4 Hypothesis Testing

Rather than looking at confidence intervals associated with model parameters, we might formulate a question associated with the data in terms of a hypothesis. In particular, we have a so-called null hypothesis which refers to some basic premise which to we will adhere unless evidence from the data causes us to abandon it.

Example 4.1 In a clinical treatment data may be collected to compare two treatments (old v. new).

The *null hypothesis* is likely to be

no difference between treatments

The *alternative hypothesis* might be:-

a) treatments are different (*2-sided*),
b) new treatment is better (*1-sided*),
c) old treatment is better (*1-sided*).

■

In general we are often in a position to specify the form of the p.m.f., $p(x;\theta)$, say, or the p.d.f., $f(x;\theta)$, but there is doubt about the value of the parameter θ. All that is known is that θ is some element of a specified parameter space Θ. We assume that the null hypothesis of interest specifies that θ is an element of some subset Θ_0 of Θ, and so is true if $\theta \in \Theta_0$ but false if $\theta \notin \Theta_0$.

Example 4.2

A coin is tossed and we hypothesise that it is fair. Hence Θ_0 is the set $\{\frac{1}{2}\}$ containing just one element of the parameter space $\Theta = [0,1]$.

■

As a convention we shall denote the complement of Θ_0 in Θ by Θ_1. We call the original hypothesis that $\theta \in \Theta_0$ the *null hypothesis* and denote it by H_0. The hypothesis that $\theta \in \Theta_1$ is referred to as the *alternative hypothesis* and denoted by H_1.

4.1 Data and questions

Data set 2.3 (which we have seen before) *Silver content of Byzantine coins*

A number of coins from the reign of King Manuel I, Comnemus (1143 - 80) were discovered in Cyprus. They arise from four different coinages at intervals throughout his reign. The question of interest is whether there is any significant difference in their silver

content with the passage of time; there is a suspicion that it was deliberately and steadily reduced. The data give the silver content (%Ag) of the coins.

Table 2.3 Silver content of coins

First	Second	Third	Fourth
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

On the face of it the suspicion could be correct in that the fourth coinage would seem to have lower silver content than, say, the first coinage, but there is a need for firm statistical evidence if it is to be confirmed. Suppose the true percentage of silver in coinage i is μ_i. The null hypothesis would be

$$H_0 : \mu_1 = \mu_4$$

versus the alternative hypothesis

$$H_1 : \mu_1 \neq \mu_4.$$

If we believed, right from the start, that no monarch of the period would ever *increase* the silver content, and that the only possibility of its changing would be in the direction of reduction, then the alternative hypothesis would be

$$H_1 : \mu_1 > \mu_4.$$

□

Data set 4.1 *Patients with glaucoma in one eye*

The following data (Ehlers, N., *Acta Opthalmologica*, **48**) give corneal thicknesses in microns for patients with one glaucomatous eye and one normal eye.

Table 4.1 Glaucoma in one eye

Corneal thickness		
Glaucoma	Normal	Difference
488	484	4
478	478	0
480	492	−12
426	444	−18
440	436	4
410	398	12
458	464	−6
460	476	−16

Is there a difference in corneal thickness between the eyes? To answer this we take the differences *Glaucoma* − *Normal* for each patient and test for the mean of those differences being zero.

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0.$$

□

Data set 4.2 *Shoshoni rectangles*

Most individuals, if asked to draw a rectangle, would produce something instinctively "not too square and not too long" - something pleasing to the eye.

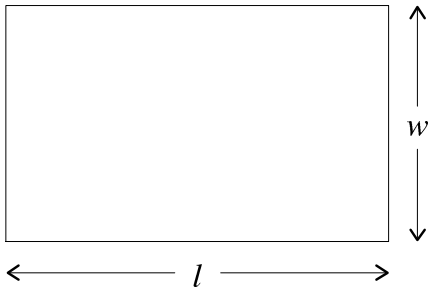

Figure 4.1

The ancient Greeks called a rectangle *golden* if the ratio of its width to its length was

$$\frac{w}{l} = \frac{1}{2}\left(\sqrt{5} - 1\right) = 0.618$$

This ratio is called the *golden ratio*.

Shoshoni Indians used beaded rectangles to decorate their leather goods. The table below gives width-to-length ratios for 20 rectangles, analysed as part of a study in experimental aesthetics.

Table 4.2 Shoshoni bead rectangles

Width-to-length ratios			
0.693	0.670	0.654	0.749
0.606	0.553	0.601	0.609
0.672	0.662	0.606	0.615
0.844	0.570	0.933	0.576
0.668	0.628	0.690	0.611

The data are taken from Lowie's Selected Papers in Anthropology. ed Dubois, University of California Press.

Did the Shoshoni instinctively make their rectangles conform to the golden ratio? In terms of hypothesis testing, we would like to test

$$H_0 : \theta = 0.618 \quad \text{against} \quad H_1 : \theta \neq 0.618.$$

□

Data set 4.3 *Etruscan and Italian skull widths*

The data comprise measurements of maximum head breadth (in cm) which were made in order to ascertain whether Etruscans were native Italians. They were published by Barnicott, N.A. and Brothwell, D.R. (1959), *The Evaluation of Metrical Data in the Comparison of Ancient and Modern Bones in Medical Biology and Etruscan Origins*. Little, Brown and Co.

Table 4.3 Ancient Etruscan and modern Italian skull widths

Ancient Etruscan skulls							Modern Italian skulls					
141	147	126	140	141	150	142	133	124	129	139	144	140
148	148	140	146	149	132	137	138	132	125	132	137	130
132	144	144	142	148	142	134	130	132	136	130	140	137
138	150	142	137	135	142	144	138	125	131	132	136	134
154	149	141	148	148	143	146	134	139	132	128	135	130
142	145	140	154	152	153	147	127	127	127	139	126	148
150	149	145	137	143	149	140	128	133	129	135	139	135
146	158	135	139	144	146	142	138	136	132	133	131	138
155	143	147	143	141	149	140	136	121	116	128	133	135
158	141	146	140	143	138	137	131	131	134	130	138	138
150	144	141	131	147	142	152	126	125	125	130	133	
140	144	136	143	146	149	145	120	130	128	143	137	

The question asked by the archaeologists is whether there is evidence of different race from these two sets of measurements. The null hypothesis is that the two samples of measurements come from the same distribution; the alternative hypothesis is that there is a difference between the two samples.
□

Data set 4.4 *Pielou's data on Armillaria root rot in Douglas fir trees*

The data below were collected by the ecologist E.C. Pielou, who was interested in the pattern of healthy and diseased trees. The subject of her research was *Armillaria* root rot in a plantation of Douglas firs. She recorded the lengths of 109 runs of diseased trees and these are given below.

Table 4.4 Run lengths of diseased trees

Run length	1	2	3	4	5	6
Number of runs	71	28	5	2	2	1

On biological grounds, Pielou proposed a geometric distribution as a probability model. Is this plausible? Can we formally test the goodness of fit of a hypothesised distribution to data?
□

Data set 4.5 *Flying bomb hits on London*

The following data give the number of flying bomb hits recorded in each of 576 small areas of $\frac{1}{4}km^2$ in the south of London during World War II.

Table 4.5 Flying bomb hits on London

Number of hits in an area	0	1	2	3	4	5	≥ 6
Frequency	229	211	93	35	7	1	0

Propaganda broadcasts claimed that the weapon could be aimed accurately. If, however, this was not the case, the hits should be randomly distributed over the area and should therefore be fitted by a Poisson distribution. Is this the case?
□

Data set 4.6 *A famous and historic data set*

These are Pearson's 1909 data on crime and drinking.

Table 4.6 Crime and drinking

Crime	*Drinker*	*Abstainer*
Arson	50	43
Rape	88	62
Violence	155	110
Stealing	379	300
Coining	18	14
Fraud	63	144

Is crime drink related?
□

Data set 4.7 *Snoring and heart disease*

The data in the table below come from a study which investigated whether snoring was related to heart disease (Norton, P.G. and Dunn, E.V. (1985), *British Medical Journal*, **291**, pages 630-632).

Table 4.7 Snoring frequency and heart disease

Heart disease	Non-snorers	Occasional snorers	Snore nearly every night	Snore every night	Total
Yes	24	35	21	30	110
No	1355	603	192	224	2374
Total	1379	638	213	254	2484

Is there an association between snoring frequency and heart disease?
□

4.2 Basic ideas

The first, and main idea, is that we need to use statistics which contain all of the relevant information about the parameter (or parameters) we are going to test: in other words we will be looking towards using *sufficient statistics*. Therefore it is hardly surprising that we usually use the same statistics as we would in calculating confidence intervals. Let us try to work out how we might do this by applying common-sense to an example.

Example 4.3 *Shoshoni bead rectangles*

We want to test whether the Shoshoni instinctively made their rectangles conform to the golden ratio. That is we want to test

$$H_0 : \theta = 0.618 \quad \text{against} \quad H_1 : \theta \neq 0.618.$$

Let us start by assuming the data are normally distributed (we haven't checked this, but let us proceed anyway) with mean θ and variance σ^2.

$$X_i \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \Rightarrow \quad \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1)$$

Since we do not know σ^2 we use the t-statistic,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t\,(n-1)\,.$$

We have 20 measurements so, under the null hypothesis $\theta = 0.618$ gives

$$\frac{\sqrt{20}(\overline{X} - 0.618)}{S} \sim t\,(19)\,,$$

where $S^2 = \frac{1}{20}\sum_{i=1}^{20}(X_i - \overline{X})^2$.

For these data, $\overline{x} = 0.660$, $s = 0.093$. Let us look for some evidence which might support the null hypothesis.

Look at the graph of the t-statistic with 19 degrees of freedom.

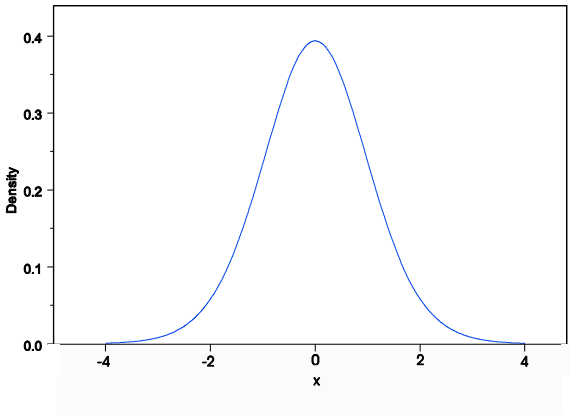

Figure 4.2 Graph of $t(19)$ p.d.f.

Suppose we place the observed value of the above t-statistic on the graph. Where should it be? The observed value of T is

$$\frac{\sqrt{20}(\overline{x} - 0.618)}{s} = \frac{\sqrt{20}(0.660 - 0.618)}{0.093} = 2.019$$

The t-distribution is symmetric and the observed value is to the right. Under the assumption that the null hypothesis holds as above, we can calculate the probability that a measurement of T gives a value at least as extreme as the observed value. Because the alternative hypothesis is *2-sided* this means calculating the following probability

$$P(|T| \geq 2.019) = 0.058.$$

This says that (for a $t\,(19)$ distribution) *the probability of a measurement of T being at least as far into the tails as the observed value is* 0.058. We write $p = 0.058$, and refer to p as the *p-value* or *significance level* of the test. It tells us how far into the tails of the distribution our observed value of the test statistic T lies under the null hypothesis; in this case $H_0 : \theta = 0.618$. *A small p-value gives grounds for rejecting the null hypothesis in favour of the alternative.* However $p = 0.058$ (interpreted as a roughly 1 in 17 chance) is not particularly small. There is some evidence for rejecting the null hypothesis, but the case for it is very weak.
■

Example 4.4 *Patients with glaucoma in one eye*

Here is Table 4.1 again, and we ask "Is there a difference in corneal thickness between the eyes?"

Table 4.1 Glaucoma in one eye

Corneal thickness		
Glaucoma	Normal	Difference
488	484	4
478	478	0
480	492	−12
426	444	−18
440	436	4
410	398	12
458	464	−6
460	476	−16

Formally we are testing the difference θ between the corneal thicknesses.

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0.$$

Assuming the data to be normally distributed (for the sake of this example), the mean difference is $\overline{x} = -4$ and the estimated standard deviation is $s = 10.744$. Under H_0 we obtain a t-statistic of

$$t = \sqrt{n}\frac{\overline{x}}{s} = \frac{-4\sqrt{8}}{10.744} = -1.053.$$

The t-statistic has 7 degrees of freedom for a p-value of 0.327.
We cannot reject the null hypothesis of no difference in corneal thickness.

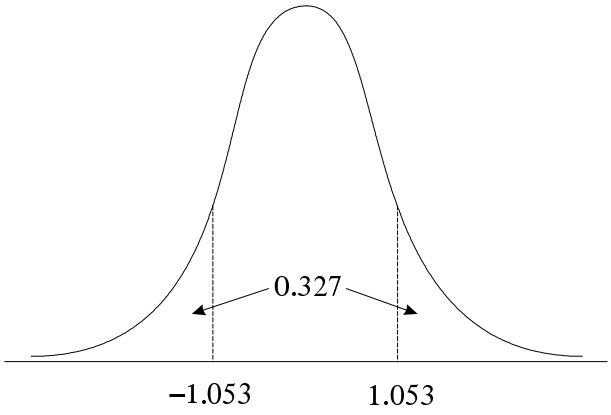

Figure 4.3 Graph of $t(7)$ p.d.f.

■

The p-value is different under different alternative hypotheses.

In Example 4.4 above the natural alternative hypothesis is $H_1 : \theta \neq 0$. However it is possible that we may believe that glaucoma can only reduce corneal thickness, and that no other outcome is possible. In such a case the alternative hypothesis would be $H_1 : \theta < 0$. Does this affect the p-value?

In such a case the tail of interest in the t-distribution would be the *lower tail*. For the one-sided alternative the upper tail no longer provides evidence against the null-hypothesis, so the p-value becomes

$$p = P(T < -1.053) = 0.1635.$$

In this example even in the case of the alternative hypothesis $H_1 : \theta < 0$ there is no strong evidence to suggest that the null hypothesis is false. Whatever the alternative, we have no grounds to reject the hypothesis of no difference in the corneal thickness.

Definition 4.1 *Hypothesis test*

A *hypothesis test* is conducted using a test statistic whose distribution is known under the null hypothesis H_0 , and is used to consider the likely truth of the null hypothesis as opposed to a stated alternative hypothesis H_1.

□

Definition 4.2 *p-value*

The *p-value* (or *significance level* or *size*) is the probability of the test statistic taking a value, in the light of the alternative hypothesis, at least as extreme as its observed value. It is calculated under the assumption that the test statistic has the distribution which it would have if the null hypothesis were true.

If the alternative hypothesis is two-sided it will usually be the case that extreme values occur in two disjoint regions, referring to two tails of the distribution under the null hypothesis.

□

4.3 Testing Normally distributed samples

The considerations of whether Shoshoni bead rectangles had proportions which conformed with the golden ratio and whether glaucoma has an effect on corneal thickness were each an example of a t-test, which is defined below.

Definition 4.3 *t-test*

A *t-test* is used for observations, independently drawn from a normal distribution $N(\mu, \sigma^2)$ with unknown parameters. Given the sample mean $\overline{X}$ and the sample variance S^2, the test statistic is

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \sim t\,(n-1)$$

under the null hypothesis $H_0 : \mu = \mu_0$.

□

Definition 4.4 *Z-test*

For observations drawn from a normal distribution $N(\mu, \sigma^2)$, but with σ^2 known, we use a Z-test of $H_0 : \mu = \mu_0$ with test statistic

$$Z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \sim N(0, 1)$$

under H_0.
□

Definition 4.5 *Paired t-test*

Suppose that we have pairs of random variables (X_i, Y_i) and that $D_i = X_i - Y_i$, $i = 1, \ldots, n$, is a random sample from a normal distribution $N(\mu, \sigma^2)$ with unknown parameters. We use the test statistic

$$\frac{\sqrt{n}(\overline{D} - \mu_0)}{S_D} \sim t\,(n-1)$$

under the null hypothesis $H_0 : \mu = \mu_0$. Here S_D^2 is the sample variance of the differences D_i.
□

Example 4.4 (revisited) *Patients with glaucoma in one eye*

You have already seen how this operates by forming a single sample from the differences in corneal thickness between Glaucomatous and Normal eyes. Then the sample of differences is tested for zero mean. Note, however, that the differences need to be normally distributed, a point we rather glossed over earlier. We can check the validity of this assumption with a Normal probability plot.

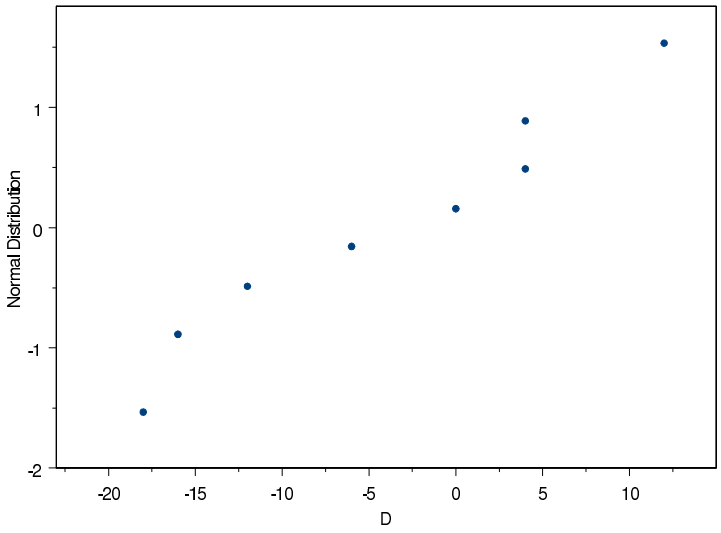

Figure 4.4 Normal probability plot of differences for glaucoma data

It's a bit rough, but it could be worse.
■

One particularly important use of a t-statistic occurs when we have two samples and we wish to compare the means under the assumption of each sample having the same unknown variance.

Definition 4.6 *The two-sample t-test*

Consider two random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ which are independent, normally distributed with the same variance. The null hypothesis is $H_0 : \mu_X = \mu_Y$. Under H_0 we can construct a test statistic T such that

$$T \sim t\,(m+n-2)\,.$$

For the two-sample test, T is constructed under H_0 as follows.

Step 1: Under H_0,

$$\overline{X} - \overline{Y} \sim N\left(0, \sigma^2\left(\frac{1}{m}+\frac{1}{n}\right)\right), \text{ where } \sigma^2 \text{ is the common variance.}$$

Step 2:

$$\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2(m-1), \quad \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\Longrightarrow \quad \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi^2(m+n-2)$$

Step 3: Thus, writing

$$S^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2},$$

we obtain

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)}} \sim t\,(m+n-2)\,,$$

under the null hypothesis, H_0.
□

Example 4.5 *Etruscan and Italian skull widths*

Table 4.3 Ancient Etruscan and modern Italian skull widths

Ancient Etruscan skulls							Modern Italian skulls					
141	147	126	140	141	150	142	133	124	129	139	144	140
148	148	140	146	149	132	137	138	132	125	132	137	130
132	144	144	142	148	142	134	130	132	136	130	140	137
138	150	142	137	135	142	144	138	125	131	132	136	134
154	149	141	148	148	143	146	134	139	132	128	135	130
142	145	140	154	152	153	147	127	127	127	139	126	148
150	149	145	137	143	149	140	128	133	129	135	139	135
146	158	135	139	144	146	142	138	136	132	133	131	138
155	143	147	143	141	149	140	136	121	116	128	133	135
158	141	146	140	143	138	137	131	131	134	130	138	138
150	144	141	131	147	142	152	126	125	125	130	133	
140	144	136	143	146	149	145	120	130	128	143	137	

The width measurements are taken with the aim of comparing modern day Italians with ancient Etruscans. The null hypothesis is therefore that the mean skull width is the same. In what follows X refers to Ancient Etruscan measurements and Y refers to Modern Italian.

$$\overline{x} - \overline{y} = 11.33,\ m = 84,\ n = 70.$$

Using the formula in Definition 4.6 above, the value of the test statistic turns out to be 11.92. As we are just asking "*is there a difference?*", we need a 2-sided alternative hypothesis, and so we test against a $t\,(152)$ distribution and obtain

$$P(|T| \geq 11.92) = 0.0000.$$

The test provides overwhelming evidence to suggest that the two populations are ancestrally of different origin. Of course, we need to verify the plausibility of the data being normally distributed. This is easily done by calculating $X_i - \overline{X}$ for $i = 1, \ldots, m$ and $Y_i - \overline{Y}$ for $i = 1, \ldots, n$ (so that both samples now have the same mean, namely zero) and combining the whole lot into a single sample: then just look at a normal probability plot. Figure 4.5 shows such a plot for the skulls data, which look convincingly normal.

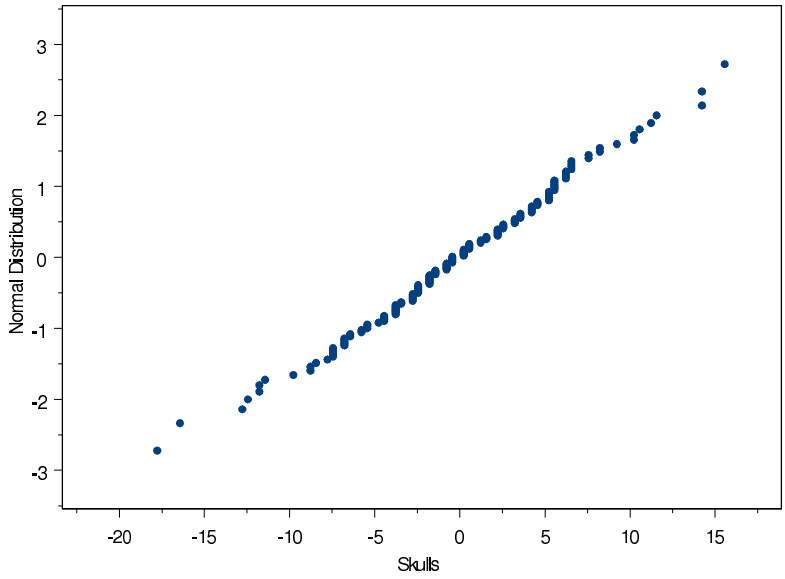

Figure 4.5 Normal probability plot for skulls data

Note that

$$P\left(-t_{\alpha/2}\left(m+n-2\right) < \frac{\overline{X}-\overline{Y}-\left(\mu_X-\mu_Y\right)}{S\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)}} < t_{\alpha/2}\left(m+n-2\right)\right) = 1-\alpha$$

the confidence interval for $\mu_X - \mu_Y$ at the 95% level is given by

$$\left(\left(\overline{X}-\overline{Y}\right) - St_{\alpha/2}\left(m+n-2\right)\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)}, \left(\overline{X}-\overline{Y}\right) + St_{\alpha/2}\left(m+n-2\right)\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)}\right).$$

For the skull data this works out to be $(9.45, 13.21)$. Note that we have a p-value less than 0.05 and the 95% confidence interval does not contain 0, the value of $\mu_X - \mu_Y$ under the null hypothesis. This is not a coincidence - there is a link here.

■

4.4 Hypothesis testing and confidence intervals

The connection here can best be illustrated with an example. Basically when considering random samples with mean μ and null hypothesis $\mu = \mu_0$, *a p-value of less than* α is *equivalent to the appropriate confidence interval at the* $(1-\alpha)100\%$ *level not containing* μ_0.

Example 4.6 *Normal distribution*

Suppose that we have a normal random sample of size n with sample mean $\overline{X}$ and sample variance S^2, and suppose also that the alternative hypothesis is *2-sided*. We can either (i) test for the mean $\mu = \mu_0$ using a t-test and calculate a p-value; or (ii) construct a *central* confidence interval for the mean and see whether or not it contains μ_0. We use a central interval because the test is two-sided.

(i) Carry out a t-test,

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \sim t\,(n-1)\,, \text{ under } H_0 : \mu = \mu_0.$$

Let's suppose that it has observed value t_0 and corresponding p-value of $p < \alpha$. Then

$$P(|T| \geq |t_0|) = p.$$

(ii) Construct a $(1-\alpha)100\%$ confidence interval. Here remember we make no assumptions about μ. Given that X_i has a normal distribution $N(\mu, \sigma^2)$ then $\dfrac{\sqrt{n}(\overline{X} - \mu)}{S}$ has a $t\,(n-1)$ distribution. Choose $t(>0)$ such that

$$P(\left|\frac{\sqrt{n}(\overline{X} - \mu)}{S}\right| \geq t) = \alpha.$$

Then the observed confidence interval for μ is

$$\left(\overline{x} - t\frac{s}{\sqrt{n}}, \overline{x} + t\frac{s}{\sqrt{n}}\right)$$

where $\overline{x}, s$ are the observed values of $\overline{X}, S$.
We must have

$$t_0 = \frac{\sqrt{n}(\overline{x} - \mu_0)}{s}$$

as t_0 is the observed value of the t-statistic. Since $p < \alpha$ we know

$$|t_0| > t \quad \Leftrightarrow \quad t_0 \notin (-t, t) \quad \Leftrightarrow \quad \mu_0 = \overline{x} + t_0\frac{s}{\sqrt{n}} \notin \left(\overline{x} - t\frac{s}{\sqrt{n}}, \overline{x} + t\frac{s}{\sqrt{n}}\right).$$

Hence we can see that there is an equivalence between the test and the interval.
■

In each of the following we can construct a test statistic for $H_0 : \mu = \mu_0$ or construct a confidence interval for μ.

(a) Basic t-test:

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t\,(n-1)\,.$$

(b) Paired t-test: $D_i = X_i - Y_i, \quad S^2 = \dfrac{1}{n-1}\sum \left(D_i - \overline{D}\right)^2$

$$\frac{\sqrt{n}(\overline{D} - \mu)}{S} \sim t\,(n-1)\,.$$

(c) Two-sample t-test, equal variance: $S^2 = \dfrac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S\sqrt{\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}} \sim t\,(m+n-2)$$

For all of these test statistics we have, for each H_1,

2-tailed	upper tailed	lower tailed
$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$

4.5 The magic 5% significance level (or p-value of 0.05)

The question arises in each example considered so far: *what is the critical level for the p-value? Is there some generally accepted level at which null hypotheses are automatically rejected?* Alas, the literature is filled with what purports to be the definitive answer to this question, which is so misleading and ridiculous that it needs special mention.

A significance level of $p < 0.05$ is often taken to be of interest, because it is below the "magic" level of 0.05**.** For example suppose that we had tested a new drug (new drug versus standard drug), which under the null hypothesis of no difference between the two drugs, gave $p = 0.04$. This says that the apparent difference between the two drugs being due to chance is less than 1 in 20. The p-value of 0.05 is the watershed used by the American control board (the FDA, which stands for Food and Drugs Administration) which licences new drugs from pharmaceutical companies**.** As a result it has been almost universally accepted right across the board in all walks of life.

However this level can be, to say the least, inappropriate and possibly even catastrophic.

Suppose, for example, we were considering test data for safety critical software for a nuclear power station, N representing the number of faults detected in the first 10 years. Would we be happy with a p-value on trials which suggests that

$$P(N \geq 1) = 0.05?$$

We might be more comfortable if $p = 0.0001$, but even then, given the number of power stations (over 1000 in Europe alone) we would be justified in worrying. The significance level which should be used in deciding whether or not to reject a null hypothesis ought to depend entirely on the question being asked; it quite properly should depend upon the consequences of being wrong. At the very least we should qualify our rejection with something like the following.

$0.05 < p \leq 0.06$	"Weak evidence for rejection"
$0.03 < p \leq 0.05$	"Reasonable evidence for rejection"
$0.01 < p \leq 0.03$	"Good evidence for rejection"
$0.005 < p \leq 0.01$	"Strong evidence for rejection"
$0.001 < p \leq 0.005$	"Very strong evidence for rejection"
$0.0005 < p \leq 0.001$	"Extremely strong evidence for rejection"
$p \leq 0.0005$	"Overwhelming evidence for rejection"

4.6 The critical region

Suppose we have data $\mathbf{x} = x_1, x_2, \ldots, x_n,\ x \in \mathbb{R}_x$, which constitute evidence about the truth or falsehood of a null hypothesis H_0. Suppose further that we have decided to formulate our test as a decision rule by electing a p-value in advance, say α, and rejecting the null hypothesis in situations where the data lead to a p-value less than or equal to α. In such circumstances we can decide, in advance, on a region $C_1 \subset \mathbb{R}_x$ such that H_0 is rejected if $\mathbf{x} \in C_1$. Should $\mathbf{x} \in C_0$, the complement of C_1, H_0 is not rejected.

C_1 is called the *critical region* of the test and α is called the *significance level*.

Note that α is the probability of rejection of H_0 given that it is true. In other words

$$P(\mathbf{x} \in C_1 \mid H_0) = \alpha.$$

Example 4.4 (revisited) *Patients with glaucoma in one eye*

The significance level is set at 0.05 and H_0 is rejected if $t \leq -2.365$ or $t \geq 2.365$. Here $t = -1.053$ and the null hypothesis is not rejected.

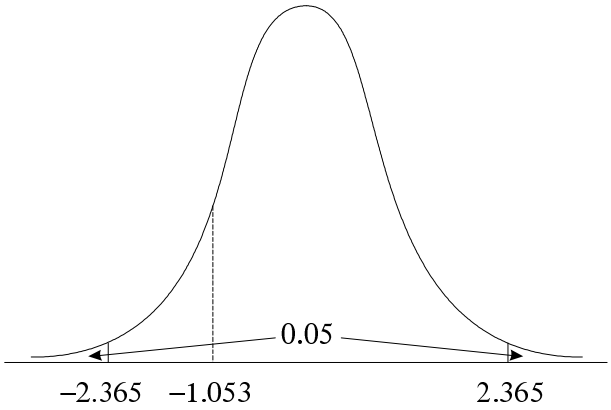

Figure 4.6 The critical region

■

4.7 Errors in hypothesis testing

There are two types of possible error.
A *Type I error* is the error of rejecting the null hypothesis when it is, in fact, true.
A *Type II error* is the error of not rejecting the null hypothesis when it is, in fact, false.

	H_0 not rejected	H_0 rejected
H_0 true	no error	Type I error
H_0 false	Type II error	no error

Thus

$$\begin{aligned} P(\text{Type I error}) &= P(x \in C_1 \mid H_0) = \alpha \\ P(\text{Type II error}) &= P(x \in C_0 \mid H_1) = \beta. \end{aligned}$$

The probability that H_0 is correctly rejected, $P(x \in C_1 \mid H_1) = 1 - \beta$ is called the *power* of the test.

Example 4.7 *Do air bags save lives?*

Suppose that deaths in crashes involving a particular make of car have been at an average rate of 6 per week and that the company has introduced air bags. They want to use the figures over the next year (*i.e.* 52 weeks) to test their effectiveness. Assume the data are from a Poisson distribution with mean μ. The company plans to test

$$H_0 : \mu = 6 \quad \text{against} \quad H_1 : \mu < 6,$$

using a significance level of 0.05.

$$Y = \sum_{i=1}^{52} X_i \sim Poisson(52\mu)$$

and we use a critical region of the form

$$C_1 = \{y : y \leq k\}.$$

Now the distribution of Y may be approximated by $N(52\mu, 52\mu)$ or, under H_0, $N(312, 312)$.

$$\begin{aligned} 0.05 &= P(Y \leq k) \\ &\simeq P\left(Z \leq \frac{k-312}{\sqrt{312}}\right) \end{aligned}$$

where $Z \sim N(0,1)$, so that

$$\frac{k-312}{\sqrt{312}} \simeq -1.645$$

giving $k = 283$ to the nearest integer.
The power of the test is $P(Y \leq 283)$ where $Y \sim Poisson(52\mu)$. Thus

$$\text{Power} \simeq P\left(Z \leq \frac{283-52\mu}{\sqrt{52\mu}}\right) = \Phi\left(\frac{283-52\mu}{\sqrt{52\mu}}\right).$$

Note that at $\mu = 6$ the power has value 0.05 and the power increases as μ decreases, approaching 1 as μ approaches 0.

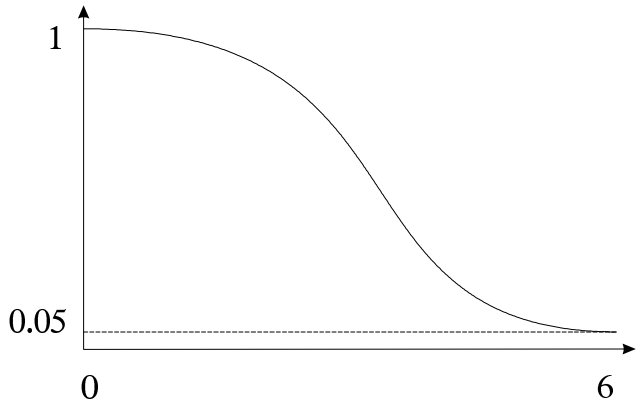

Figure 4.7 The power function

■

4.8 The Neyman-Pearson Lemma

Jerzy Neyman (1894 - 1981)

Lemma 4.1 *The Neyman-Pearson Lemma*

Let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ, where $\theta \in \Theta = \{\theta_0, \theta_1\}$, and let $L(\theta; \mathbf{x})$ be the likelihood function. If there exists a test at significance level α such that, for some positive constant k,

(i)
$$\frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})} \leq k \quad \text{for each } \mathbf{x} \in C_1$$

and

(ii)
$$\frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})} > k \quad \text{for each } \mathbf{x} \in C_0$$

then this test is most powerful at significance level α for testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$.

□

Proof The proof is given for sampling from continuous distributions.

Suppose there exists a critical region C_1 with the properties in the statement of the lemma. Let A_1 be the critical region of any other test at significance level α. Then

$$\int_{C_1} f(\mathbf{x},\theta_0) d\mathbf{x} = \int_{A_1} f(\mathbf{x}, \theta_0) d\mathbf{x} = \alpha,$$

where $f(\mathbf{x}, \theta_0)$ is the joint p.d.f. of the random sample $\mathbf{X}$.

Now
$$\begin{aligned} A_1 \cup C_1 &= A_1 \cup (C_1 \cap \overline{A_1}) \\ &= C_1 \cup (A_1 \cap \overline{C_1}). \end{aligned}$$

It follows therefore that

$$\int_{A_1} f(\mathbf{x}, \theta_1) d\mathbf{x} + \int_{C_1 \cap \overline{A_1}} f(\mathbf{x}, \theta_1) d\mathbf{x}$$

$$= \int_{C_1} f(\mathbf{x}, \theta_1) d\mathbf{x} + \int_{A_1 \cap \overline{C_1}} f(\mathbf{x}, \theta_1) d\mathbf{x}$$

or

$$\int_{C_1} f(\mathbf{x}, \theta_1) d\mathbf{x} - \int_{A_1} f(\mathbf{x}, \theta_1) d\mathbf{x}$$

$$= \int_{C_1 \cap \overline{A_1}} f(\mathbf{x}, \theta_1) d\mathbf{x} - \int_{A_1 \cap \overline{C_1}} f(\mathbf{x}, \theta_1) d\mathbf{x}$$

Now, by property (i), $f(\mathbf{x}, \theta_1) \geq \frac{1}{k} f(\mathbf{x}, \theta_0)$ for each point of C_1 and hence for each point of $C_1 \cap \overline{A_1}$. Furthermore, by property (ii), $f(\mathbf{x}, \theta_1) < \frac{1}{k} f(\mathbf{x}, \theta_0)$ for each point of $\overline{C_1}$ and hence for each point of $A_1 \cap \overline{C_1}$. Therefore

$$\int_{C_1 \cap \overline{A_1}} f(\mathbf{x}, \theta_1) d\mathbf{x} \geq \frac{1}{k} \int_{C_1 \cap \overline{A_1}} f(\mathbf{x}, \theta_0) d\mathbf{x}$$

and

$$\int_{A_1 \cap \overline{C_1}} f(\mathbf{x}, \theta_1) d\mathbf{x} < \frac{1}{k} \int_{A_1 \cap \overline{C_1}} f(\mathbf{x}, \theta_0) d\mathbf{x}.$$

Substituting we obtain

$$\begin{aligned} &\textstyle\int_{C_1} f(\mathbf{x}, \theta_1) d\mathbf{x} - \int_{A_1} f(\mathbf{x}, \theta_1) d\mathbf{x} \\ &\geq \frac{1}{k} \left(\textstyle\int_{C_1 \cap \overline{A_1}} f(\mathbf{x}, \theta_0) d\mathbf{x} - \int_{A_1 \cap \overline{C_1}} f(\mathbf{x}, \theta_0) d\mathbf{x} \right) \\ &= \frac{1}{k} \left(\textstyle\int_{C_1} f(\mathbf{x}, \theta_0) d\mathbf{x} - \int_{A_1} f(\mathbf{x}, \theta_0) d\mathbf{x} \right) = 0 \end{aligned}$$

which is what we set out to prove.
■

Example 4.8 *Insect traps*

Gilchrist (1984) refers to an experiment in which a total of 33 insect traps were set out across sand dunes and the numbers of insects caught in a fixed time were counted. The table gives the number of traps containing various numbers of the taxa *Staphylinoidea*.

Count	0	1	2	3	4	5	6	≥ 7
Frequency	10	9	5	5	1	2	1	0

Assuming the data to come from a Poisson distribution with mean μ, we wish to test the null hypothesis $H_0 : \mu = \mu_0 = 1$ against the alternative $H_1 : \mu = \mu_1 > 1$.

The likelihood is

$$L(\mu; \mathbf{x}) = \frac{e^{-n\mu} \mu^{\sum x_i}}{\prod x_i!}$$

so the Neyman-Pearson Lemma gives

$$\frac{e^{-n\mu_0} \mu_0^{\sum x_i}}{\prod x_i!} \Bigg/ \frac{e^{-n\mu_1} \mu_1^{\sum x_i}}{\prod x_i!} \leq k$$

or

$$\frac{e^{-n\mu_0} \mu_0^{\sum x_i}}{e^{-n\mu_1} \mu_1^{\sum x_i}} \leq k.$$

Taking logs of both sides

$$-n\mu_0 + n\mu_1 + (\log \mu_0 - \log \mu_1) \sum x_i \leq \log k$$

and, rearranging,

$$(\log \mu_0 - \log \mu_1) \sum x_i \leq \log k + n\mu_0 - n\mu_1.$$

Since $\mu_1 > \mu_0$, we obtain from the Neyman-Pearson Lemma a best critical region of the form

$$\sum x_i \geq C.$$

$\sum X_i \sim Poisson(33\mu)$, so, under H_0, $\sum X_i \sim Poisson(33)$.

In fact, the total number of insects counted was 54 and

$$P\left(\sum X_i \geq 54 \mid \mu_0 = 1\right) = 0.000487.$$

This is strong evidence for rejection.

Note that the observed mean is $54/33 = 1.64$, which does not appear to be all that far from 1, but appearances can be deceptive.

■

Example 4.9 *Teenagers with anorexia*

An experiment in the treatment of anorexia by cognitive behavioural therapy was carried out on 29 teenage females. Weights in kilograms before treatment and after 6 weeks of treatment were recorded and the change in weight of each individual , namely (*Weight after treatment−weight before treatment*), was calculated. The mean weight difference of the sample was 3.007 kg with standard deviation 7.309 kg. We want to test

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta = \theta_1 > 0.$$

Note that the test is one-sided because we are actually asking the question "does the therapy have a beneficial effect?" We assume that it either results in weight gain or makes no difference.

Let us look at the question from a Neyman-Pearson point of view. Start by assuming the data are normally distributed (we haven't checked this, but let us proceed anyway) with mean θ and variance σ^2.

$$X_i \sim N(\theta, \sigma^2) \quad \Rightarrow \quad L\left(\theta, \sigma^2; \mathbf{x}\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_i - \theta\right)^2.$$

From the Neyman-Pearson lemma, the test has the form

$$\frac{\left(2\pi\sigma^2\right)^{-n/2} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2}{\left(2\pi\sigma^2\right)^{-n/2} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_i - \theta_1\right)^2} \leq k.$$

Taking logs,

$$-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[x_i^2 - \left(x_i - \theta_1\right)^2\right] \leq \log k$$

or

$$\sum_{i=1}^{n} \left[x_i^2 - \left(x_i - \theta_1\right)^2\right] \geq -2\sigma^2 \log k.$$

This may be written

$$\sum_{i=1}^{n} \theta_1 \left(2x_i - \theta_1\right) \geq -2\sigma^2 \log k$$

or, since $\theta_1 > 0$,

$$\sum_{i=1}^{n} x_i \geq \text{ constant}$$

or, equivalently,

$$\overline{x} \geq \text{ constant.}$$

Under H_0 the random variable $\overline{X} \sim N(0, \sigma^2/n)$. Since we do not know σ^2 we use the t-statistic,

$$\frac{\sqrt{n}(\overline{X})}{S} \sim t(n-1).$$

For a test at the 5% level of significance, the upper 5% tail of $t(28)$ is cut off by 1.701, so the critical region is given by

$$\overline{x} \geq \frac{1.701s}{\sqrt{n}} = \frac{1.701 \times 7.309}{\sqrt{29}} = 2.309.$$

The observed value of $\overline{x}$ is 3.007 so we reject the null hypothesis at the 5% level. Note that the 2% tail is cut off by 2.154 giving a critical region of

$$\overline{x} \geq 2.923$$

so, with an observed value of 3.007, we can actually reject at the 2% level.
■

Notice that we have said absolutely nothing about the alternative hypothesis in the example above. Apart from it's being positive, so that the direction of the inequality was not altered when we divided through by it, it hasn't figured in the calculation. We shall have more to say about this in the next section.

4.9 Uniformly most powerful tests

We have seen that the Neyman-Pearson test applies to a single point null hypothesis against a single point alternative. You might think that this is of no practical use whatsoever because just about all applications involve *composite* alternatives and even sometimes a *composite* null hypothesis as well. You saw this with Example 4.8 on insect traps, where the alternative was a point value of $\mu_1 > 1$, and with Example 4.9 on anorexia, where the alternative was $\theta_1 > 0$.

But suppose we were to regard a composite alternative hypothesis as being made up of a set of simple alternatives and we were to test each simple hypothesis in turn according to the theorem. Could we obtain a *uniformly most powerful test*?

Definition 4.6 *Uniformly most powerful test*

A uniformly most powerful test at significance level α is a test such that its power function Q satisfies

(i) $Q(\theta_0) = \alpha$;

(ii) $Q(\theta)$ is at least as large as the power of any other test at significance level α for each $\theta \in \Theta_1$.

This is illustrated below, where the solid curve represents the power function of a uniformly most powerful test and the broken curve relates to any other test at the same significance level.

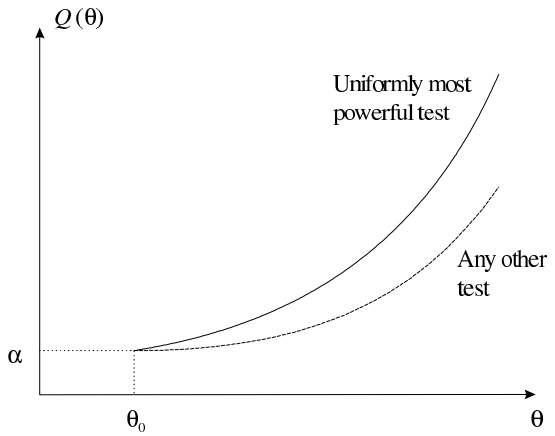

Figure 4.8 Power functions

4.9.1 Tests involving one-sided alternative hypotheses

Let us re-examine Example 4.8 involving insect traps. We obtained the critical region

$$\sum x_i \geq \text{ constant}$$

where, under H_0, $\sum X_i \sim Poisson(33\mu_0)$. Writing

$$\psi_\mu(k) = \sum_{j=0}^{k} \frac{e^{-\mu}\mu^j}{j!}$$

we have a critical region of size α

$$C_1 = \left\{ \mathbf{x} : \sum x_i \geq k \right\}, \quad \alpha = 1 - \psi_{\mu_0}(k).$$

The crucial feature is that this critical region *does not depend upon the value of* θ_1.

The test is simultaneously most powerful at significance level α for testing

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu = \mu_0 + \delta$$

as δ ranges over $\mathbb{R}^+$. In other words the test is for $H_0 : \mu = \mu_0$ against the composite alternative $H_1 : \mu > \mu_0$.

Example 4.10 *Uniform distribution*

$\mathbf{X} = X_1, \ldots, X_n$ constitute a random sample from $U(0, \theta)$. Let us construct a test at significance level 0.1 of $H_0 : \theta = 1$ against the composite alternative $H_1 : \theta > 1$. We have

$$L(\theta; x) = \begin{cases} \theta^{-n}, & 0 < x_i < \theta, \quad 1 \leq i \leq n, \\ 0, & \text{otherwise.} \end{cases}$$

The Neyman-Pearson lemma gives

$$\frac{L(1; \mathbf{x})}{L(\theta_1; \mathbf{x})} = \begin{cases} \theta_1^n, & 0 < x_{(1)}, x_{(n)} < 1, \\ 0, & 0 < x_{(1)}, \ 1 < x_{(n)} < \theta_1. \end{cases}$$

and a critical region $C_1 = \{\mathbf{x} : x_{(n)} \geq c\}$. We need

$$P(\mathbf{X} \in C_1; \theta = 1) = 0.1,$$

that is,

$$P(X_{(n)} \geq c; \theta = 1) = 0.1.$$

$X_{(n)}$ has c.d.f. $F_{(n)}(x) = x^n, \ x \in (0, 1)$ so

$$\begin{aligned} 1 - c^n &= 0.1, \\ c &= 0.9^{1/n}. \end{aligned}$$

The test with critical region $C_1 = \{\mathbf{x} : x_{(n)} \geq 0.9^{1/n}\}$ is most powerful for the test. Since this critical region does not depend upon θ_1, the test is uniformly most powerful against the composite hypothesis $H_1 : \theta > 1$.

The power function of this test is

$$\begin{aligned} Q(\theta) &= P(X_{(n)} \geq 0.9^{1/n}; \theta), \quad \theta \in [1, \infty), \\ &= 1 - F_{(n)}\left(0.9^{1/n}\right) \\ &= 1 - \left[0.9^{1/n} / \theta\right]^n \\ &= 1 - 0.9 / \theta^n \ . \end{aligned}$$

The graphs of this function for $n = 1$ and $n = 5$ are given below. Note that the larger the value of n, the better the power.

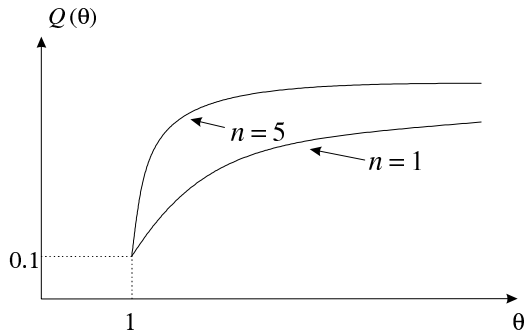

Figure 4.9 Power functions for $n = 1$ and $n = 5$

■

4.9.2 Tests involving two-sided alternative hypotheses

Now consider the problem of testing $H_0 : \theta = \theta_0$ against the two-sided alternative $H_1 : \theta \neq \theta_0$.

It is easy to see that failure to specify the direction of the alternative hypothesis makes the methods we have been using inapplicable. There is no uniformly most powerful test. When the true value of θ is greater than θ_0, then a two-tailed test cannot be more powerful than a one-tailed test which takes account of the information.

We can, however, have an *unbiased* test.

A test at significance level α for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is said to be *unbiased* if

$$Q(\theta) \geq \alpha \quad \text{for all } \theta \in \Theta_1.$$

This means that the power is never less than the significance level; the probability of rejection of H_0 at any element of Θ_0 is necessarily no greater than the probability of rejection at any element of Θ_1.

4.10 Summary of hypothesis testing

There are four main ingredients to a test.

- The critical region C_1.
- The sample size n.
- The significance level (or size) $\alpha = P(x \in C_1 | H_0)$
- The power $Q = P(x \in C_1 | H_1)$.

If any two of these are known, the other two may be determined.

Example 4.8 (revisited) *Sample size calculation*

Suppose that, before carrying out the test for the insect traps $H_0 : \mu = \mu_0 = 1$ against $H_1 : \mu = \mu_1 > 1$, we wanted to determine a suitable sample size. Suppose further that we wanted to specify a significance level of $\alpha = 0.01$ and that we wished to ensure that the test would be powerful enough to reject the null hypothesis by specifying a power of 0.95 for a value of $\mu_1 = 1.5$. We know that $\sum X_i \sim Poisson(n\mu)$, so, under H_0, $\sum X_i \sim Poisson(n)$, and under H_1 with $\mu_1 = 1.5$ we know that $\sum X_i \sim Poisson(1.25n)$.

A normal approximation would give us, under H_0, $\sum X_i \dot{\sim} N(n, n)$ so that

$$\frac{\sum X_i - n}{\sqrt{n}} \dot{\sim} N(0,1) \quad \Rightarrow \quad P\left(\frac{\sum X_i - n}{\sqrt{n}} \geq 2.326\right) \simeq 0.01$$

and the critical region is

$$\sum x_i \geq n + 2.326\sqrt{n}.$$

For a power of 0.95, we require

$$P\left(\sum X_i \geq n + 2.326\sqrt{n} \mid \mu_1 = 1.5\right) = 0.95,$$

which may be re-written

$$P\left(\frac{\sum X_i - 1.5n}{\sqrt{1.5n}} \geq \frac{-0.5n + 2.326\sqrt{n}}{\sqrt{1.5n}}\right) = 0.95.$$

For a standard normal distribution, $P(Z \geq -1.645) = 0.95$, so the approximate sample size can be calculated from

$$\frac{-0.5n + 2.326\sqrt{n}}{\sqrt{1.5n}} \simeq -1.645$$

giving

$$\sqrt{n} = 8.681, \quad n = 75.367,$$

so the recommended sample size would be 76.
■

4.11 The Likelihood Ratio Test

4.11.1 The likelihood ratio

We often want to test in situations where the adopted probability model involves several unknown parameters. Thus we may denote an element of the parameter space by

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$$

Some of these parameters may be *nuisance* parameters, (*e.g.* testing hypotheses on the unknown mean of a normal distribution with unknown variance, where the variance is regarded as a nuisance parameter).

We use the *likelihood ratio*, $\lambda(\mathbf{x})$, defined as

$$\lambda(\mathbf{x}) = \frac{\sup\{L(\boldsymbol{\theta}; \mathbf{x}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}}{\sup\{L(\boldsymbol{\theta}; \mathbf{x}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}}, \quad \mathbf{x} \in \mathbb{R}_X^n.$$

The informal argument for this is as follows.

For a realisation x, determine its best chance of occurrence under H_0 and also its best chance overall. The ratio of these two chances can never exceed unity, but, if small, would constitute evidence for rejection of the null hypothesis.

A *likelihood ratio test* for testing $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ against $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ is a test with critical region of the form

$$C_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\},$$

where k is a real number between 0 and 1.

Clearly the test will be at significance level α if k can be chosen to satisfy

$$\sup\{P(\lambda(\mathbf{X}) \leq k;\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0)\} = \alpha.$$

If H_0 is a simple hypothesis with $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0\}$, we have the simpler form

$$P(\lambda(\mathbf{X}) \leq k;\ \boldsymbol{\theta}_0) = \alpha.$$

To determine k, we must look at the c.d.f. of the random variable $\lambda(\mathbf{X})$, where the random sample $\mathbf{X}$ has joint p.d.f. $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}_0)$.

Example 4.11 *Exponential distribution*

Test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Here $\boldsymbol{\Theta}_0 = \{\theta_0\}$, $\boldsymbol{\Theta}_1 = [\theta_0, \infty)$.

The likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta) = \theta^n e^{-\theta \sum x_i}.$$

The numerator of the likelihood ratio is

$$L(\theta_0; \mathbf{x}) = \theta_0^n e^{-n\theta_0 \overline{x}}.$$

We need to find the supremum as θ ranges over the interval $[\theta_0, \infty)$. Now

$$l(\theta; \mathbf{x}) = n \log \theta - n\theta \overline{x}$$

so that

$$\frac{\partial l(\theta; \mathbf{x})}{\partial \theta} = \frac{n}{\theta} - n\overline{x}$$

which is zero only when $\theta = 1/\overline{x}$. Since $L(\theta; \mathbf{x})$ is an increasing function for $\theta < 1/\overline{x}$ and decreasing for $\theta > 1/\overline{x}$,

$$\sup\{L(\theta; \mathbf{x}) : \theta \in \Theta\} = \begin{cases} \overline{x}^{-n} e^{-n}, & \text{if } 1/\overline{x} \geq \theta_0 \\ \theta_0^n e^{-n\theta_0 \overline{x}} & \text{if } 1/\overline{x} < \theta_0 \end{cases}.$$

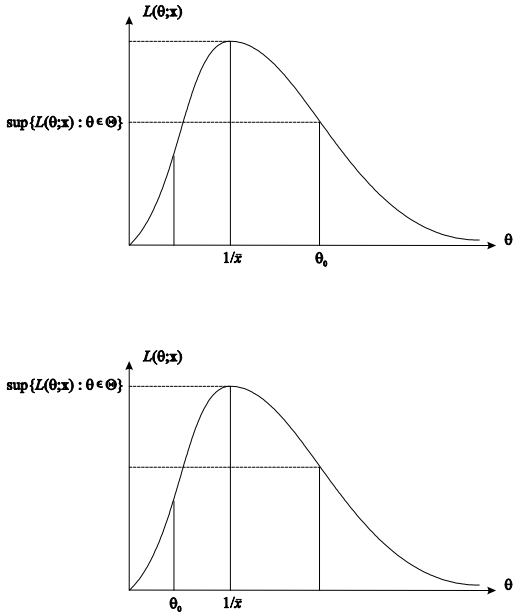

Figure 4.10 Likelihood function

$$\lambda(\mathbf{x}) = \begin{cases} \dfrac{\theta_0^n e^{-n\theta_0 \overline{x}}}{\overline{x}^{-n} e^{-n}}, & 1/\overline{x} \geq \theta_0 \\ 1, & 1/\overline{x} < \theta_0 \end{cases}$$

$$= \begin{cases} \theta_0^n \overline{x}^n e^{-n\theta_0 \overline{x}} e^n, & 1/\overline{x} \geq \theta_0 \\ 1, & 1/\overline{x} < \theta_0 \end{cases}$$

Since

$$\frac{d}{d\overline{x}} \left(\overline{x}^n e^{-n\theta_0 \overline{x}} \right) = n\overline{x}^{n-1} e^{-n\theta_0 \overline{x}} \left(1 - \theta_0 \overline{x}\right)$$

is positive for values of $\overline{x}$ between 0 and $1/\theta_0$ where $\theta_0 > 0$, it follows that $\lambda(\mathbf{x})$ is a non-decreasing function of $\overline{x}$. Therefore the critical region of the likelihood ratio test is of the form

$$C_1 = \left\{ \mathbf{x} : \sum_{i=1}^{n} x_i \leq c \right\}.$$

■

Example 4.12 *The one-sample t-test*

The null hypothesis is $H_0 : \theta = \theta_0$ for the mean of a normal distribution with unknown variance σ^2.

We have

$$\begin{aligned} \boldsymbol{\Theta} &= \{(\theta, \sigma^2) : \theta \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\} \\ \boldsymbol{\Theta}_0 &= \{(\theta, \sigma^2) : \theta = \theta_0,\ \sigma^2 \in \mathbb{R}^+\} \end{aligned}$$

and

$$f(x; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \theta)^2 \right), \quad x \in \mathbb{R}.$$

The likelihood function is

$$L(\theta, \sigma^2; \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

Since

$$l(\theta_0, \sigma^2; \mathbf{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0)^2$$

and

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \theta_0)^2,$$

which is zero when

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta_0)^2$$

we conclude that

$$\sup\left\{L(\theta_0, \sigma^2; \mathbf{x})\right\} = \left(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i - \theta_0)^2\right)^{-n/2} e^{-n/2}.$$

For the denominator, we already know from previous examples that the m.l.e. of θ is $\overline{x}$, so

$$\sup\left\{L(\theta, \sigma^2; \mathbf{x})\right\} = \left(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{-n/2} e^{-n/2}$$

and

$$\lambda(\mathbf{x}) = \left(\frac{\sum_{i=1}^{n}(x_i - \theta_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^{-n/2}.$$

This may be written in a more convenient form. Note that

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \theta_0)^2 &= \sum_{i=1}^{n}\left((x_i - \overline{x}) + (\overline{x} - \theta_0)\right)^2 \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \theta_0)^2\end{aligned}$$

so that

$$\lambda(\mathbf{x}) = \left(1 + \frac{n(\overline{x} - \theta_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^{-n/2}.$$

The critical region is

$$C_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\}$$

so it follows that H_0 is to be rejected when the value of

$$\frac{|\overline{x} - \theta_0|}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

exceeds some constant.

Now we have already seen that

$$\frac{\overline{X}-\theta}{S/\sqrt{n}} \sim t(n-1)$$

where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Therefore it makes sense to write the critical region in the form

$$C_1 = \left\{\mathbf{x} : \frac{|\overline{x}-\theta_0|}{s/\sqrt{n}} \geq c\right\}$$

which is the standard form of the two-sided t-test for a single sample.
■

4.11.2 The likelihood ratio statistic

Since the function $-2\log\lambda(\mathbf{x})$ is a decreasing function, it follows that the critical region of the likelihood ratio test can also be expressed in the form

$$C_1 = \{\mathbf{x} : -2\log\lambda(x) \geq c\}.$$

Writing

$$\Lambda(\mathbf{x}) = -2\log\lambda(\mathbf{x}) = 2\left[l(\widehat{\theta} : \mathbf{x}) - l(\theta_0 : \mathbf{x})\right]$$

the critical region may be written as

$$C_1 = \{\mathbf{x} : \Lambda(\mathbf{x}) \geq c\}$$

and $\Lambda(\mathbf{X})$ is called the *likelihood ratio statistic*.

We have been using the idea that values of θ close to $\widehat{\theta}$ are well supported by the data so, if θ_0 is a possible value of θ, then it turns out that, for large samples,

$$\Lambda(\mathbf{X}) \xrightarrow{D} \chi^2_p$$

where $p = \dim(\theta)$.

Let us see why.

4.11.3 The asymptotic distribution of the likelihood ratio statistic

Write

$$l(\theta_0) = l(\widehat{\theta}) + (\widehat{\theta}-\theta_0)l'(\widehat{\theta}) + \frac{1}{2}(\widehat{\theta}-\theta_0)^2 l''(\widehat{\theta}) + \ldots$$

and, remembering that $l'(\widehat{\theta}) = 0$, we have

$$\begin{aligned}\Lambda &\simeq (\widehat{\theta} - \theta_0)^2 \left[-l''(\widehat{\theta})\right] \\ &= (\widehat{\theta} - \theta_0)^2 J(\widehat{\theta}) \\ &= (\widehat{\theta} - \theta_0)^2 I(\theta_0) \frac{J(\widehat{\theta})}{I(\theta_0)}.\end{aligned}$$

But

$$(\widehat{\theta} - \theta_0) I(\theta_0)^{1/2} \xrightarrow{D} N(0,1) \quad \text{and} \quad \frac{J(\widehat{\theta})}{I(\theta_0)} \xrightarrow{P} 1$$

so

$$(\widehat{\theta} - \theta_0)^2 I(\theta_0) \xrightarrow{D} \chi_1^2$$

or

$$\Lambda \xrightarrow{D} \chi_1^2$$

provided θ_0 is the true value of θ.

Example 4.13 *Poisson distribution*

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a Poisson distribution with parameter θ, and test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at significance level 0.05.

The p.m.f. is

$$p(x; \theta) = \frac{e^{-\theta}\theta^x}{x!}, \quad x = 0, 1, \ldots$$

so that

$$l(\theta : \mathbf{x}) = -n\theta + \sum_{i=1}^{n} x_i \log \theta - \log \prod_{i=1}^{n} x_i!$$

and

$$\frac{\partial l(\theta : \mathbf{x})}{\partial \theta} = -n + \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

giving $\widehat{\theta} = \overline{x}$.

Therefore

$$\Lambda = 2n \left[\theta_0 - \overline{x} + \overline{x} \log\left(\frac{\overline{x}}{\theta_0}\right)\right].$$

The distribution of Λ under H_0 is approximately χ_1^2 and $\chi_1^2(0.95) = 3.84$, so the critical region of the test is

$$C_1 = \left\{\mathbf{x} : 2n \left[\theta_0 - \overline{x} + \overline{x} \log\left(\frac{\overline{x}}{\theta_0}\right)\right] \geq 3.84\right\}.$$

■

4.11.4 Testing goodness-of-fit for discrete distributions

Example 4.14 *Pielou's data on Armillaria root rot in Douglas fir trees*

You have already seen the data below as Data set 4.4. They were collected by the ecologist E.C. Pielou, who was interested in the pattern of healthy and diseased trees. The subject of her research was *Armillaria* root rot in a plantation of Douglas firs. She recorded the lengths of 109 runs of diseased trees.

Table 4.4 Run lengths of diseased trees

Run length	1	2	3	4	5	6
Number of runs	71	28	5	2	2	1

On biological grounds, Pielou proposed a geometric distribution as a probability model. Is this plausible?
□

Let's try to answer this by first looking at the general case.

Suppose we have k groups with n_i in the i^{th} group. Thus

Group	1	2	3	4	$\cdots$	k
Number	n_1	n_2	n_3	n_4	$\cdots$	n_k

where $\sum_i n_i = n$.

Suppose further that we have a probability model such that $\pi_i(\theta),\ i = 1, 2, \ldots, k$, is the probability of being in the i^{th} group. Clearly $\sum_i \pi_i(\theta) = 1$.

The likelihood is

$$L(\theta) = n! \prod_{i=1}^{k} \frac{\pi_i(\theta)^{n_i}}{n_i!}$$

and the log-likelihood is

$$l(\theta) = \sum_{i=1}^{k} n_i \log \pi_i(\theta) + \log n! - \sum_{i=1}^{k} \log n_i!$$

Suppose $\widehat{\theta}$ maximises $l(\theta)$, being the solution of $l'(\widehat{\theta}) = 0$.

The general alternative is to take π_i as unrestricted by the model and subject only to $\sum_i \pi_i = 1$. Thus we maximise

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{k} n_i \log \pi_i + \log n! - \sum_{i=1}^{k} \log n_i! \quad \text{with} \quad g(\boldsymbol{\pi}) = \sum_i \pi_i = 1.$$

Using Lagrange multiplier γ we obtain the set of k equations

$$\frac{\partial l}{\partial \pi_i} - \gamma \frac{\partial g}{\partial \pi_i} = 0, \quad 1 \le i \le k,$$

or

$$\frac{n_i}{\pi_i} - \gamma = 0, \quad 1 \leq i \leq k.$$

Writing this as

$$n_i - \gamma\pi_i = 0, \quad 1 \leq i \leq k$$

and summing over i we find $\gamma = n$ and

$$\widehat{\pi}_i = \frac{n_i}{n}.$$

The likelihood ratio statistic is

$$\begin{aligned} \Lambda &= 2\left[\sum_{i=1}^{k} n_i \log \frac{n_i}{n} - \textstyle\sum_{i=1}^{k} n_i \log \pi_i(\widehat{\theta})\right] \\ &= 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right). \end{aligned}$$

General statement of asymptotic result for the likelihood ratio statistic

Testing $H_0 : \theta \in \Theta_0 \subset \Theta$ against $H_1 : \theta \in \Theta$, the likelihood ratio statistic

$$\Lambda = 2\left[\sup_{\theta\in\Theta} l(\theta) - \sup_{\theta\in\Theta_0} l(\theta)\right] \xrightarrow{D} \chi^2_p,$$

where

$$p = \dim\Theta - \dim\Theta_0$$

In the case above where we are looking at the fit of a one-parameter distribution

$$\Lambda = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right),$$

the restriction $\sum_{i=1}^{k} \pi_i = 1$ means that $\dim\Theta = k - 1$. Clearly $\dim\Theta_0 = 1$ so $p = k - 2$ and

$$\Lambda \xrightarrow{D} \chi^2_{k-2}.$$

Example 4.14 (revisited) *Pielou's data on Armillaria root rot in Douglas fir trees*

The data are

Run length	1	2	3	4	5	6
Number of runs	71	28	5	2	2	1

and Pielou proposed a geometric model with p.m.f.

$$p(x) = (1-\theta)^{x-1}\theta, \quad x = 1, 2, \ldots$$

where x is the observed run length. Thus, if $x_j,\ 1 \leq j \leq n$, are the observed run lengths, the log-likelihood for Pielou's model is

$$l(\theta) = \sum_{j=1}^{n}(x_j - 1)\log(1-\theta) + n\log\theta$$

and, maximising,

$$\frac{\partial l(\theta)}{\partial \theta} = -\frac{\sum_{j=1}^{n} x_j - n}{(1-\theta)} + \frac{n}{\theta}$$

which gives

$$\widehat{\theta} = \frac{1}{\overline{x}}.$$

By the invariance property of m.l.e.'s

$$\pi_i(\widehat{\theta}) = (1-\widehat{\theta})^{i-1}\widehat{\theta} = \frac{(\overline{x}-1)^{i-1}}{\overline{x}^i}.$$

The data give $\overline{x} = 1.523$. We can therefore use the expression for $\pi_i(\widehat{\theta})$ to calculate

$$\Lambda = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right) = 3.547.$$

There are six groups, so $p = 6 - 1 - 1 = 4$.

The approximate distribution of Λ is therefore χ^2_4 and

$$P(\Lambda \geq 3.547) = 0.471.$$

There is no evidence against Pielou's conjecture that a geometric distribution is an appropriate model.

■

Example 4.15 *Flying bomb hits on London*

Data set 4.5 gave the number of flying bomb hits recorded in each of 576 small areas of $\frac{1}{4}km^2$ in the south of London during World War II.

Table 4.5 Flying bomb hits on London

Number of hits in an area	0	1	2	3	4	5	≥ 6
Frequency	229	211	93	35	7	1	0

Propaganda broadcasts claimed that the weapon could be aimed accurately. If, however, this was not the case, the hits should be randomly distributed over the area and should therefore be fitted by a Poisson distribution. Is this the case?

□

The first thing to do is to calculate the m.l.e. of the Poisson parameter. The likelihood function for a sample of size n is

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!},$$

so that the log-likelihood is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} x_i \log \theta - n\theta - \sum_{i=1}^{n} \log x_i! \\ \frac{\mathrm{d}l}{\mathrm{d}\theta} &= \frac{\sum_{i=1}^{n} x_i}{\theta} - n = 0 \end{aligned}$$

and

$$\widehat{\theta} = \overline{x} = \frac{535}{576} = 0.928.$$

Using the Poisson probability mass function with $\theta = 0.929$ we therefore obtain

i	0	1	2	3	4	5	≥ 6
$\pi_i\left(\widehat{\theta}\right)$	0.3949	0.3669	0.1704	0.0528	0.0123	0.0023	0.0004

and hence

$$\Lambda = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right) = 1.4995.$$

This is tested against $\chi^2(\nu)$ where $\nu = k - 2 = 7 - 2 = 5$. This gives $P(\Lambda \geq 1.4995) = 0.913$. Clearly there is not a shred of evidence in favour of rejection.
■

4.11.5 The approximate χ^2 distribution

The tests carried out in Examples 4.14 and 4.15 are not, strictly speaking, correct. The reason for this is that the χ^2-distribution we have used to calculate the p-values is an approximation, and the quality of that approximation depends upon the sample size. Happily there is a general rule of thumb you can use.

Rule of thumb for the χ^2 approximation

An approximate χ^2-distribution may be used for testing count data provided that the expected value of each cell in the table is at least 5. If the expected value of a cell is less than 5, it should be pooled with an adjacent cell or cells to obtain a suitable value.
□

Example 4.14 (revisited) *Pielou's data on Armillaria root rot in Douglas fir trees*

Look at the table with the expected values written in.

Run length	1	2	3	4	5	6	≥7
Number of runs n_i	71	28	5	2	2	1	0
Expected number of runs $n\pi_i(\widehat{\theta})$	71.569	24.577	8.440	2.898	0.995	0.342	0.218

Clearly we need to pool cells to obtain

Run length	1	2	≥ 3
Number of runs n_i	71	28	10
Expected number of runs $n\pi_i(\widehat{\theta})$	71.569	24.577	12.893

The test statistic is now re-calculated to obtain $\Lambda = 1.087$, which is tested as $\chi^2(1)$ to give a p-value of 0.297. The conclusion that there is no evidence against Pielou's conjecture that the underlying distribution is geometric is unaltered.

■

Example 4.15 (revisited) *Flying bomb hits on London*

Including the expected values in the table for flying bomb hits, we obtain the table below.

Number of hits in an area	0	1	2	3	4	5	≥ 6
Frequency	229	211	93	35	7	1	0
Expected frequency	227.462	211.334	98.150	30.413	7.027	1.325	0.230

After pooling, we obtain

Number of hits in an area	0	1	2	3	≥ 4
Frequency	229	211	93	35	8
Expected frequency	227.462	211.334	98.150	30.413	8.582

The test statistic is now re-calculated to obtain $\Lambda = 1.101$, which is tested as $\chi^2(3)$ to give a p-value of 0.777. Again we find no evidence for rejection of the null hypothesis. We have therefore found no evidence that V1 flying bomb could be aimed with any degree of precision.

■

4.11.6 Two-way contingency tables

Data are obtained by cross-classifying a fixed number of individuals according to two criteria. They are therefore displayed as n_{ij} in a table with r rows and c columns as follows.

$$\begin{array}{ccc|c} n_{11} & \cdots & n_{1c} & n_{1.} \\ \vdots & \ddots & \vdots & \vdots \\ n_{r1} & \cdots & n_{rc} & n_{r.} \\ \hline n_{.1} & \cdots & n_{.c} & n \end{array}$$

The aim is to investigate the independence of the two classifications.

Example 4.16 *A famous and historic data set*

These are Pearson's 1909 data on crime and drinking. The data were introduced in Data set 4.6.

Karl Pearson

Table 4.6 Crime and drinking

Crime	*Drinker*	*Abstainer*
Arson	50	43
Rape	88	62
Violence	155	110
Stealing	379	300
Coining	18	14
Fraud	63	144

Is crime drink related?
□

Suppose the k^{th} individual goes into cell (X_k, Y_k), $k = 1, 2, \ldots, n$, and that individuals are independent. Let

$$P\left((X_k, Y_k) = (i, j)\right) = \theta_{ij}, \quad i = 1, 2, \ldots, r; \quad j = 1, 2, \ldots, c,$$

where $\sum_{ij} \theta_{ij} = 1$. The null hypothesis of independence of classifiers can be written $H_0 : \theta_{ij} = \phi_i \rho_j$.

This is on *Problem Sheet 6* so here are a few hints.

The likelihood function is

$$L(\boldsymbol{\theta}) = n! \prod_{i,j} \frac{\theta_{ij}^{n_{ij}}}{n_{ij}!}$$

so the log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{i,j} n_{ij} \log \theta_{ij} + \log n! - \sum_{i,j} \log n_{ij}!$$

Under H_0, put $\theta_{ij} = \phi_i \rho_j$ and maximise with respect to ϕ_i and ρ_j subject to $\sum_i \phi_i = \sum_j \rho_j = 1$. You will obtain

$$\widehat{\phi}_i = \frac{n_{i.}}{n}, \quad \widehat{\rho}_j = \frac{n_{.j}}{n}$$

Under H_1, maximise with respect to θ_{ij} subject to $\sum_{ij} \theta_{ij} = 1$. You will obtain

$$\widehat{\theta}_{ij} = \frac{n_{ij}}{n}$$

and, finally

$$\Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log \left(\frac{n_{ij} n}{n_{i.} n_{.j}} \right).$$

Example 4.16 (continued) *A famous and historic data set*

For these data, $\Lambda = 50.52$.

Under H_0, $\Lambda \sim \chi^2_p$, where $p = \dim \Theta - \dim \Theta_0$. In the notation used earlier, there are apparently 6 values of ϕ_i to estimate, but in fact there are only 5 values because $\sum_i \phi_i = 1$. Similarly there are $2 - 1 = 1$ values of ρ_j. Thus $\dim \Theta_0 = 6$.
Because $\sum_{ij} \theta_{ij} = 1$, $\dim \Theta = 12 - 1 = 11$ so, therefore, $p = 11 - 6 = 5$.

Testing against a χ^2-distribution with 5 degrees of freedom, note that the 0.9999 quantile is 25.75 and we can reject at the 0.0001 level of significance. There is overwhelming evidence that crime and drink are related.
■

Degrees of freedom

It is clear from the above that, when testing contingency tables, the number of degrees of freedom of the resulting χ^2-distribution is given, in general, by

$$\begin{aligned} p &= rc - 1 - (r-1) - (c-1) \\ &= rc - r - c + 1 \\ &= (r-1)(c-1). \end{aligned}$$

4.11.7 Pearson's statistic

For testing independence in contingency tables, let O_{ij} be the observed number in cell (i,j), $i = 1,2,\ldots,r$; $\quad j = 1,2,\ldots,c$, and E_{ij} be the expected number in cell (i,j). Pearson's statistic is

$$P = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}.$$

The expected number E_{ij} in cell (i,j) is calculated under the null hypothesis of independence.

If $n_{i.}$ is the total for the i^{th} row and the overall total is n, then the probability of an observation being in the i^{th} row is estimated by

$$P(i^{th} \text{ row}) = \frac{n_{i.}}{n}.$$

Similarly

$$P(j^{th} \text{ column}) = \frac{n_{.j}}{n}$$

and

$$\begin{aligned} E_{ij} &= n \times P(i^{th} \text{ row}) \times P(j^{th} \text{ column}) \\ &= = \frac{n_{i.} n_{.j}}{n}. \end{aligned}$$

Example 4.16 (revisited) *A famous and historic data set*

These are the data on crime and drinking with the row and column totals.

Crime	*Drinker*	*Abstainer*	*Total*
Arson	50	43	93
Rape	88	62	150
Violence	155	110	265
Stealing	379	300	679
Coining	18	14	32
Fraud	63	144	207
Total	753	673	1426

The E_{ij} are easily calculated.

$$E_{11} = \frac{93 \times 753}{1426} = 49.11, \text{ and so on.}$$

Pearson's statistic turns out to be $P = 49.73$, which is tested against a χ^2-distribution with $(6-1) \times (2-1) = 5$ degrees of freedom and the conclusion is, of course, the same as before.

■

4.11.8 Pearson's statistic and the likelihood ratio statistic

$$P = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Consider the Taylor expansion of $x \log(x/a)$ about $x = a$.

$$x \log\left(\frac{x}{a}\right) = (x-a) + \frac{(x-a)^2}{2a} - \frac{(x-a)^3}{6a^2} + \cdots$$

Now put $x = n_{ij}$ and $a = \dfrac{n_{i.}n_{.j}}{n}$ so that

$$n_{ij} \log\left(\frac{n_{ij}n}{n_{i.}n_{.j}}\right) = \left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right) + \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{2\frac{n_{i.}n_{.j}}{n}} + \cdots$$

Thus

$$\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}n}{n_{i.}n_{.j}}\right)$$

$$= n - n\sum_i \frac{n_{i.}}{n} \sum_j \frac{n_{.j}}{n} + \tfrac{1}{2}\sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} + \cdots \simeq \tfrac{1}{2}P$$

or

$$\Lambda \simeq P$$

Example 4.17 *Snoring and heart disease*

You saw the data in the table below in Data set 4.7.

Table 4.7 Snoring frequency and heart disease

Heart disease	Non-snorers	Occasional snorers	Snore nearly every night	Snore every night	Total
Yes	24	35	21	30	110
No	1355	603	192	224	2374
Total	1379	638	213	254	2484

Is there an association between snoring frequency and heart disease?

You might like to practise on this data set by calculating both $\Lambda = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\dfrac{n_{ij}n}{n_{i.}n_{.j}}\right)$ and $P = \sum_{i,j} \dfrac{(O_{ij} - E_{ij})^2}{E_{ij}}$. You should get

$$\Lambda = 65.904 \quad \text{and} \quad P = 72.782.$$

Each is approximately distributed $\chi^2(3)$ and in each case the p-value is 0. The conclusion is that there is an association between snoring and heart disease.

■