) Use the χ^2 distribution and a 10% level of significance to test the adequacy of the binomial distribution as a model for these data.

(c) On the evidence before you, and assuming that the point of entering is to win money, would you advise Miss Fortune to continue with this competition and why? (AEB)

4. The table below gives the distribution of the number of hits by flying bombs in 450 equally sized areas in South London during World War II.

Number of hits (x)	0	1	2	3	4	5
Frequency (f)	180	173	69	20	6	2

(a) Find the expected frequencies of hits given by a Poisson distribution having the same mean and total as the observed distribution.

(b) Use the χ^2 distribution and a 10% level of significance to test the adequacy of the Poisson distribution as a model for these data. (AEB)

5.3 Continuous probability models

For a continuous random variable, probabilities for precise values do not exist so a comparison of observed and expected frequencies of individual values is not possible.

However, it is possible to calculate, by integration or using tables, the probability that the value of a continuous random variable falls within some specified interval.

Hence for continuous probability models, χ^2 goodness of fit tests are based upon a comparison of observed and expected frequencies in non-overlapping intervals which together constitute the complete range for the random variable.

Example

The following table summarises the waiting times, in minutes, of a random sample of 200 people at a taxi rank.

Waiting time (x)	0 –	0.5 –	1.0 –	1.5 – 2.5
Number of people (f)	77	60	35	28

Test the claim that the waiting time, X, can be modelled by the probability density function

$$f(x) = \begin{cases} 0.8 - 0.32x & 0 \le x < 2.5 \\ 0 & \text{otherwise} \end{cases}$$

Solution

$$P(X < a) = \int_0^a (0.8 - 0.32x)\,dx = 0.16a(5 - a),$$

so that

$$P(X < 2.5) = 1.00 \qquad P(0.0 \le X < 0.5) = 0.36$$

$$P(X < 1.5) = 0.84 \qquad P(0.5 \le X < 1.0) = 0.28$$

$$P(X < 1.0) = 0.64 \qquad P(1.0 \le X < 1.5) = 0.20$$

$$P(X < 0.5) = 0.36 \qquad P(1.5 \le X < 2.5) = 0.16$$

Using $E_i = (200 \times \text{probability})$ gives the following table of calculations.

x	$O_i = f_i$	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0.0 –	77	72	5	25	0.347
0.5 –	60	56	4	16	0.286
1.0 –	35	40	–5	25	0.625
1.5 – 2.5	28	32	–4	16	0.500
					1.758

H_0: suggested model is appropriate

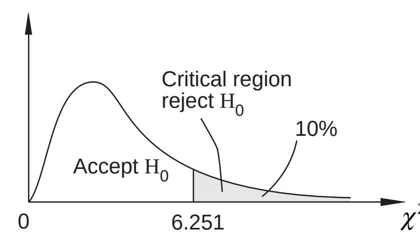
H_1: suggested model is not appropriate

Significance level, $\alpha = 0.10$ (say)

Degrees of freedom, $v = 4 - 1 = 3$

(4 classes; 1 constraint: $\sum E_i = \sum O_i$)

Critical region is $\chi^2 > 6.251$



Test statistic is $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 1.758$

This value does not lie in the critical region. There is no evidence, at the 10% significance level, to suggest that waiting times cannot be modelled by the suggested probability density function.

Activity 7 Response times

Using a stopwatch or a watch with a stopwatch facility in hundredths of a second, set the watch going and try to stop at exactly 5 seconds. Record the exact time on the stopwatch. At this stage it is easier to work in pairs. Repeat this at least 100 times and then construct a grouped frequency distribution using the following classes:

4.5 – , 4.7 – , 4.8 – , 4.9 – , 5.0 – , 5.1 – , 5.2 – , 5.3 – 5.5.

{Discard the few, if any, times outside the range 4.5 to 5.5.}

Hence test the goodness of fit of your data to the distribution defined by

$$f(x)=\begin{cases}4(x-4.5) & 4.5<x\le 5.0\\ 4(5.5-x) & 5.0<x\le 5.5\\ 0 & \text{otherwise}\end{cases}$$

Draw a relative frequency histogram of your grouped frequency distribution and superimpose the graph of the above distribution.

Comment on the similarities and/or differences.

Example (exponential)

The table below shows the time intervals, in seconds, between successive white cars in free flowing traffic on an open road. Can these times be modelled by an exponential distribution?

Time	0 –	20 –	40 –	60 –	90 –	120 –180
Frequency	41	19	16	13	9	2

Solution

The p.d.f. of an exponential distribution is given by

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & \text{otherwise}\end{cases}$$

where λ^{-1} is the mean.

Hence the sample mean can be used as an estimate of λ^{-1}.

Using class mid-points (x) of 10, 30, 50, 75, 105 and 150 gives

$$\bar{x}=\hat{\lambda}^{-1}=\frac{\sum fx}{\sum f}=\frac{4000}{100}=40$$

Hence $$P(X < a) = \int_0^a \frac{e^{-\frac{x}{40}}}{40}\,dx = 1 - e^{-\frac{a}{40}}$$

The following probabilities and expected frequencies can now be calculated.

	Class probability		E_i
$P(X < \infty) = 1.000$	$P(0 \le X < 20)$	$= 0.393$	39.3
$P(X < 180) = 0.989$	$P(20 \le X < 40)$	$= 0.239$	23.9
$P(X < 120) = 0.950$	$P(40 \le X < 60)$	$= 0.145$	14.5
$P(X < 90) = 0.895$	$P(60 \le X < 90)$	$= 0.118$	11.8
$P(X < 60) = 0.777$	$P(90 \le X < 120)$	$= 0.055$	5.5
$P(X < 40) = 0.632$	$P(120 \le X < 180)$	$= 0.039$	3.9*
$P(X < 20) = 0.393$	$P(180 \le X < \infty)$	$= 0.011$	1.1*
		1.000	100.0

* Combine classes so that all $E_i \ge 5$

What will be the value now for the degrees of freedom?

The calculation of the value of χ^2 is completed as follows.

x	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0 –	41	39.3	1.7	2.89	0.074
20 –	19	23.9	–4.9	24.01	1.005
40 –	16	14.5	1.5	2.25	0.155
60 –	13	11.8	1.2	1.44	0.122
90 –	9	5.5	3.5	12.25	2.227
120 –	2	5.0	–3.0	9.00	1.800
					5.383

H_0: exponential distribution is appropriate

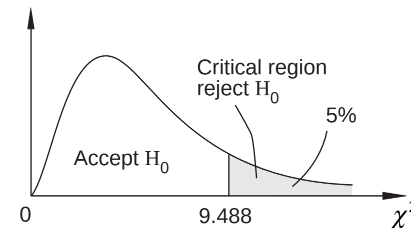
H_1: exponential distribution is not appropriate

Significance level, $\alpha = 0.05$

Degrees of freedom, $v = 6 - 2 = 4$

(6 classes; 2 constraints: $\sum E_i = \sum O_i$ and $\sum E_i x_i = \sum O_i x_i$ from estimation of λ)

Critical region is $\chi^2 > 9.488$



Test statistic is $\chi^2 = \sum \dfrac{\left(O_i - E_i\right)^2}{E_i} = 5.383$

This value does not lie in the critical region. There is no evidence, at the 5% significance level, to suggest that an exponential distribution is not appropriate.

What changes would need to be made to this solution if the question had suggested an exponential distribution with $\lambda = 0.025$?

Example (normal)

An analysis of the fat content, X%, of a random sample of 175 hamburgers of a particular grade resulted in the following summarised information.

Fat content	Number of hamburgers (f)
$26 \le x < 28$	7
$28 \le x < 30$	22
$30 \le x < 32$	36
$32 \le x < 34$	45
$34 \le x < 36$	33
$36 \le x < 38$	28
$38 \le x < 40$	4

Can it be assumed that the fat content of this grade of hamburger is normally distributed?

Solution

What parameters will need to be estimated?

Here it is necessary to estimate the parameters μ and σ^2 of the normal distribution by $\bar{x}$ and $\hat{\sigma}^2$ respectively.

What will now be the number of constraints?

Using class mid-points of 27, 29, ..., 39 results in

$$\sum fx = 5775 \quad \text{and} \quad \sum fx^2 = 192047$$

so that $$\bar{x} = \frac{5775}{175} = 33$$

and $$\hat{\sigma} = \sqrt{\frac{1}{174}\left[192047 - \frac{5775^2}{175}\right]} = 2.91$$

Using the standardisation $z = \dfrac{x-\mu}{\sigma} = \dfrac{x-33}{2.91}$ gives the following probabilities and hence expected frequencies.

	Class probability	E_i
$P(X<\infty) = P(Z<\infty) = 1.000$	$P(-\infty < X < 26) = 0.008$	1.4*
$P(X<40) = P(Z<2.405) = 0.992$	$P(26 \le X < 28) = 0.035$	6.1*
$P(X<38) = P(Z<1.718) = 0.957$	$P(28 \le X < 30) = 0.108$	18.9
$P(X<36) = P(Z<1.031) = 0.849$	$P(30 \le X < 32) = 0.214$	37.5
$P(X<34) = P(Z<0.344) = 0.635$	$P(32 \le X < 34) = 0.270$	47.2
$P(X<32) = P(Z<-0.344) = 0.365$	$P(34 \le X < 36) = 0.214$	37.5
$P(X<30) = P(Z<-1.031) = 0.151$	$P(36 \le X < 38) = 0.108$	18.9
$P(X<28) = P(Z<-1.718) = 0.043$	$P(38 \le X < 40) = 0.035$	6.1*
$P(X<26) = P(Z<-2.405) = 0.008$	$P(40 \le X < \infty) = 0.008$	1.4*
	1.000	175.0

* Combine classes so that all $E_i \ge 5$

The calculation of the value of χ^2 is now completed as follows.

Class	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\dfrac{(O_i - E_i)^2}{E_i}$
$-\infty < x < 28$	7	7.5	–0.5	0.25	0.033
$28 \le x < 30$	22	18.9	3.1	9.61	0.508
$30 \le x < 32$	36	37.5	–1.5	2.25	0.060
$32 \le x < 34$	45	47.2	–2.2	4.84	0.103
$34 \le x < 36$	33	37.5	–4.5	20.25	0.540
$36 \le x < 38$	28	18.9	9.1	82.81	4.381
$38 \le x < \infty$	4	7.5	–3.5	12.25	1.633
					7.258

H_0: normal distribution is appropriate

H_1: normal distribution is not appropriate

Significance level, $\alpha = 0.10$ (say)

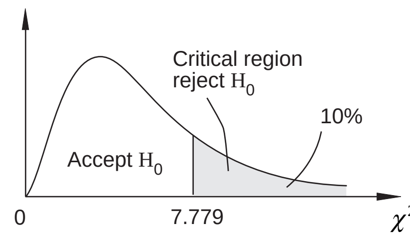
Degrees of freedom, $v = 7 - 3 = 4$

(7 classes; 3 constraints:

$\sum E_i = \sum O_i$,

$\sum E_i x_i = \sum O_i x_i$ from estimation of μ ,

and $\sum E_i x_i^2 = \sum O_i x_i^2$ from estimation of σ^2)



Critical region is $\chi^2 > 7.779$

Test statistic is $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 7.258$

This value does not lie in the critical region. There is no evidence, at the 10% significance level, to suggest that a normal distribution is not appropriate.

What changes would need to be made to this solution if the question had suggested a normal distribution with

(a) $\mu = 33$ **(b)** $\sigma = 3$ **(c)** $\mu = 33$ **and** $\sigma = 3$**?**

Activity 8 Customer arrivals revisited

Construct a grouped frequency distribution of the inter-arrival times recorded in Activity 6.

Draw a histogram of your results and observe its shape.

Examine whether or not inter-arrival times may be modelled by an exponential distribution.

Activity 9 Textbook weights

Using a suitable random selection procedure, obtain and weigh, in grams, a sample of at least 100 textbooks from your school or college library. A set of electronic kitchen scales should give sufficient accuracy.

Construct a grouped frequency distribution of your results and draw the corresponding histogram.

Investigate the claim that textbook weights are normally distributed.

Exercise 5C

1. A sample of 300 electronic circuit components is selected at random from a production process. The lifetime, in hours, of each component is measured by testing it to destruction with the following summarised results.

Lifetime	0-	50-	100-	150-	200-	300-	400-	500-
Frequency	63	47	55	34	29	27	24	21

Use a goodness of fit test to test the hypothesis that the lifetimes of components from this production process follow an exponential distribution with mean 200 hours.

2. The duration, in hours, of the effect of the standard dose of a certain drug on a healthy adult female is thought to be exponentially distributed. The table below shows the results for a random sample of 200 healthy females all given the standard dose.

Duration	0-	3-	6-	9-	12-	18-	24-36
Females	40	31	31	22	23	22	31

Test the hypothesis that the duration of the effect is exponentially distributed.

3. The table below summarises the times taken, in seconds, by 250 nine-year-old children to tie their shoe laces.

Time	<10	10-	20-	30-	40-	50-	60-
Frequency	9	16	51	72	79	11	12

Apply a χ^2 goodness of fit test to investigate whether or not the times taken by nine-year-old children to tie their shoe laces are normally distributed with a mean of 35 seconds and a standard deviation of 10 seconds.

4. The shape of the human head was the subject of an international project financed by the World Council for Health and Welfare. Observations were taken in many countries and the nose lengths, to the nearest millimetre, of 150 Italians are summarised below.

Nose lengths (mm) x	Frequency f
$-\infty < x \le 44$	4
$45 \le x \le 47$	12
$48 \le x \le 50$	63
$51 \le x \le 53$	59
$54 \le x \le 56$	10
$57 \le x < \infty$	2

Estimate the mean and the standard deviation of the population from which these observations were taken. (For these calculations you should assume that the lower and upper classes have the same width as the other classes.)

Use the χ^2 distribution and a 1% level of significance to test the adequacy of the normal distribution as a model for these data.

(AEB)

5.4 Miscellaneous Exercises

1. At a vinegar bottling plant, samples of five bottles are taken at regular intervals and their contents measured. During a particular week 170 samples were taken and the number of bottles in each sample containing less than 1160 ml recorded. The results are given in the following table.

Number of bottles containing less than 1160 ml	0	1	2	3	4	5
Number of samples	41	62	49	12	5	1

Estimate p, the proportion of bottles produced containing less than 1160 ml.

Find the expected frequencies of a binomial distribution with the same sample size, value of p and total frequency as the observed data.

Applying a goodness of fit test, investigate whether the data support the view that the numbers of bottles in samples of size 5, containing less than 1160 ml, have a binomial distribution. (AEB)

2. A factory operates four production lines. Maintenance records show that the daily number of stoppages due to mechanical failure were as follows (it is possible for a production line to break down more than once on the same day).

Number of stoppages (x)	**Number of days** (f)
0	728
1	447
2	138
3	48
4	26
5	13
6 or more	0

You may assume that $\sum f = 1400 \qquad \sum fx = 1036$

(a) Use a χ^2 distribution and a 1% significance level to determine whether the Poisson distribution is an adequate model for the data.

(b) The maintenance engineer claims that breakdowns occur at random and that the mean rate has remained constant throughout the period. State, giving a reason, whether your answer to (a) is consistent with this claim.

(c) Of the 1036 breakdowns which occurred, 230 were on production line A, 303 on B, 270 on C and 233 on D. Test at the 5% significance level whether these data are consistent with breakdowns occurring at an equal rate on each production line. (AEB)

3. In a European country, registration for military service is compulsory for all eighteen-year-old males. All males must report to a barracks where, after an inspection, some, including all those less than 1.6 m tall, are excused service. The heights of a sample of 125 eighteen-year-olds measured at the barracks were as follows.

Height	1.2–	1.4–	1.6–	1.8–	2.0–2.2
Frequency	6	34	31	42	12

(a) Use a χ^2 test and a 5% significance level to confirm that the normal distribution is not an adequate model for these data.

(b) Show that, if the second and third classes (1.4– and 1.6–) are combined, the normal distribution does appear to fit the data. Comment on this apparent contradiction in the light of the information at the beginning of the question. (AEB)

4. (a) The number of books borrowed from a library during a certain week were 518 on Monday, 431 on Tuesday, 485 on Wednesday, 443 on Thursday and 523 on Friday.

Is there any evidence that the number of books borrowed varies between the five days of the week? Use a 1% level of significance.

Interpret fully your conclusions.

(b) The following 50 observations were believed to be a random sample from the discrete probability distribution defined by

$$P(X=r) = (r-1)p^2(1-p)^{r-2} \quad r = 2, 3, \ldots$$

r	2	3	4	5	≥ 6
Frequency	18	17	12	3	0

An appropriate estimate of p is $\frac{2}{\bar{x}}$, where $\bar{x}$ is the sample mean. Use a χ^2 test, at the 5% significance level, to investigate whether these data are consistent with the postulated distribution. (AEB)

5. (a) The table below summarises the values of 600 pseudo random numbers as generated by a particular model of calculator.

Value	0.0–	0.2–	0.4–	0.6–	0.8–1.0
Frequency	105	141	112	107	135

Can it be assumed that pseudo random numbers, as generated by this model of calculator, are distributed uniformly over the interval zero to unity?

(b) The table below shows the results of the measurement of the lifetime, in thousands of hours, of each of a random sample of 200 Type A transistors.

Lifetime	0–	5–	10–	15–
Frequency	47	34	28	25

Lifetime	20–	30–	50–
Frequency	22	33	11

Perform a χ^2 goodness of fit test of the hypothesis that the lifetimes of Type A transistors follow an exponential distribution with mean 20000 hours.

Outline, without calculation, the modifications necessary to your analysis, if the hypothesis did not specify the mean value.

6. A mill weaves cloth in standard lengths. When a length of cloth contains a serious blemish, the damaged section is cut out and the two remaining parts stitched together. This is known as a string. An analysis of the number of strings in 220 lengths of a particular type revealed the following data.

Number of strings	0	1	2	3	4	5	6	7
Frequency	14	29	57	48	31	41	0	0

(a) Test whether a Poisson distribution is an adequate model for these data, using a 5% significance level.

(b) On seeing the analysis, the manager pointed out that lengths of cloth containing more than 5 strings were unsaleable. If necessary, larger sections of cloth would be removed so that no length contained more than 5 strings. Without this restriction, she estimated that there would be an average of 3 strings per length.

If a Poisson distribution with mean 3 is fitted to the data the expected numbers are as follows.

Number of strings	0	1	2	3	4	≥ 5
Expected number	10.96	32.85	49.29	49.29	36.98	40.63

Test whether a Poisson distribution with mean 3 is an adequate model for the data provided all observations of 5 or more are classified together (as is the case in this part). Use a 5% significance level.

(c) In the light of your calculations in (a) and (b), discuss whether it is likely that serious blemishes occur at random at a constant average rate through the cloth. (AEB)

7. A company sells clothes by mail order catalogue. The size of clothes is defined by the hip size; thus the height of customers of a particular size may vary considerably.

Data sent in by female customers of size 18 showed the following distribution of heights, in centimetres.

Class	**Class mid-mark** (x)	**Frequency** (f)
130–	135	8
140–	145	129
150–	155	61
160–	165	34
170–	175	22
180–	185	11

(a) Given that

$\sum f = 265 \quad \sum fx = 40735 \quad \sum fx^2 = 6299425$

estimate the mean and standard deviation of heights.

(b) Test, at the 1% significance level, whether the normal distribution provides an adequate model for the heights.

(c) The company decides, for economic reasons, that it is not possible to produce a range of garments of a particular size suitable for customers of different heights. A single height must be chosen and it is proposed that this should be the mean height. Comment on this suggestion as it applies to customers of size 18 and make an alternative proposal. (AEB)

8. A biased coin is tossed 5 times and the number of heads obtained is recorded. This is repeated 200 times. The table below summarises the results.

Number of heads (x)	0	1	2	3	4	5
Frequency (f)	5	39	70	52	25	9

Find the frequencies of the number of heads given by a binomial distribution having the same mean and total as the observed distribution.

Instead of calculating the usual goodness of fit statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$, calculate the values of the following:

(a) $\chi_1^2 = \sum \frac{(O-E)}{E}$

(b) $\chi_2^2 = \sum \frac{|O-E|}{E}$

How useful do you think these new statistics may be in measuring the goodness of fit in the above situation? (AEB)

9. The owner of a small country inn observes that during the holiday season the demand for rooms is as follows.

Rooms required	0	1	2	3	4	5	6	7	8
Number of nights	2	9	16	26	33	25	20	11	5

Calculate the mean demand for rooms per night. Use a χ^2 test with a 5% significance level to determine whether the Poisson distribution is an adequate model for the data.

The inn has only four rooms available to let. Assuming demand follows a Poisson distribution with mean 4.17, calculate the mean and variance of the number of rooms occupied per night. (AEB)

10. (a) As part of a statistics project, students observed five private cars passing a college and counted the number which were carrying the driver only, with no passengers. This was repeated 80 times.

The results of a particular student were as follows.

Number of cars with driver only	0	1	2	3	4	5
Number of times observed	0	3	12	27	26	12

Use the χ^2 distribution and a 5% significance level to test whether the binomial distribution provides an adequate model for the data.

(b) In a further part of the project the students counted the number of cars passing the college in 130 intervals each of length 5 seconds. The following table shows the results obtained by the same student together with the expected numbers if a Poisson distribution, with the same mean as the observed data, is fitted.

Number of cars passing a point in a 5 second interval	**Number of intervals**	
	Observed	**Expected**
0	28	25.85
1	40	41.75
2	32	33.72
3	19	18.16
4	7	7.33
5	3	2.37
6	1	0.64
7 or more	0	0.18

Use a χ^2 distribution and a 5% significance level to test whether the Poisson distribution provides an adequate model for the data.

(c) The teacher suspected that this student had not observed the data but invented them. Explain why the teacher was suspicious and comment on the strength of the evidence supporting her suspicions.

(AEB)