Chapter 4

Goodness–of–fit tests

4.1 Introduction

In this section, we address the question of whether our data follow any *pattern*, or fit a specified or assumed *probability distribution*. More precisely, we ask: Is there any evidence against the (null) hypothesis that the data are a random sample from a particular distribution, such as the Poisson distribution? The method first forms a frequency table of the data and then compares the observed frequencies with those that would be expected if the hypothesis was correct. If there is a large difference between observed and expected frequencies then this sheds doubt on the hypothesised distribution.

A simple example: traffic accidents

Consider the following data which is the number of traffic accidents involving children within two kilometres of schools, listed by days of the week.

Day	No. of accidents (Observed)
Monday	23
Tuesday	18
Wednesday	17
Thursday	19
Friday	23

We could use a hypothesis test from chapter 2 to find out, for example, if the mean number of accidents per day is equal to some specified value. However, in goodness–of–fit tests, we are more interested in the *distribution* of the data rather than just a single parameter. Looking at the data, there appears to be more accidents at the start and end of the week than during mid–week days; however, the above dataset is only a sample of all accidents that have taken place, and we need to decide whether or not this apparent effect is simply a result of sampling variation, or whether a pattern *really does exist* in the data.

Let's consider a hypothesis which states that no pattern exists in the data (i.e. there are no "peaks" in the data, and everything occurs *uniformly*). If this were true, we'd ex-

pect there to be (roughly) the same number of accidents on each day of the week. Thus, we would obtain the *expected frequencies*:

Day	No. of accidents (Expected)
Monday	20
Tuesday	20
Wednesday	20
Thursday	20
Friday	20

Although there *did* appear to be more accidents at the start of, and towards the end of, the week, comparing the expected frequencies in the above table with the real–life data (the *observed frequencies*), we see that they are fairly close. But how close do these observed and expected frequencies have to be before we conclude that traffic accidents occur uniformly throughout the week? To avoid subjectivity, we should compare these two sets of frequencies through a formal hypothesis test, known as a *chi–squared ($\boldsymbol{\chi^2}$) goodness–of–fit test*. The hypothesis tests we have looked at so far (tests for one mean and tests for two means) have compared a calculated test statistic to the standard normal distribution or the t–distribution; goodness–of–fit tests use the chi–squared (χ^2) distribution. This distribution, like the t–distribution, has one parameter called the degrees of freedom (ν); its critical values can be found in table 4.1.

4.2 Goodness–of–fit tests: basic framework

The procedure for carrying out a goodness–of–fit test is as follows:

1. **State the null hypothesis ($\boldsymbol{H_0}$)**
   If we think our frequencies occur uniformly, then, in the context of the last example (traffic accidents), this would be

   $$H_0 \ : \ \text{There are the same number of accidents each day of the week.}$$

   If we think our data might follow a particular distribution, then the null hypothesis might be, for example,

   $$H_0 \ : \ \text{The number of accidents follows a Poisson distribution.}$$

   You have already done some work on probability distributions for continuous and discrete data; we will review some of these distributions shortly.

2. **State the alternative hypothesis ($\boldsymbol{H_1}$)**
   This (as is often the case) is just the opposite statement to the null hypothesis; so in the context of goodness–of–fit tests, this might be

   H_1 : There are *not* the same number of accidents each day of the week, or
   H_1 : The number of accidents does *not* follow a Poisson distribution.

3. **Calculate the test statistic**
   The test statistic is

$$X^2 \;=\; \sum \frac{(O-E)^2}{E},$$

   where O and E represent the *observed* and *expected* frequencies respectively. This is most easily calculated by drawing up a table:

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
⋮	⋮	⋮

   and then summing the final column. **For the goodness–of–fit test to work, all expected frequencies must be ≥ 5; to achieve this, adjacent categories can be "pooled"**.

4. **Find the *p*–value of the test**
   As before, a range for our p–value can be found by comparing our test statistic to statistical tables. The test statistic X^2 follows a χ^2 distribution for which critical values are given in table 4.1. As with the t–distribution, the row we use depends on the degrees of freedom, which can be calculated as

$$\nu \;=\; (\text{number of categories after pooling}) - (\text{number of parameters estimated}) - 1.$$

   Once we have the 10%, 5% and 1% critical values, we can find out where our test statistic lies and so obtain a corresponding range for our p–value.

5. **Reach a conclusion**
   This bit's the same as before – use table 2.1 to interpret your p–value. Remember, generally, to reject H_0, we need a p–value of less than 5%. Don't forget to write a sentence in the context of the question, i.e. "the data appear to follow a Poisson distribution".

Back to the traffic accidents example

Let us now test the hypothesis that the number of traffic accidents occurs uniformly throughout the week; i.e. on average, there are an equal number of traffic accidents each day.

Steps 1 and 2 *(hypotheses)*
Our hypotheses are

H_0 : There are the same number of accidents each day of the week versus
H_1 : There are *not* the same number of accidents each day of the week.

Step 3 *(calculating the test statistic)*
This is the hard bit! Remember, the test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

Drawing up a table usually helps! Thus,

Day	Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
Monday	23	20	0.45
Tuesday	18	20	0.2
Wednesday	17	20	0.45
Thursday	19	20	0.05
Friday	23	20	0.45

So,

$$\begin{aligned} X^2 &= 0.45 + 0.2 + 0.45 + 0.05 + 0.45 \\ &= 1.6. \end{aligned}$$

Notice we didn't have to pool any adjacent categories since all expected frequencies were ≥ 5.

Step 4 *(finding the p–value)*
Using table 4.1, with degrees of freedom

$$\begin{aligned} \nu &= (\text{number of categories after pooling}) - (\text{number of parameters estimated}) - 1 \\ &= 5 - 0 - 1 \\ &= 4, \end{aligned}$$

we obtain the following critical values:

Significance level	10%	5%	1%
Critical value	7.78	9.49	13.28

In goodness–of–fit tests, there are no two–tailed tests, and so you don't have to worry about your choice!

Since our test statistic $X^2 = 1.6$ lies to the left of the first critical value, our p–value is bigger than 10%.

Step 5 *(conclusion)*
Using table 2.1 to interpret our p–value, we see that there is no evidence to reject the null hypothesis. Thus, we retain H_0, and conclude that there is insufficient evidence to suggest that accidents do not occur uniformly, i.e. we conclude that the number of accidents is, on the whole, the same for each day of the week (this is because the observed and expected frequencies were so close).

If we are testing the null hypothesis that the data occur uniformly (as in this example), the expected frequencies are easy to calculate. However, things get a little more tricky if we want to test the hypothesis that the data adhere to a more complicated distribution. Before we look at this, the next section will quickly review some probability models introduced in semester 1.

4.3 Probability distributions

A *probability distribution* of a discrete random variable X is the list of all possible values X can take and the probabilities associated with them. For example, if the random variable X is the outcome of a roll of a (fair, six–sided) dice, then the probability distribution for X is

x	1	2	3	4	5	6
$\Pr(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

This is an example of a discrete probability distribution, since the random variable X can only take integer values. There are a number of 'standard' probability distributions which data often adopt; the two *discrete* probability distributions you studied in semester 1 are the *binomial distribution* and the *Poisson distribution*.

4.3.1 The binomial distribution

The binomial distribution is used to model the number of successes in a series of n independent trials. If each trial results in either a "success" or a "failure" and the probability that any particular trial is a "success" is p, then the random variable X = number of successes out of n trials has a *binomial distribution* with probabilities of the form

$$\Pr(X = r) \quad = \quad {}^{n}\mathrm{C}_{r} \times p^{r} \times (1-p)^{n-r},$$

where ${}^{n}\mathrm{C}_{r} = \frac{n!}{r!(n-r)!}$ (or the nCr button on your calculator), and r takes any of the values $0, 1, \ldots, n$ (remember that $r!$ is the factorial function, and $r! = r \times (r-1) \times (r-2) \times \cdots \times 3 \times 2 \times 1$). The mean of a binomial distribution is given by $n \times p$, and the variance is given by $n \times p \times (1-p)$.

The following are examples of an experiment containing a binomial random variable:

(i) 100 students of equal ability sit an exam. The total number passing is recorded;

(ii) A fair dice is thrown 20 times and the number of sixes obtained is recored;

(iii) 50 people are asked in a survey if they would buy a new type of perfume. The number of people saying they would is recorded.

Why is the random variable X = *number of rainy days in a week* unlikely to have a binomial distribution?

In the context of goodness–of–fit tests, we can use the the formula for calculating probabilities from a binomial distribution to calculate expected frequencies based on this distribution; *the expected frequency is just the sample size multiplied by the associated probability.*

4.3.2 The Poisson distribution

This distribution is used to model data which are counts of (random) events in a certain area or time interval, without a fixed upper limit. The probability distribution has one parameter – λ – and its probabilities take the form

$$\Pr(X = r) \quad = \quad \frac{\lambda^r e^{-\lambda}}{r!},$$

where r can take values $0, 1, 2, \ldots$. The mean and variance of a Poisson distribution are both equal to λ.

The following are practical examples in which a Poisson distribution might be appropriate:

(i) Number of cars passing a street corner in an hour period;

(ii) Number of telephone calls made to a call centre in a day.

Again, in the context of goodness–of–fit tests, we can use the formula for calculating probabilities from a Poisson distribution to calculate expected frequencies based on this distribution.

4.4 A More Complex Example

Consider the following data:

Number of claims	Observed frequency (O)
0	144
1	91
2	32
3	11
4	2
5 +	0
	280

The data represents the number of small factories in northern England in which industrial injuries resulted in claims for compensation between April 2003 and March 2004 (Whitaker, L., Biometrika, 2005, **10**, 55). Clearly, here we are not expecting the same number of claims in each class – we would expect the number of factories to decrease as the number of claims increases. We could, however, use a probability distribution to represent the number of accidents we could expect. But which probability distribution might we use?

Well, the data are discrete (since we are looking at counts), and for discrete data we have looked at two probability distributions: the binomial distribution and the Poisson distribution. Recall the set–up for a binomial distribution to be appropriate: we need n trials, each with two outcomes ("success" and "failure"), and the probability of "success", p, should be constant across all trials. In this example, it is difficult to see how we would use a binomial distribution; however, the Poisson distribution might be more appropriate. Remember that for a Poisson distribution to be valid, all we really need are data which are counts over a certain time interval, without a fixed upper limit. Thus, the Poisson distribution might be used here – we have counted the number of factories with 0 claims, 1 claim, 2 claims, etc., and there's no natural fixed upper limit to the number of claims (the last class is just 5+).

Recall that the mean of a Poisson random variable is equal to the rate parameter λ. Thus, we can estimate λ by setting it equal to the sample mean, i.e.

$$\begin{aligned}\lambda &= \frac{0 \times 144 + 1 \times 91 + 2 \times 32 + 3 \times 11 + 4 \times 2}{280} \\ &= \frac{196}{280} \\ &= 0.7.\end{aligned}$$

Once we have this we can proceed as before; the expected probabilities based on the Poisson distribution will be calculated using the Poisson formula on the previous page, and the expected frequencies found by multiplying these probabilities by the sample size.

Steps 1 and 2 *(hypotheses)*
Since we think the Poisson distribution might be an appropriate model for our data, we test

$$\begin{aligned} H_0 &: \text{Claims follow a Poisson distribution} \quad \text{against} \\ H_1 &: \text{Claims do } \textit{not} \text{ follow a Poisson distribution.} \end{aligned}$$

Step 3 *(calculating the test statistic)*
Recall that, for goodness–of–fit tests, the test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

We already have the O's – these are just the observed frequencies. What we need to calculate are the E's (the expected frequencies). Unlike the traffic accidents example, we are now testing the hypothesis that the data follow a particular distribution – the Poisson distribution. Thus, we first of all use the formula for the Poisson distribution to obtain expected *probabilities*, and then multiply these by the the total number of accidents to obtain the expected frequencies. From earlier, we know that Poisson probabilities are found using

$$\Pr(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}.$$

We have estimated λ as 0.7; thus, we just need to substitute this into the formula to calculate our probabilities for different values of r. For example, the expected probability of no deaths is

$$\begin{aligned} \Pr(X = 0) &= \frac{e^{-0.7} \times 0.7^0}{0!} \\ &= 0.4966. \end{aligned}$$

Similarly,

$$\begin{aligned} \Pr(X = 1) &= \frac{e^{-0.7} \times 0.7^1}{1!} \\ &= 0.3476. \end{aligned}$$

We can do this for all our categories and then convert these to frequencies by multiplying by the total number of accidents we have observed (280). This gives

Number of claims	Expected probability	Expected frequency (E)
0	0.4966	139.048
1	0.3476	97.328
2	0.1217	34.076
3	0.0284	7.952
4	0.0050	1.4
5 +	0.0007	0.196
		280

For the χ^2 test to be valid, the expected frequencies must be at least 5; hence, we have to adjust our categories to allow this to happen. We can do this by "pooling" the last three categories, to give:

Number of claims	Observed frequency (O)	Expected frequency (E)
0	144	139.048
1	91	97.328
2	32	34.076
3+	13	9.548

Notice that the observed frequencies must also be "pooled" across the last three categories. The test statistic can then be found by calculating $\frac{(0-E)^2}{E}$ for each category, and then summing over all categories, i.e.

Number of claims	O	E	$\frac{(O-E)^2}{E}$
0	144	139.048	0.176
1	91	97.328	0.411
2	32	34.076	0.126
3+	13	9.548	1.248

Thus,

$$\begin{aligned} X^2 &= \sum \frac{(O-E)^2}{E} \\ &= 0.176 + 0.411 + 0.126 + 1.248 \\ &= 1.961. \end{aligned}$$

Step 4 *(finding the p–value)*
Recall that for goodness–of–fit tests, we use the χ^2 distribution to obtain our p–value. Thus, using table 4.1 with degrees of freedom

$$\begin{aligned} \nu &= (\text{number of categories after pooling}) - (\text{number of parameters estimated}) - 1 \\ &= 4 - 1 - 1 \\ &= 2, \end{aligned}$$

we obtain the following values:

Significance level	10%	5%	1%
Critical value	4.61	5.99	9.21

Our test statistic $X^2 = 1.961$ lies to the left of the first critical value, and so our p–value is bigger than 10%.

Step 5 *(conclusion)*
Using table 2.1 to interpret our p–value, we find that there is no evidence against the null hypothesis, and so we should retain H_0. We can say say that the observed values match those we would expect to see from a Poisson distribution with a value of $\lambda = 0.7$ quite well, and there is insufficient evidence to suggest that our data do not follow the proposed distribution.

4.5 Exercises

1. The following table contains data on the number of customers a shop has on each day of the week. Does the number of customers occur uniformly throughout the week?

Day	Frequency
Monday	95
Tuesday	81
Wednesday	84
Thursday	120
Friday	117
Saturday	110
Sunday	93
Total	700

2. The following table contains data on the number of complaints received per day at a major retail bank's branches.

Number of Complaints	Frequency
0	270
1	140
2	65
3	14
4 +	5
Total	494

Propose an appropriate distribution for these data and test to see if it is consistent with the data.

3*. The Elves Toy Co. makes toy trains. For quality control purposes, toy trains coming off the production line are regularly inspected for defects; defective trains are always thrown away. Once every week, five trains are randomly selected from the production line. The probability that a train has a defect and is thrown away is 0.25. The table below shows the number of weeks in which 0, 1, ..., 5 trains were defective and thus thrown away in 2007.

No. of trains thrown away	No. of weeks
0	10
1	21
2	14
3	6
4	1
5	0

Propose an appropriate probability distribution for the number of defective trains, and test to see if it is consistent with the data observed.

* Prize question

		p				
		50%	10%	5%	1%	0.1%
	1	0.45	2.17	3.84	6.63	10.83
	2	1.39	4.61	5.99	9.21	13.82
	3	2.37	6.25	7.82	11.34	16.27
	4	3.36	7.78	9.49	13.28	18.47
	5	4.34	9.24	11.07	15.09	20.52
	6	5.35	10.64	12.59	16.81	22.46
	7	6.35	12.02	14.07	18.48	24.32
ν	8	7.34	13.36	15.51	20.09	26.13
	9	8.34	14.68	16.92	21.67	27.88
	10	9.34	15.99	18.31	23.21	29.59
	12	11.34	18.55	21.03	26.22	32.91
	15	14.34	22.31	25.00	30.58	37.70
	20	19.34	28.41	31.41	37.57	45.32
	25	24.34	34.38	37.65	44.31	52.62
	30	29.34	40.26	43.77	50.89	59.70

Table 4.1: This table contains values of x for which $\Pr(X^2 > x) = p$, where X^2 has a χ^2–distribution with ν degrees of freedom