Probability and Statistics II
Statistics part

Contents

1 Sampling distributions related to the normal distribution

Up to now, we have generally assumed that all parameters of a probability distribution we are working with are known. We will now begin the study of situations where the parameters of a probability distribution are not known to us, and must be inferred on the basis of samples from the probability distribution; this is known as *statistical inference*. In this part of the course, we will focus mainly on inference with the use of confidence intervals and hypothesis tests. A large part of statistical inference involves estimation of unknown parameters of probability distributions, however we will not consider this aspect of statistical inference in detail.

In statistical inference, a *population* is defined to be a set of objects (e.g. the population of a country) which we are interested in studying in order to determine certain properties of it (e.g. the average height or income). Since a population is often impractical to survey in its entirety, we base our inferences about a population on a *sample*, defined to be subset of a population. A property over a population will be assumed to have a probability distribution (e.g. heights in a population have a normal distribution) and we will use random variables to model samples from a population. We will use upper case, i.e. Y_i to denote random variables and lower case i.e. y_i to denote observed values of random variables.

Let $Y_1, \ldots, Y_n$ be random variables. A *statistic* $T = T(Y_1, \ldots, Y_n)$ is a function of $Y_1, \ldots, Y_n$ that does not depend on any unknown parameters. Specifically, it can only depend on $Y_1, \ldots, Y_n$ and constants we know. A statistic T is a random variable and has its own probability distribution, known as the *sampling distribution* of T. The probability distribution of T may (or may not) depend on unknown parameters. We will use t_{obs} to denote the observed value of a statistic. We will later see examples of statistics in the context of hypothesis testing.

We next define the notion of a *confidence interval*. Let $Y_1, \ldots, Y_n$ be random variables with some joint distribution that depends on a parameter θ. Let $L(Y_1, \ldots, Y_n) < U(Y_1, \ldots, Y_n)$ be two statistics. We call the (random) interval

$$[L(Y_1, \ldots, Y_n), U(Y_1, \ldots, Y_n)] \tag{1}$$

a $100(1-\alpha)\%$ confidence interval for θ if

$$P(L(Y_1, \ldots, Y_n) < \theta < U(Y_1, \ldots, Y_n)) = 1 - \alpha. \tag{2}$$

For constructing confidence intervals, we will rely on a general technique known as the *pivotal method*. Let $Y_1, \ldots, Y_n$ have a joint distribution which depends on an unknown parameter θ. Next, let $g(Y_1, \ldots, Y_n, \theta)$ be a function of $Y_1, \ldots Y_n$ and θ with a sampling distribution that we know and that *does not* depend on θ. Since we know the sampling distribution of g, we can find a and b with

$$P(a < g(Y_1, \ldots, Y_n, \theta) < b) = 1 - \alpha \tag{3}$$

and rewrite as an inequality in terms of the unknown parameter θ

$$P(L(Y_1, \ldots, Y_n) < \theta < U(Y_1, \ldots Y_n)) = 1 - \alpha \tag{4}$$

to obtain a $100(1-\alpha)\%$ confidence interval for θ. We will see examples of this technique in applications later.

We review some basic terminology and motivation of *hypothesis testing*. Suppose we observe data $y_1, \ldots, y_n$ from an experiment and we have a collection of possible distributions for $Y_1, \ldots, Y_n$, indexed by a parameter θ, which may be vector-valued. Now suppose we have a theory about what θ is. We will denote this theory by H_0 and term it the *null hypothesis*. H_0

states that θ is in a certain set. H_1, which we term the *alternative hypothesis*, states that θ is not in the set defined by H_0. A hypothesis is termed as *simple* if it contains a single point. Otherwise, the hypothesis is known as a *composite* one.

Our goal is to determine whether the data provides evidence against H_0. A *test statistic* $T = T(Y_1, \ldots, Y_n)$ will be used to measure evidence against H_0. Large values of the test statistic will be taken to be evidence against H_0. Whenever the value of T is sufficiently large, we will *reject* the null hypothesis H_0. To formally do this, we divide the set of possible values of T into two regions. If the observed value of T falls into the first one, we will not reject H_0. If it falls into the other one, we will reject H_0, this is known as the *critical region*, or *rejection region*.

When conducting a hypothesis test, two types of errors are possible. A *Type I* error occurs when the null hypothesis is falsely rejected given that it is true — that is, the data comes from a distribution with θ in H_0, yet the observed T lies in the critical region. A *Type II* error occurs when we falsely accept H_0. The *power function* $\pi(\theta)$ associated to a hypothesis test is the probability of rejecting H_0 when the data comes from a distribution with parameter θ. We are particularly interested in values of $\pi(\theta)$ when θ is in H_1, as this tells us how well the test is able correctly reject H_0, when it is false.

The *significance level* of a hypothesis test is defined as the probability of falsely rejecting the null hypothesis given that it is true, i.e. the probability of a Type I error. In testing a hypothesis, we want to minimize the probability of a Type I error. However, doing this alone is insufficient, since we can always make this probability 0 by never rejecting the null hypothesis! This would in turn lead to a lot of Type II errors. Therefore, we need a compromise between the probabilities of each of these errors in choosing how to construct a hypothesis test. A significance level is usually denoted by α and often-used values are 0.05, 0.1, 0.01.

In this chapter we will look at results relating to a normal distribution. You have already seen some in the first half of the course and some others in the first year. These results will enable us to make tests about the parameters of the normal distribution, the mean and variance, and find confidence intervals.

We will define a *random sample* $Y_1, \ldots, Y_n$ to be random variables which are independent and have the same distribution. Another way this is described is to say the Y_is are independent and identically distributed, abbreviated as iid.

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a normal distribution mean μ and variance σ^2. We define the sample mean as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and the sample variance as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Remember the difference between (a) $\bar{Y}$, this is a random variable which is used as an estimator of the population mean μ, it has a distribution and we can find its mean and variance; (b) μ, which is an unknown parameter; (c) $\bar{y}$, the estimate of μ, this is a number which we can calculate once we have measured all the sample values.

Similarly the random variable S^2 is the estimator of the parameter σ^2 and the number s^2 is the estimate.

Lemma 1.1. *The distribution of* $\bar{Y}$ *is* $N(\mu, \sigma^2/n)$.

This follows from the example at the bottom of page 35 in the lecture notes of the probability part of this module since

$$\bar{Y} = \frac{1}{n}Y_1 + \cdots + \frac{1}{n}Y_n$$

so using the results for the distribution of U in that example with $a_i = 1/n$, $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ we see that

$$\mathrm{E}[\bar{Y}] = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu$$

and

$$\mathrm{Var}[\bar{Y}] = \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}.$$

So if

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma}$$

then $Z \sim N(0, 1)$.

Recall also, from an example on page 29 in the lecture notes of the probability part of this module, that if $Z_i = (Y_i - \mu)/\sigma$ then $\sum Z_i^2 \sim \chi_n^2$.

1.1 Test of hypothesis for the mean when the variance is known

In the lecture I discussed testing if a mean is equal to a particular value when the variance is known. This is revision from the first year. Here are some notes on it and an example.

General procedure for hypothesis tests of the mean value

Suppose that μ is the unknown mean of some large population with variance σ^2; that H_0 and H_1 are phrased in terms of μ; that $\bar{X}$ is the mean of a random sample of size n; and that we can assume that $\bar{X}$ is (approximately) normal. Remember that $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{var}(\bar{X}) = \sigma^2/n$.

1. Population variance σ^2 known

(a) One-sided tests (also called one-tailed tests)

(i) H_0: $\mu \leq \mu_0$, where μ_0 is known;
H_1: $\mu > \mu_0$.

$$\text{We know that} \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \text{under } \mathrm{H}_0,$$

so we take $\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ as the test statistic. For significance level α, the rejection region is $\{z : z > z_\alpha\}$.

Given the sample mean $\bar{x}$ of the data, reject H_0 if $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$.

(ii) H_0: $\mu \geq \mu_0$, where μ_0 is known;
H_1: $\mu < \mu_0$.

By a similar argument, reject H_0 if $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$.

(b) Two-sided tests (also called two-tailed tests)

H_0: $\mu = \mu_0$, where μ_0 is known;
H_1: $\mu \neq \mu_0$.

Use the same test statistic $\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ but with the symmetric rejection region $\{z : z < -z_{\alpha/2} \cup z > z_{\alpha/2}\}$.

Reject H_0 if $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{\alpha/2}$ or $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha/2}$.

2. σ^2 unknown, but n large ($n \geq 50$)

As Case 1, but replace the known σ by the sample standard deviation s.

Thus the test statistic is

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

but the rejection region does not change.

These tests are all called "one-sample z-tests".

Example 1.1. *Drills being manufactured are supposed to have a mean length of 4cm. From past experience we know the standard deviation is equal to 1cm and the lengths are normally distributed. A random sample of 10 drills had a mean of 4.5cm. Test the hypothesis that the mean is 4.0 with significance level* $\alpha = 0.05$.

We have

$$H_0 : \mu = 4.0 \quad \textit{versus} \quad H_1 : \mu \neq 4.0$$

We know that

$$\bar{X} \sim N\left(\mu, \frac{1}{10}\right)$$

so if H_0 *is true*

$$Z = \frac{\bar{X} - 4}{\sqrt{1/10}} \sim N(0, 1).$$

The observed value of Z is

$$\frac{4.5 - 4}{\sqrt{1/10}} = 1.58$$

For a 2 sided test with $\alpha = 0.05$ *the rejection region is* $\{z : |z| > 1.96\}$*, as we can compute in R as follows:*

```
> qnorm(0.975)
[1] 1.959964
> qnorm(0.025)
[1] -1.959964
```

Since $z = 1.58$ *we do not reject* H_0 *at the 5% level. Alternatively, we can compute the p-value in R.*

```
> 2*(1-pnorm(1.58))
[1] 0.1141069
```

The p value is $2 \times P(Z > 1.58) = 2 \times (1 - \Phi(1.58)) = 2(1 - 0.9429) = 0.1142$ *and so there is no evidence against* H_0.

1.2 The distribution of the sample variance

Theorem 1.1. *If $Z_1, Z_2, \ldots, Z_n$ are a random sample from a standard normal distribution $N(0,1)$ then*

1. *$\bar{Z}$ has a $N(0, 1/n)$ distribution.*
2. *$\bar{Z}$ and $\sum(Z_i - \bar{Z})^2$ are independent.*
3. *$\sum(Z_i - \bar{Z})^2$ has a chi-square distribution with $n-1$ degrees of freedom.*

These results allow us to find the distribution of the sample variance.

Corollary 1.1. *Suppose $X_1, \ldots, X_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution.*

Let $Z_i = (X_i - \mu)/\sigma$ so that $Z_i \sim N(0,1)$, $\bar{Z} = (\bar{X} - \mu)/\sigma$ and $\bar{Z} \sim N(0, 1/n)$. Then $Z_i - \bar{Z} = (X_i - \bar{X})/\sigma$ and $\sum(Z_i - \bar{Z})^2 = \sum(X_i - \bar{X})^2/\sigma^2$.

It follows that $(\bar{X} - \mu)/\sigma$ and $\sum(X_i - \bar{X})^2/\sigma^2$ are independent which implies that $\bar{X}$ and $\sum(X_i - \bar{X})^2$ are independent and hence $\bar{X}$ and S^2 are independent.

Moreover, it follows that $\sum(Z_i - \bar{Z})^2 = \sum(X_i - \bar{X})^2/\sigma^2 = (n-1)S^2/\sigma^2$ has a χ^2_{n-1} distribution.

From the fact that the mean of a χ^2_{n-1} is $n-1$ we see that

$$\mathrm{E}[(n-1)S^2/\sigma^2] = n - 1$$

and so $\mathrm{E}[S^2] = \sigma^2$.

We can use these results to carry out tests of hypotheses about σ^2 and find confidence intervals for it.

1.3 Test of hypothesis for the variance

Let $X_1, \ldots, X_n$ be a random sample from a population which is $N(\mu, \sigma^2)$ where μ and σ^2 are unknown. To test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ we use the test statistic

$$W = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

if H_0 is true.

Since the χ^2_ν distribution is defined on $(0, \infty)$ and is skewed two-sided rejection regions are quite complicated. The rejection region is

$$\{w : w > \chi^2_{n-1}(\alpha/2) \cup w < \chi^2_{n-1}(1 - (\alpha/2))\}.$$

For example if $\alpha = 0.05$ and $n = 9$ then from R we have

```
> qchisq(0.025, 8)
[1] 2.179731
> qchisq(0.975, 8)
[1] 17.53455
```

that is,

$$\chi^2_8(0.025) = 17.53 \qquad \chi^2_8(0.975) = 2.180$$

and we would only reject H_0 at the 5% significance level if the observed value of W was outside the interval $[2.180, 17.53]$.

Similarly for a one-sided test, for example if $H_1 : \sigma^2 > \sigma_0^2$ then we would reject H_0 at the α significance level if

$$w > \chi^2_{n-1}(\alpha)$$

Example 1.2. *It is important that the variance of the percentage impurity levels of a chemical don't exceed 4.0. A random sample of 20 consignments had a sample variance of 5.62. Test the hypothesis that the population variance is at most 4.0 at a 5% level of significance and find the P value. (If you aren't sure about P values see Section 3.5.)*

Our null and alternative hypotheses are

$$H_0 : \sigma^2 = 4.0 \qquad H_1 : \sigma^2 > 4.0$$

The test statistic

$$W = \frac{(n-1)S^2}{4.0} \sim \chi^2_{19}$$

if H_0 is true. The observed value of W is $w = \frac{19 \times 5.62}{4} = 26.695$. From R

```
> qchisq(0.95, 19)
[1] 30.14353
```

$\chi^2_{19}(0.05) = 30.14$. *Since our observed value is less than this we fail to reject H_0 at the 5% significance level.*

The P value is $P(W > 26.695)$. Now from R

```
> pchisq(26.695, 19)
[1] 0.8880396
```

with $\nu = 19$ $P(W < 26.695) = 0.8880$. Thus $P(W > 26.695) = 1 - 0.8880 = 0.1120$. Using our scale of evidence there is no evidence against H_0.

For a two sided alternative the P value is found using the formula $2 \times \min\{P(W < w_{\text{obs}}, P(W > w_{\text{obs}})\}$. This ensures the P value lies between 0 and 1.

Example 1.3. *Take the data in the last example but suppose that we want to test if the variance to equal 4.0 against an alternative it is not equal to 4.0.. Now from R*

```
> qchisq(0.025, 19)
[1] 8.906516
> qchisq(0.975, 19)
[1] 32.85233
```

and as our observed value lies between these value we fail to reject H_0. The P value is $2 \times 0.1120 = 0.2240$.

1.4 Confidence interval for the variance

A $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}, \frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}\right)$$

since

$$P\left[\chi^2_{n-1}(1-\alpha/2) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1}(\alpha/2)\right] = 1-\alpha$$

so rearranging the inequalities for σ^2

$$P\left[\frac{(n-1)S^2}{\chi^2_{n-1}(\alpha/2)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha/2)}\right]$$

This is a probability statement about the random variable S^2. Replacing the random variable by the observed value s^2 we obtain the confidence interval.

One way to interpret the confidence interval is that it is all the values of the null hypotheses we would not reject when carrying out a two sided test with significance level α.

Example 1.4. *Drills being manufactured are supposed to have a mean length of 4cm. From past experience we know the lengths are normally distributed. A random sample of 10 drills had a mean of 4.5cm and sample variance 1.2. Find 95% and 99% confidence intervals for the population variance. Note the following values from R:*

```
> qchisq(0.975, 9)
[1] 19.02277
> qchisq(0.025, 9)
[1] 2.700389
> qchisq(0.995, 9)
[1] 23.58935
> qchisq(0.005, 9)
[1] 1.734933
```

The 95% confidence interval is given by

$$\begin{aligned}\left(\frac{(n-1)s^2}{19.02}, \frac{(n-1)s^2}{2.700}\right) &= \left(\frac{9\times 1.2}{19.02}, \frac{9\times 1.2}{2.700}\right)\\ &= (0.568, 4.000)\,.\end{aligned}$$

The 99% confidence interval is given by

$$\begin{aligned}\left(\frac{(n-1)s^2}{23.59}, \frac{(n-1)s^2}{1.735}\right) &= \left(\frac{9\times 1.2}{23.59}, \frac{9\times 1.2}{1.735}\right)\\ &= (0.458, 6.225)\,.\end{aligned}$$

1.5 The t distribution

In this section we shall derive the pdf for the t distribution.

Let $W \sim N(0,1)$ and $V \sim \chi^2_r$ where W and V are independent. Remember this means $V \sim Ga(r/2, 1/2)$. It follows that the joint pdf of W and V is

$$h(w,v) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \frac{1}{\Gamma(r/2)2^{r/2}} v^{\frac{r}{2}-1} e^{-v/2}$$

for $-\infty < w < \infty,\ 0 < v < \infty$ and zero otherwise.

Now define the random variable T by

$$T = \frac{W}{\sqrt{V/r}}.$$

We make the transformation $t = \frac{w}{\sqrt{v/r}}$ $u = v$ which maps $A = \{(w,v) : -\infty < w < \infty,\ 0 < v < \infty\}$ one-to-one and onto $B = \{(t,u) : -\infty < t < \infty,\ 0 < u < \infty\}$. The inverse transformation is $w = \frac{t\sqrt{u}}{\sqrt{r}},\ v = u$ so the Jacobian is

$$J = \begin{vmatrix} \frac{\sqrt{u}}{\sqrt{r}} & \frac{t}{2\sqrt{ur}} \\ 0 & 1 \end{vmatrix} = \frac{\sqrt{u}}{\sqrt{r}}.$$

So the joint pdf of T and U is

$$\begin{aligned} g(t,u) &= \frac{1}{\sqrt{2\pi}} e^{-t^2u/2r} \frac{1}{\Gamma(r/2)2^{r/2}} u^{\frac{r}{2}-1} e^{-u/2} \frac{\sqrt{u}}{\sqrt{r}} \\ &= \frac{1}{\sqrt{2\pi}\Gamma(r/2)2^{r/2}\sqrt{r}} u^{\frac{r+1}{2}-1} e^{-u/2(1+t^2/r)} \\ &= \frac{1}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} u^{\frac{r+1}{2}-1} e^{-u/2(1+t^2/r)}. \end{aligned}$$

So the marginal pdf of T is

$$\begin{aligned} f_T(t) &= \int_0^\infty g(t,u)\, du \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} u^{\frac{r+1}{2}-1} e^{-u/2(1+t^2/r)}\, du. \end{aligned}$$

Now consider

$$\int_0^\infty u^{\frac{r+1}{2}-1} e^{-u/2(1+t^2/r)}\, du.$$

We can see this is like the pdf of a Gamma distribution with $\alpha = (r+1)/2$ and $\beta = (1+t^2/r)/2$ and therefore this integral is equal to

$$\frac{\Gamma((r+1)/2)}{((1+t^2/r)/2)^{(r+1)/2}}$$

Thus

$$f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} \frac{2^{(r+1)/2}}{((1+t^2/r))^{(r+1)/2}}.$$

The two's cancel to give us

$$f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{\pi r}\Gamma(r/2)} \frac{1}{((1+t^2/r))^{(r+1)/2}}.$$

We call the distribution with this pdf a t distribution with r degrees of freedom. Some books call it Student's t distribution because the first person to derive it published his result under the pseudonym of Student.

Note that a t distribution with 1 degree of freedom has pdf

$$\frac{\Gamma((1)}{\sqrt{\pi}\Gamma(1/2)}\frac{1}{((1+t^2/r))}.$$

But we know $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$ so the pdf reduces to

$$\frac{1}{\pi}\left(\frac{1}{1+t^2}\right)$$

which we saw before was the pdf of a Cauchy distribution.

Corollary 1.2. *If $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, the sample mean is $\bar{X} = \frac{1}{n}\sum X_i$ and the sample variance $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ then*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

that is a t distribution with $n-1$ degrees of freedom.

This is because we can write

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} / \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}$$

where $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ so T is the ratio of a $N(0,1)$ rv and the square root of a chi-square rv with $n-1$ degrees of freedom divided by $n-1$.

1.6 Test of hypothesis for the mean when the variance is unknown

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. We assume the values of both μ and σ^2 are unknown.

We want to test $H_0 : \mu = \mu_0$. When σ^2 is known we know from the first year that

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

When σ^2 is unknown we estimate it by S^2 the sample variance defined by

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}.$$

The distribution of this statistic is no longer standard normal. We have seen that

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} \sim t_{n-1}$$

a Student t distribution with $n-1$ degrees of freedom. Note the degrees of freedom is the same as the divisor in the sample variance.

A t distribution has heavier tails than a normal distribution. As the degrees of freedom tends to infinity a t distribution tends to a standard normal. We call the resulting test a t test or one sample t test.

If the alternative hypothesis is $H_1 : \mu \neq \mu_0$ then I shall write the rejection region as $\{t : |t| > t_{n-1}(\alpha/2)\}$. If the alternative hypothesis is $H_1 : \mu > \mu_0$ then the rejection region is $\{t : t > t_{n-1}(\alpha)\}$.

Example 1.5. *A gunpowder manufacturer has developed a new powder which was tested on 8 shells. The manufacturer claims that average muzzle velocity using the new powder is no less than 3000 ft/sec. Test the claim at the $\alpha = 0.025$ level of significance. The observations for the 8 shells were*

3005	*2925*	*2935*	*2965*
2995	*3005*	*2937*	*2905*

Let Y_i be the velocity of shell i. Then $\bar{y} = 2959$ and the sample standard deviation is $s = 39.1$

Our null hypothesis is $H_0 : \mu = 3000$ versus an alternative $H_1 : \mu < 3000$.

The test statistic is

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} \sim t_{n-1}$$

if H_0 is true.

The observed value of T is

$$t = \frac{(2959 - 3000)\sqrt{8}}{39.1} = -2.966.$$

For $\alpha = 0.025$ with 7 degrees of freedom the critical region is $\{t : t < -2.365\}$ as it follows from R

```
> qt(0.025, 7)
[1] -2.364624
```

and so we reject H_0 at the 2.5% significance level.

1.7 Confidence interval for the mean when the variance is unknown

Since $T = \frac{(\bar{Y} - \mu_)\sqrt{n}}{S} \sim t_{n-1}$ we have

$$P\left[-t_{n-1}(\alpha/2) < \frac{(\bar{Y} - \mu_)\sqrt{n}}{S} < t_{n-1}(\alpha/2)\right] = 1 - \alpha$$

rearranging we have

$$P\left[\bar{Y} - \frac{S}{\sqrt{N}} t_{n-1}(\alpha/2) < \mu < \bar{Y} + \frac{S}{\sqrt{N}} t_{n-1}(\alpha/2)\right] = 1 - \alpha$$

Thus replacing the rvs $\bar{Y}$ and S by their observed values a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{y} \pm t_{n-1}(\alpha/2)s/\sqrt{n}.$$

One interpretation of this interval is that it is all the values of the null hypothesis which would not be rejected in a two-sided test with significance level α.

Example 1.6. *Drills being manufactured are supposed to have a mean length of 4cm. From past experience we know the lengths are normally distributed. A random sample of 10 drills had a mean of 4.5cm and sample variance 1.2. Find 95% and 99% confidence intervals for the population mean.*

The 95% confidence interval is given by

$$\begin{aligned} \bar{x} \pm 2.262 s/\sqrt{n} &= 4.5 \pm 2.262 \times \frac{\sqrt{1.2}}{\sqrt{10}} \\ &= 4.5 \pm 0.78 \\ &= (3.72, 5.28). \end{aligned}$$

The 99% confidence interval is given by

$$\begin{aligned} \bar{x} \pm 3.25 s/\sqrt{n} &= 4.5 \pm 3.25 \times \frac{\sqrt{1.2}}{\sqrt{10}} \\ &= 4.5 \pm 1.13 \\ &= (3.37, 5.63) \end{aligned}$$

Note we are using the following values from R:

```
> qt(0.975, 9)
[1] 2.262157
> qt(0.995, 9)
[1] 3.249836
```

1.8 P values

P values can be useful as a hypothesis test with a fixed significance level α does not give any indication of the strength of evidence against the null hypothesis.

The P value depends on whether we have a one-sided or two-sided test.

Consider a one sided test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. Assuming the population variance is unknown so that we are using a t test the P value is

$$P(T > t_{\text{obs}})$$

where t_{obs} is the observed value of t.

For $H_1 : \mu < \mu_0$ the P value is $P(T < t_{\text{obs}})$.

For a two sided test with $H_1 : \mu \neq \mu_0$ it is

$$P(|T| > t_{\text{obs}}) = 2P(T > |t_{\text{obs}}|).$$

In each case we can think of the P value as the probability of obtaining a value of the test statistic more extreme than we did observe assuming that H_0 is true. What is regarded as more extreme depends on the alternative hypothesis. If the P value is small that is evidence that H_0 may not be true.

It is useful to have a scale of evidence to help us interpret the size of the P value. There is no agreed scale but the following may be useful as a first indication:

P value	Interpretation
$P > 0.10$	No evidence against H_0
$0.05 < P < 0.10$	Weak evidence against H_0
$0.01 < P < 0.05$	Moderate evidence against H_0
$0.001 < P < 0.01$	Strong evidence against H_0
$P < 0.001$	Very strong or overwhelming evidence against H_0

Note that the P value is the smallest level of significance that would lead to rejection of the null hypothesis.

Example 1.7. *For the gunpowder example the P value of the test is given by*

$$P(T < -2.966) = P(T > 2.966) \quad \textit{by symmetry}$$

where $T \sim t_7$. *From R we have*

```
> pt(2.966, 7)
[1] 0.9895369
```

so $P(T > 2.966) = 1 - 0.9895 = 0.0105$ *and we have moderate evidence against* H_0.

1.9 Hypothesis tests and confidence intervals for a binomial success probability

Assume that we are operating in the large-sample regime, with $n > 30$, so that the normal approximation can be used. First, we consider hypothesis testing and confidence intervals for a binomial success probability, p. Recall that if $X \sim$ Binomial(n, p) then $X = X_1 + \ldots + X_n$ where X_i, $i = 1, \ldots, n$, are independent and identically distributed Bernoulli(p) random variables. Let $x_1, \ldots, x_n$ be the observed values of $X_1, \ldots, X_n$. Suppose we want to test $H_0 : p = p_0$ versus an alternative. To do this, we can use the following test statistic

$$T = \frac{\bar{X} - p_0}{\sqrt{p_0(1 - p_0)/n}} \tag{5}$$

which for sufficiently large n has an approximately $N(0, 1)$ distribution. An (approximate) $100(1 - \alpha)\%$ confidence interval for p can be constructed as

$$\bar{x} \pm z_{1-\alpha/2}\sqrt{\bar{x}(1 - \bar{x})/n}. \tag{6}$$

1.10 Hypothesis tests and confidence intervals for a Poisson mean

Now suppose $X_1, \ldots, X_n$ is a random sample from the Poisson(λ) distribution with $x_1, \ldots, x_n$ the observed values. Suppose we want to test $H_0 : \lambda = \lambda_0$ versus an alternative. This can be done using the following test statistic

$$T = \frac{\bar{X} - \lambda_0}{\sqrt{\lambda_0/n}} \tag{7}$$

which is approximately $N(0, 1)$ when n is sufficiently large. An approximate $100(1 - \alpha)\%$ confidence interval for λ can be constructed as

$$\bar{x} \pm z_{1-\alpha/2}\sqrt{\bar{x}/n}. \tag{8}$$

2 Goodness of Fit Tests

In this chapter we will consider the statistical question of deciding whether a sample of data may reasonably be assumed to come from a particular distribution.

2.1 Goodness of fit tests for discrete random variables

Suppose we wish to test the hypothesis (or assumption) that a set of data follows a binomial distribution with given parameters.

For example suppose we toss three coins and count the number of heads. We want to test the hypothesis that a coin is equally likely to land head or tail. We do this 120 times and get the following data

Heads	0	1	2	3
Observed frequency	10	35	54	21

Is there any evidence to suggest that the coin is not fair (ie. not equally likely to land head or tail)?

Suppose it was equally likely. Then the number of heads in a single toss, assuming independent trials, would have a binomial distribution with $n = 3$ and $p = \frac{1}{2}$. So writing Y as the number of heads we would have $P[Y = 0] = \frac{1}{8}$, $P[Y = 1] = \frac{3}{8}$ $P[Y = 2] = \frac{3}{8}$ $P[Y = 3] = \frac{1}{8}$. Thus in 120 trials our expected frequencies under a binomial model would be

Heads	0	1	2	3
Expected frequency	15	45	45	15

Now, our observed frequencies are not the same as our expected frequencies. But this might be due to random variation. We know a random variable doesn't always take its mean value. But how surprising is the amount of variation we have here?

We make use of a test statistic X^2 defined as follows

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i are the observed frequencies, E_i are the expected frequencies and k is the number of classes, or values that Y can take.

Now it turns out that if we find the value of X^2 for lots of samples for which our hypothesis is true it has a chi-squared distribution. We can calculate the value of X^2 for our sample. If this value is big, i.e. it is in the right tail of the χ^2 distribution we might regard this as evidence that our hypothesis or assumption is false. (Note if the value of X^2 was very small we might regard this as evidence that the agreement was "too good" and that some cheating had been going on.) Since it is only values in the right tail that cast doubt on the null hypothesis these goodness of fit tests are always one tailed tests.

In our example

$$\begin{aligned} X^2 &= \frac{(10-15)^2}{15} + \frac{(35-45)^2}{45} + \frac{(54-45)^2}{45} + \frac{(21-15)^2}{15} \\ &= \frac{25}{15} + \frac{100}{45} + \frac{81}{45} + \frac{36}{15} \\ &= \frac{75+100+81+108}{45} \\ &= \frac{364}{45} \\ &= 8.08 \end{aligned}$$

To use the chi-squared distribution we need the degrees of freedom, ν. For this goodness of fit test $\nu = k - 1$ where k is the number of categories. In this example, as $k = 4$ we have $\nu = k - 1 = 3$. We need to compute the p-value $P(X^2 > 8.08)$. We find this p-value in R as follows,

```
> 1-pchisq(8.08, 3)
[1] 0.04438696
```

Thus the area to the right of 8.08 is 0.0444. This is quite a small value. It represents the probability of obtaining an X^2 value of 8.08 or more if we carry out this procedure repeatedly on samples which actually do come from a binomial distribution with $p = 0.5$. This is the p-value of the test. A p-value of 0.0444 gives moderate evidence against the hypothesis.

Alternatively, we can compute the rejection region. For example, for a significance level $\alpha = 0.05$, we find in R the following.

```
> qchisq(0.95,3)
[1] 7.814728
```

So the rejection region is $\{X^2 : X^2 > 7.815\}$. As $8.08 > 7.815$, we reject the null hypothesis that the data has a binomial distribution with $p = 0.5$ at the 5% significance level.

There are a couple of factors to complicate the goodness of fit test.

- If any of the expected frequencies (E_i) are less than 5 then we must group adjacent classes so that all expected frequencies are greater than 5.
- If we need to estimate any parameters from the data then the formula for the degrees of freedom is amended to read
$$\nu = k - d - 1$$
where k is the number of classes and d is the number of parameters estimated from the data.

We can illustrate both these ideas in the following example.

Example 2.1. *It is thought that the number of accidents per month at a junction follows a Poisson distribution. The frequency of accidents in 120 months was as follows*

Accidents	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7+*
Observed frequency (O_i)	*41*	*40*	*22*	*10*	*6*	*0*	*1*	*0*

To find the Poisson probabilities we need the mean μ. Since this isn't specified in the question we will have to estimate it from the data. A reasonable estimate is the sample mean of the data. This is

$$\hat{\mu} = \frac{0 \times 41 + 1 \times 40 + 2 \times 22 + \cdots + 6 \times 1}{120} = 1.2$$

Now using the Poisson formula

$$P[Y = y] = \frac{e^{-\hat{\mu}} \hat{\mu}^y}{y!}$$

we can compute the probabilities in the following table

Accidents	*Probability*	E_i	O_i
0	*0.3012*	*36.14*	*41*
1	*0.3614*	*43.37*	*40*
2	*0.2169*	*26.03*	*22*
3	*0.0867*	*10.40*	*10*
4	*0.0261*	*3.13*	*6*
5	*0.0062*	*0.74*	*0*
6+	*0.0015*	*0.18*	*1*

Note that the probabilities have to add to one, so the last class is six or more. Note also that this will mean the total of the expected frequencies must add to the total of the observed frequencies. Check that this holds in each example.

The last three expected frequencies are all less than 5. If we group them together into a class $4+$ *the expected frequency will be 4.05, still less than 5. So we group the last four classes into a class* $3+$ *with expected frequency 14.45 and observed frequency 17. We find* X^2 *as before.*

$$\begin{aligned} X^2 &= \frac{(36.14-41)^2}{36.14} + \frac{(43.37-40)^2}{43.37} + \frac{(26.03-22)^2}{26.03} + \frac{(14.45-17)^2}{14.45} \\ &= 0.65 + 0.26 + 0.62 + 0.45 \\ &= 1.98 \end{aligned}$$

After our grouping there are four classes, $k = 4$*, and we estimated one parameter, the mean* $\hat{\mu}$*, from the data so* $d = 1$*. Thus* $\nu = k - d - 1 = 4 - 1 - 1 = 2$*.*

We can compute the p-value in R as follows.

```
> 1-pchisq(1.98, 2)
[1] 0.3715767
```

Such a large p-value shows no evidence against the hypothesis that the data have a Poisson distribution. Alternatively, for a significance level $\alpha = 0.05$*, the rejection region is* $\{X^2 : X^2 > 5.991\}$*, computed in R as follows*

```
> qchisq(0.95,2)
[1] 5.991465
```

As 1.98<5.99, we cannot reject the hypothesis that the data have a Poisson distribution.

2.2 A goodness of fit test for a continuous random variable

Consider the following example.

Traffic is passing freely along a road. The time interval between successive vehicles is measured (in seconds) and recorded below.

Time interval	0-20	20-40	40-60	60-80	80-100	100-120	120+
No. of cars	54	28	12	10	4	2	0

Test whether an exponential distribution provides a good fit to these data.

We need to estimate the parameter λ of the exponential distribution. Since λ^{-1} is the mean of the distribution it seems reasonable to put $\lambda = 1/\bar{x}$. Now the data are presented as intervals so we will have to estimate the sample mean. It is common to do this by pretending that all the values in an interval are actually at the mid-point of the interval. We will do this whilst recognising that for the exponential distribution, which is skewed, it is a bit questionable.

The calculation for the sample mean is given below.

Midpoint x	Frequency f	fx
10	54	540
30	28	840
50	12	600
70	10	700
90	4	360
110	2	220
	110	3260

thus the estimated mean is $3260/110 = 29.6$. Thus we test if the data are from an exponential distribution with parameter $\lambda = 1/29.6$.

We must calculate the probabilities of lying in the intervals given this distribution.

$$\begin{aligned} P[X < 20] &= \int_0^{20} \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_0^{20} = \\ &= 1 - e^{-20\lambda} \\ &= 0.4912 \end{aligned}$$

$$\begin{aligned} P[20 < X < 40] &= \int_{20}^{40} \lambda e^{-\lambda x} dx \\ &= e^{-20\lambda} - e^{-40\lambda} \\ &= 0.2499 \end{aligned}$$

Similarly

$$\begin{aligned} P[40 < X < 60] &= e^{-40\lambda} - e^{-60\lambda} = 0.1272 \\ P[60 < X < 80] &= e^{-60\lambda} - e^{-80\lambda} = 0.0647 \\ P[80 < X < 100] &= e^{-80\lambda} - e^{-100\lambda} = 0.0329 \\ P[100 < X] &= e^{-100\lambda} = 0.0341 \end{aligned}$$

Multiplying these probabilities by 110 we find the expected frequencies as given in the table below.

Time interval	0-20	20-40	40-60	60-80	80-100	100+
Observed frequency	54	28	12	10	4	2
Expected frequency	54.03	27.49	13.99	7.12	3.62	3.75

We must merge the final two classes so that the expected values are greater than 5. Thus for 80+ we have 6 observed and 7.37 expected.

We find

$$X^2 = \sum \frac{(O-E)^2}{E} = 1.71.$$

Now $\nu = 5 - 1 - 1 = 3$ since after grouping there were 5 classes and we estimated one parameter from the data. From R the p-value is 0.6347 as follows.

```
> 1-pchisq(1.71,3)
[1] 0.6347129
```

Therefore, there is no evidence against the hypothesis that the data follows an exponential distribution.

Example 2.2. *64 observations on a continuous random variable X gave the following frequency table*

Interval	*0-1*	*1-2*	*2-3*	*3-4*	*4-5*	*5-6*	*6-7*	*7-8*
Frequency	*0*	*2*	*7*	*7*	*8*	*11*	*16*	*13*

Test the hypothesis that X has the pdf

$$f(x) = \begin{cases} x/32 & 0 \le x \le 8 \\ 0 & \textit{otherwise} \end{cases}$$

using the 5% significance level.

We see that

$$P(0 < X < 1) = \int_0^1 \frac{x}{32} = \left[\frac{x^2}{64}\right]_0^1 = 1/64$$

and

$$P(1 < X < 2) = \int_1^2 \frac{x}{32} = \left[\frac{x^2}{64}\right]_1^2 = 3/64$$

. We find the other probabilities similarly so we have

Interval	*0-1*	*1-2*	*2-3*	*3-4*	*4-5*	*5-6*	*6-7*	*7-8*
Probability	*1/64*	*3/64*	*5/64*	*7/64*	*9/64*	*11/64*	*13/64*	*15/64*
Expected Frequency	*1*	*3*	*5*	*7*	*9*	*11*	*13*	*15*
Observed Frequency	*0*	*2*	*7*	*7*	*8*	*11*	*16*	*13*

We have to group so that the smallest class is $0-3$ with observed and expected frequency 9. The observed value of X^2 is 1.87. There are $6-0-1=5$ degrees of freedom. We can compute the rejection region in R as follows.

```
> qchisq(0.95,5)
[1] 11.0705
```

If we have a 5% significance level the rejection region is an observed value of $X^2 > 11.07$. Thus we cannot reject the null hypothesis that X has this pdf.

3 Hypothesis tests for two samples

In this chapter we consider examples with two samples. We would want to test that two means or two variances are equal.

3.1 Two independent samples - the two sample t test

Consider the situation where we have two independent samples and we want to test if they come from the same population. In particular if they have the same mean. We shall use the following notation.

We assume that the first sample $X_1, \ldots, X_{n_1}$ is of size n_1 and is normally distributed with mean μ_1 and variance σ^2. We shall denote the sample mean and variance by $\bar{X}$ and S_1^2. We assume that the second sample $Y_1, \ldots, Y_{n_2}$ is of size n_2 and is normally distributed with mean μ_2 and variance σ^2. We shall denote the sample mean and variance by $\bar{Y}$ and S_2^2. Note we are assuming that the samples come from populations with the same variance.

We want to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against an alternative which is often two sided $H_1 : \mu_1 \neq \mu_2$ but which could be one sided.

Because we are assuming the population variances are the same we estimate the variance by what is called the pooled estimate of variance. This is

$$S_0^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{9}$$

It can be shown that in these circumstances the pooled estimate of variance is an unbiased estimator of σ^2. It can also be shown that

$$\frac{(n_1 + n_2 - 2)S_0^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2},$$

since the denominator of the right hand side of (9) divided by σ^2 is the sum of two independent chi-squared random variables with degrees of freedom $n_1 - 1$ and $n_2 - 1$.

Note that if σ^2 were known $\bar{X} \sim N(\mu_1, \sigma^2/n_1)$ and $\bar{Y} \sim N(\mu_2, \sigma^2/n_2)$ it follows that

$$\bar{X} - \bar{Y} \sim N\left((\mu_1 - \mu_2), \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})\right)$$

since the samples are assumed independent. Thus

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0, 1)$$

and so if σ^2 were known we could base a test of $\mu_1 = \mu_2$ on the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

which would have a $N(0, 1)$ distribution if H_0 were true.

Since σ^2 is unknown we replace it by the pooled estimate S_0^2 and as in the one sample case the distribution changes from a normal to a t. The degrees of freedom are $n_1 + n_2 - 2$ since as noted above the distribution of $(n_1 + n_2 - 2)S_0^2/\sigma^2$ is $\chi^2_{n_1+n_2-2}$. Thus our test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S_0\sqrt{1/n_1 + 1/n_2}}$$

which has a $t_{n_1+n_2-2}$ distribution if H_0 is true.

Example 3.1. *Two random samples were independently drawn from two populations. The first sample of size 6 had mean 49.5 and variance 280.3 and the second of size 5 had mean 64.4 and variance 310.3. Is there evidence to indicate a difference in population means?*

We are testing

$$H_0 : \mu_1 = \mu_2 \quad \textit{versus} \quad H_1 : \mu_1 \neq \mu_2$$

The test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S_0\sqrt{1/n_1 + 1/n_2}}$$

which has a $t_{n_1+n_2-2}$ distribution if H_0 is true.

The pooled estimate of variance is given by

$$s_0^2 = \frac{5 \times 280.3 + 4 \times 310.3}{6 + 5 - 2} = 293.63$$

so $s_0 = 17.14$. The observed value of T is therefore

$$t = \frac{49.5 - 64.4}{17.14\sqrt{\frac{1}{6} + \frac{1}{5}}} = -1.40.$$

We can compare this value to a t_9 distribution. The p value will be given by $2 \times P(T < -1.40)$. From R we find

```
2*(pt(-1.4, 9))
[1] 0.1950286
```

so the p value is 0.195. So there is no evidence against H_0.

Example 3.2. *The mean reaction times, in hundredths of a second, of two groups of subjects taking a flashing-light stimulus are given below. The first group consisted of subjects who were new to the project while the subjects in the second group had taken part in previous experiments. Test if experience has had an effect on the mean response at the 5% significance level.*

New	*2.7*	*3.0*	*3.3*	*2.9*	*3.5*	*2.7*	*3.0*	*3.1*	*2.8*	*3.0*
Experienced	*2.7*	*2.5*	*3.0*	*2.7*	*2.6*	*2.5*	*2.9*	*2.7*		

We are going to assume that the data is normally distributed, and therefore that the response time, X, of the new subjects is $N(\mu_1, \sigma^2)$ and the response time, Y, of the experienced subjects is $N(\mu_2, \sigma^2)$. The null and alternative hypothesis are given by

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$

The test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S_0\sqrt{1/n_1 + 1/n_2}}$$

which has a $t_{n_1+n_2-2}$ distribution if H_0 is true.

$$n_1 = 10, \bar{x} = 3.0, (n_1 - 1)s_1^2 = 0.58, s_1^2 = 0.064$$
$$n_2 = 8, \bar{y} = 2.7, (n_2 - 1)s_2^2 = 0.22, s_2^2 = 0.031$$

The pooled estimate of variance is given by

$$s_0^2 = \frac{0.58 + 0.22}{10 + 8 - 2} = 0.05$$

The observed value of T is therefore

$$t = \frac{3.0 - 2.7}{\sqrt{0.05}\sqrt{\frac{1}{10} + \frac{1}{8}}} = 2.828.$$

We can compare this value to a t_{16} distribution. In R we obtain the following.

```
> qt(0.025,16)
[1] -2.119905
```

Therefore, the rejection region for a 5% significance test is $|t| > 2.120$ so we reject the null hypothesis at the 5% level and conclude that experience does have an effect on the mean response.

3.1.1 Confidence interval for the difference in means

If we are asked to estimate the difference in means between two independent normal samples with the same variance we would also want the corresponding confidence interval. This is given by

$$\bar{x} - \bar{y} \pm t_{n_1+n_2-2}(1 - \alpha/2)s_0\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Example 3.3. *Compute a 95% confidence interval for the difference in means for the two samples in Example 3.1.*

In R we obtain the following.

```
> qt(0.025,9)
[1] -2.262157
```

Therefore, the confidence interval is computed as follows.

$$\begin{aligned} 49.9 - 64.4 \pm 2.262 \times 17.14\left(\frac{1}{6} + \frac{1}{5}\right)^{1/2} &= -14.5 \pm 23.48 \\ &= (-37.98, 8.98) \end{aligned}$$

Example 3.4. *Compute a 95% confidence interval for the difference in population means for the data in Example 3.2.*

The confidence interval is computed as follows.

$$\begin{aligned} 3.0 - 2.7 \pm 2.120 \times \sqrt{0.05}\left(\frac{1}{10} + \frac{1}{8}\right)^{1/2} &= 0.3 \pm 0.225 \\ &= (0.075, 0.525) \end{aligned}$$

Note that 0 does not belong to the 95% confidence interval agreeing with our finding that we could reject H_0 at the 5% significance level.

For the two sample t-test we have to make the assumption that the population variances are the same. Is this reasonable? In the next section we test this assumption.

3.2 F test for comparing two variances

It is often desirable to compare two variances. To do this, we introduce the following theorem.

Theorem 3.1. *If the random variables C_1 and C_2 are independent and $C_1 \sim \chi^2_{\nu_1}$ and $C_2 \sim \chi^2_{\nu_2}$ then*

$$\frac{C_1/\nu_1}{C_2/\nu_2} \sim F^{\nu_1}_{\nu_2}$$

that is, the ratio follows an F-distribution with ν_1 degrees of freedom and ν_2 degrees of freedom.

We know that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \qquad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

and they are independent. Therefore, using the theorem above, it follows that

$$\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F^{n_1-1}_{n_2-1}$$

We can use this result to test a null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$. The test statistic is

$$F = \frac{S_1^2}{S_2^2}$$

and it can be shown that $F \sim F^{n_1-1}_{n_2-1}$ if H_0 is true where $F^{n_1-1}_{n_2-1}$ is an F distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

We can find the relevant percentage points by using R.

It is important to note that the validity of the F test relies heavily on the underlying populations of our samples being normally distributed. If they are not the results can be misleading. If possible we should check the normality assumption using suitable plots and tests.

Example 3.5. *Find the rejection region in terms of F if $n_1 = 6$ and $n_2 = 11$.*

Using R we find

```
> qf(0.025, 5, 10)
[1] 0.1510767
> qf(0.975, 5, 10)
[1] 4.236086
```

So the rejection region is the set of observed F $\{F : F < 0.151 \cup F > 4.236\}$.

Example 3.6. *For the data in Example 3.1 the observed value of F is $280.3/310.3 = 0.9033$. Compute the rejection region for the test of $H_0 : \sigma_1^2 = \sigma_2^2$ versus a two-sided alternative*

The rejection region is found as follows. Using R,

```
> qf(0.025, 5, 4)
[1] 0.1353567
> qf(0.975, 5, 4)
[1] 9.364471
```

so we would reject the null hypothesis if observed $F < 0.135$ or $F > 9.364$. Thus we are not rejecting the null hypothesis for the purposes of the test on equality of means.

Example 3.7. *Two random samples were independently drawn from two normal populations. The first sample of size 13 had mean 9.5 and variance 93.3 and the second of size 11 had mean 14.0 and variance 25.2. Test the hypothesis that the populations have the same variance at the 5% significance level.*

The observed value of F *is* $93.3/25.2 = 3.70$*. To carry out a test of* $H_0 : \sigma_1^2 = \sigma_2^2$ *versus a two-sided alternative the rejection region would be* $F > 3.621$ *or* $F < 0.296$*, as we can compute in R as follows.*

```
> qf(0.025, 12, 10)
[1] 0.2964234
> qf(0.975, 12, 10)
[1] 3.620945
```

Thus we reject the null hypothesis at the 5% significance level.

3.2.1 Confidence interval for the ratio of two variances

We can find a confidence interval for the ratio σ_1^2/σ_2^2. As seen in the previous section,

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_2-1}^{n_1-1}$$

Therefore

$$P\left(F_{n_2-1}^{n_1-1}(.025) < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{n_2-1}^{n_1-1}(.975)\right) = 0.95$$

Rearranging so that σ_1^2/σ_2^2 is the subject, we have the 95% random interval

$$P\left(\frac{S_1^2/S_2^2}{F_{n_2-1}^{n_1-1}(.975)} < \frac{\sigma_1^2}{\sigma_2^2} < (\frac{S_1^2/S_2^2}{F_{n_2-1}^{n_1-1}(.025)}\right) = 0.95$$

Thus the 95% confidence interval for σ_1^2/σ_2^2 is

$$\left(\frac{s_1^2/s_2^2}{F_{n_2-1}^{n_1-1}(.975)}, \frac{(s_1^2/s_2^2)}{F_{n_2-1}^{n_1-1}(.025)}\right)$$

Example 3.8. *For the data in Example 3.7 the 95% confidence interval for* σ_1^2/σ_2^2 *is*

$$\left(\frac{3.70}{3.620945}, \frac{3.70}{0.2964234}\right) = (1.022, 12.482).$$

This F test is often used to test whether two variances are equal before carrying out the t test for two independent samples discussed in the previous section, as the equality of variances is an assumption for the t test. If the null hypothesis of the F test is not rejected, one does not have enough evidence to reject the hypothesis that the variances of the two samples are equal, and therefore can carry out the t test. If the null hypothesis of the F test is rejected, one cannot carry out the t test for two independent samples as the assumption of equality of variance for the two samples is not satisfied.

So the next question is: what happens when one would like to test the difference in means between two independent samples, but the null of hypothesis of equal variances is rejected by an F test? This is indeed what happened in Examples 3.1 and 3.6. Since we cannot assume the population variances are equal we cannot use the two sample t-test. We discuss this in the next section.

3.3 An approximate test when variances are unequal

If we want to test equality of two means when we know that the two samples have (known) population variances, we can use the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

which would has a standard normal distribution if the null hypothesis is true. If the variances are unknown, and we have rejected the hypothesis that they are equal, we can use the test statistic

$$T^* = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

Unfortunately the distribution of T^* is not known exactly. We can, however, approximate it by a t distribution with ν^* degrees of freedom where

$$\nu^* = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{s_1^4/n_1^2}{n_1 - 1} + \frac{s_2^4/n_2^2}{n_2 - 1}\right)}.$$

Note that in general ν^* is not an integer.

Example 3.9. *We saw in Example 3.7 that we could not assume that the population variances were equal. We can test the equality of the population means using T^*.*

$$H_0 : \mu_1 = \mu_2 \quad \textit{versus} \quad H_1 : \mu_1 \neq \mu_2$$

The test statistic is

$$T^* = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

which has an approximate t_{ν^*} distribution if H_0 is true where

$$\nu^* = \frac{(93.3/13 + 25.2/11)^2}{\left(\frac{93.3^2/13^2}{12} + \frac{25.2^2/11^2}{10}\right)} = 18.6.$$

Note that

```
> qt(.975,18.6)
[1] 2.096075
```

The observed value of the test statistic is $t^* = -4.5/3.077 = -1.462$. If we use a two sided test with $\alpha = 0.05$ the rejection region is $\{t^* : |t^*| > 2.096\}$ so we don't reject H_0 at the 5% significance level.

There is some dispute about the use of this approximate t-test. Some books recommend that it is always used, however close the sample variances, because the F test relies heavily on normality. Others argue that the approximate test has lower power than the two sample test and if the sample variances are close together it is better to use the two sample test. In this course we will adopt the latter position, checking the equality of variances by the F test and only using the approximate procedure if there is evidence against the variances being equal.

In practice unless the two sample variances are very different, in which case we will probably use the approximate test, the difference in answers between the two methods is minimal.

We can find the approximate confidence interval for the difference in means as

$$\bar{x} - \bar{y} \pm t_{\nu^*}(1 - \alpha/2)\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}.$$

Example 3.10. *For the data in Example 3.7 the approximate 95% confidence interval for* $\mu_1 - \mu_2$ *is*

$$-4.5 \pm 2.096 \times 3.077 = -4.5 \pm 6.45 = (-10.95, 1.95).$$

3.4 Matched pairs t-test

One of the assumptions we make in the two sample t-test is that the two samples are independent. If they are not we can use another test called the matched pairs t-test. This test is appropriate if measurements are taken of pairs of similar subjects. For example, we might have pairs of twins, pigs from the same litter, a pair of measurements on the same individual or pairs of patients who have been matched to be similar. We would expect the measurements on such similar individuals to be similar. This violates the independence assumption needed for a two sample t-test. How do we analyse such data? We find the differences for each pair and then do a 1 sample t-test on the differences. We are assuming that the differences are normally distributed with an unknown mean and variance. We test the null hypothesis that this mean is zero.

Example 3.11. *Sixteen patients sampled at random were matched by age and weight. One of each pair were assigned at random to treatment A and the other to treatment B. A blood test of a certain chemical produced the following results*

A	*14.0*	*5.0*	*8.6*	*11.6*	*12.1*	*5.3*	*8.9*	*10.3*
B	*13.2*	*4.7*	*9.0*	*11.1*	*12.2*	*4.7*	*8.7*	*9.6*

Test whether there is a difference in the two treatments. Find a 90% confidence interval for the mean difference in the treatments.

The differences are $+0.8, +0.3, -0.4, +0.5, -0.1, +0.6, +0.2, +0.7$. *The mean difference is* $\bar{d} = 0.325$, *the variance of the differences is* $s_d^2 = 0.1707$ *so the standard deviation is* $s_d = 0.413$. *The null hypothesis is* $\mu_d = 0$ *versus an alternative that* $\mu_d \neq 0$. *The test statistic is*

$$T = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{\bar{d}\sqrt{n}}{s_d}$$

which has a t distribution with 7 degrees of freedom if H_0 *is true. The observed value of* $t = 2.226$. *In R we obtain*

```
> pt(2.226,7)
[1] 0.9693341
> 1-pt(2.226,7)
[1] 0.0306659
> 2*(1-pt(2.226,7))
[1] 0.0613318
> qt(0.05,7)
[1] -1.894579
```

Comparing this with a t_7 *distribution* $P(t_7 < 2.226) = .9692$ *so the p value is* $2(1 - 0.9692) = 0.0616$ *so there is weak evidence against the null hypothesis.*

A 90% confidence interval is of the form

$$\begin{aligned} \bar{d} \pm t_7(.95)\frac{s_d}{\sqrt{n}} &= 0.325 \pm 1.895 \times \frac{0.413}{\sqrt{8}} \\ &= 0.325 \pm 0.277 \\ &= (0.048, 0.602) \end{aligned}$$

What would be the conclusion if we had wrongly ignored the pairing? We would use a 2 sample t-test. The summary statistics for the two samples are

	A	*B*
n	*8*	*8*
Mean	*9.475*	*9.15*
Variance	*10.16*	*9.93*

The pooled estimate of variance is 10.047 so the observed value of the test statistic is

$$T = \frac{9.475 - 9.15}{\sqrt{10.047}\left(\frac{1}{8} + \frac{1}{8}\right)^{1/2}} = 0.205$$

we are comparing with a t_{14} distribution. In R we find

```
> 2*(1-pt(0.205,14))
[1] 0.8405227
```

so the conclusion would be that there is no difference in means.

Such matching is a simple example of a *designed experiment* with *blocking*. Here we have blocks of size 2 but in more complicated examples we might want, for example, to compare 5 animal feeds. We could do this using 5 animals from the same litter. It is important that biases are not introduced into the experiment so we pay careful attention to allocating diets to animals at random. If we are using the same person twice in a study, once with each treatment, it is important to choose the order they receive the treatments randomly. With a drug treatment it may be necessary to allow time between the two treatments so that the first drug is not still affecting the subject when the second drug is taken. If the subject is a patient with a long term illness requiring continuous treatment this could be a problem. In such a clinical trial it is also important, if practically possible, that the patient receiving the treatment does not know which treatment he is receiving and the doctor assessing their improvement also does not know as again this might introduce biases. The whole subject of experimental design is a huge one in its own right.

3.5 Test of two proportions

Suppose we have collected data in an opinion poll on whether the budget was good for the country from men and women and we want to test the hypothesis that the proportions thinking it was good are equal. Suppose we question n_1 men and n_2 women and x_1 men and x_2 women say it was good. The estimate of the proportions thinking it was good will be $\hat{p_1} = x_1/n_1$ and $\hat{p_2} = x_2/n_2$. We can estimate the difference in proportions by $\hat{p_1} - \hat{p_2}$. To test the hypothesis that the population proportions are equal $H_0 : p_1 = p_2$ we need a test statistic with known distribution if H_0 is true. If n_1 and n_2 are large then by the central limit theorem the distribution of $\hat{p_1} - \hat{p_2}$ is normal. The variance of $\hat{p_1} - \hat{p_2}$ is

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

To estimate this quantity note that if H_0 is true then $p_1 = p_2 = p$ and the best estimate of p is $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$. Thus our test statistic is

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which has a standard normal distribution if H_0 is true.

Example 3.12. *Of 1000 men asked 450 thought the budget was good for the country and of 950 women 390 thought it was good. Test the hypothesis at the 5% level that the same proportion of men and women thought it was good.*

The null hypothesis is $H_0 : p_1 = p_2$ *against* $H_1 : p_1 \neq p_2$.

$\hat{p_1} = 450/1000 = 0.45$, $\hat{p_2} = 390/950 = 0.4105$, $\hat{p} = 840/1950 = .4308$. *The test statistic* Z *given above has a standard normal distribution if* H_0 *is true. The observed value of* Z *is*

$$z = \frac{0.45 - 0.4105}{\sqrt{0.4308 \times 0.5692 \times \left(\frac{1}{1000} + \frac{1}{950}\right)}} = 1.759$$

In R we find

```
> qnorm(0.025)
[1] -1.959964
```

The rejection region is $\{z : |z| > 1.96\}$ *so we don't reject* H_0.

The confidence interval for the difference in proportions is not quite what you would expect from the test. Because we are not assuming that $p_1 = p_2$ we estimate the variance differently. The 95% confidence interval is given by

$$\hat{p_1} - \hat{p_2} \pm 1.96 \times \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}.$$

Example 3.13. *For the opinion poll data the 95% confidence interval is given by*

$$\begin{aligned} .45 - .4105 \pm 1.96\sqrt{\frac{(.45)(.55)}{1000} + \frac{(.4105)(.5895)}{950}} &= .0395 \pm .0439 \\ &= (-0.0044, 0.0834) \end{aligned}$$

As a proportion cannot be negative, the confidence interval becomes (0, 0.0834).

3.6 Comparing two correlation coefficients

Suppose we have data from two normally distributed random variables and we are interested in their correlation ρ. The sample correlation is r. The quantity

$$Z' = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$$

is approximately normally distributed with mean

$$\frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right]$$

and variance $1/(n-3)$ where n is the number of pairs of observations from which r was calculated. The approximation is only valid if $n > 50$. So

$$\frac{\frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] - \frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right]}{\sqrt{1/(n-3)}} \sim N(0,1)$$

Example 3.14. *A correlation coefficient is calculated based on 84 pairs of observations. The null hypothesis is that* $\rho = 0.5$*. Test this hypothesis if* $r = 0.34$*.*

The alternative hypothesis is $\rho \neq 0.5$*. The observed value of* Z' *is* $0.5\ln(1.34/0.66) = 0.3541$*. Also* $0.5\ln(1.5/0.5) = 0.5493$ *and* $\sqrt{1/(84-3)} = 0.1111$*. So if the null hypothesis is true* Z' *has a normal distribution with mean 0.5493 and standard deviation 0.1111. Therefore our test statistic* $(Z' - 0.5493)/0.1111$ *will have a* $N(0,1)$ *distribution. The observed value of the test statistic is* $(0.3541 - 0.5493)/0.1111 = -1.76$ *which is not significant at the 5% significance level. So we cannot reject the null hypothesis.*

We can easily extend this idea to comparing two correlation coefficients. Again we assume the underlying distributions are normal.

Example 3.15. *The correlation coefficient between* X*, the mathematics mark, and* Y *the science mark in Year 10 classes in a large school is* 0.67 *for a group of 75 boys and* 0.42 *for a group of 63 girls. Test the hypothesis that the true correlation coefficient in the whole population of Year 10 girls is the same as that in the population of year 10 boys.*

We have a null hypothesis that says that $\rho_1 = \rho_2$ *so* Z_i' *will have the same mean in each population. So we test the null hypothesis that* $Z_1' - Z_2'$ *has mean 0. Assuming independence of the two samples*

$$\begin{aligned} \mathrm{Var}[Z_1' - Z_2'] &= \mathrm{Var}[Z_1'] + \mathrm{Var}[Z_2'] \\ &= \frac{1}{75-3} + \frac{1}{63-3} \\ &= 0.0306. \end{aligned}$$

The observed values of z_1' *and* z_2' *are* $z_1' = 0.5\ln(1.67/0.33) = 0.8107$ *and* $z_2' = 0.5\ln(1.42/0.58) = 0.4497$*. So the test statistic*

$$\frac{(Z_1' - Z_2') - 0}{\sqrt{\mathrm{Var}[Z_1' - Z_2']}}$$

will be $N(0,1)$ *if* H_0 *is true. The value is*

$$\frac{0.8107 - 0.4497}{\sqrt{0.0306}} = \frac{0.3610}{0.1749} = 2.064$$

and we can reject the null hypothesis of equal correlations at the 5% significance level.

4 Contingency tables

If our data consist of two categorical variables we form a contingency table. We are often interested in whether there is some form of association or lack of independence between the two variables. Exactly what form this association takes depends on the way we collect the data.

For example consider the following:

Example 4.1. *227 randomly selected males were classified by eye and hair colour*

	Eye colour			
Hair colour	*Brown*	*Green/grey*	*Blue*	*Total*
Black	*10*	*24*	*8*	*42*
Brown	*16*	*41*	*26*	*83*
Fair/Red	*5*	*32*	*65*	*102*
Total	*31*	*97*	*99*	*227*

Note that in this example we selected 227 males at random and then classified them according to hair and eye colour. Apart from the grand total of 227 none of the other entries in the table were fixed. We may ask if there is an association, or lack of independence, between the two factors (hair colour and eye colour). Do the proportions (or probabilities) of the three eye colours differ among the sub-populations comprising the three hair colours? Equivalently do the proportions (or probabilities) of the three hair colours differ among the three eye colours? To answer this question we need a test of INDEPENDENCE.

Compare this with the following example.

Example 4.2. *A survey of smoking habits in a sixth form sampled 50 boys and 40 girls at random and the frequencies were noted in the following table.*

	Smoking			
	None	*Light*	*Heavy*	*Total*
Boys	*16*	*20*	*14*	*50*
Girls	*24*	*10*	*6*	*40*
Total	*40*	*30*	*20*	*90*

In this example we chose to sample 50 boys and 40 girls. Before we classified their smoking habits we knew that the row totals would be 50 and 40. We want to know if there is a difference between the sexes. We are comparing two distributions (over smoking habits) so the test is one of SIMILARITY or HOMOGENEITY. The hypothesis we are testing is that the population proportions of boys and girls in each smoking category are the same.

The method of sampling is important and that this determines the hypothesis that we want to test. However it turns out that whatever the method of sampling the method we use to analyse the contingency table is the same. As with goodness of fit problems we find the Expected frequencies under the null hypothesis, calculate X^2 and compare this to an appropriate χ^2 value.

Consider the hair and eye colour example. The null hypothesis is that

$$P(\text{eye colour and hair colour}) = P(\text{eye colour}) \times P(\text{hair colour}).$$

We can estimate $P(\text{brown eyes})$, for example, by the number of people with brown eyes divided by the total number of people (31/227). Similarly we can estimate $P(\text{black hair})$ by the total number of people with black hair divided by the total number of people (42/227). So

if the hypothesis of independence is true $P(\text{brown eyes and black hair})$ will be estimated by $(31/227) \times (42/227)$ and we would expect the number of people in our sample with brown eyes and black hair to be $227 \times (31/227) \times (42/227)$. Similarly the expected number of people in our sample with a particular combination of hair colour and eye colour if the hypothesis of independence is true will be

$$\begin{aligned} E_k &= n \times \text{Row total}/n \times \text{Column total}/n \\ &= \frac{\text{Row total} \times \text{ Column total}}{\text{overall sample size } (n)} \end{aligned}$$

Using this rule the table of expected frequencies is as follows:

	Eye colour			
Hair colour	Brown	Green/grey	Blue	Total
Black	5.74	17.95	18.32	42
Brown	11.33	35.47	36.20	83
Fair/Red	13.93	43.59	44.48	102
Total	31	97	99	227

We calculate X^2 as before as

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where the sum is over all the cells of the contingency table.

The number of degrees of freedom is

$$\nu = (\text{no. of rows} - 1)(\text{no. of columns} - 1).$$

Basically given the row and column totals we only need to know the values of ν cells in the table to determine the rest.

As before $X^2 \sim \chi^2_\nu$ under the null hypothesis of independence and a large value of X^2 gives evidence against the hypothesis.

Here $X^2 = 34.9$ on $\nu = (3-1) \times (3-1) = 4$ degrees of freedom. The P value is $P(X^2 > 34.9)$. From R we find

```
> 1-pchisq(34.9, 4)
[1] 4.870322e-07
```

so there is overwhelming evidence against the hypothesis that hair colour and eye colour are independent.

Consider now the smoking example. Since the row totals are fixed, under the hypothesis of similarity the row proportions or probabilities are the same for each row. It follows that

$$\frac{E_k}{\text{Row total}} = \frac{\text{Column total}}{n}$$

or

$$E_k = \frac{\text{Row total} \times \text{ Column total}}{n}$$

Using this rule the table of expected frequencies is as follows:

	Smoking			
	None	Light	Heavy	Total
Boys	22.22	16.67	11.11	50
Girls	17.78	13.33	8.89	40
Total	40	30	20	90

For the example we find that $X^2 = 7.11$. The degrees of freedom is $(2-1)(3-1) = 2$. Using R we find

```
> 1-pchisq(7.11, 2)
[1] 0.02858137
```

Hence there is moderate evidence against the hypothesis of similarity, moderate evidence that smoking habits differ between the boys and girls.

As we saw with the goodness of fit test X^2 will only have a well approximated χ^2 distribution if all the $E_k > 5$. It may be possible to group rows or columns to achieve this if one of the variables is ordinal (e.g. smoking habits) but if it both are categorical any such grouping is arbitrary. In the case of contingency tables we will relax our condition to say that not more than 20% of the cells of the table should have $E_k < 5$ and none should have $E_k < 1$.

4.1 2×2 contingency tables

For 2×2 tables we can find a formula for the value of X^2 in terms of the entries in the table. If the table is

	Presence	Absence	Total
Group 1	a	b	$a+b$
Group 2	c	d	$c+d$
Total	$a+c$	$b+d$	n

Then

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Example 4.3. *Two areas of heathland are examined; in the larger area 66 sampling units are examined and 58 of them contain a particular species of heather, while in the smaller area 22 units are examined and 12 of these contain that species. Is the species occurring at the same density over the two areas?*

The null hypothesis is that the proportion of units containing the species is the same in the two areas. The 2×2 table obtained from these data is

	Presence	*Absence*	*Total*
Area 1	*58*	*8*	*66*
Area 2	*12*	*10*	*22*
Total	*70*	*18*	*88*

The value of X^2 according to the formula is

$$X^2 = \frac{88(58 \times 10 - 8 \times 12)^2}{66 \times 22 \times 70 \times 18} = 11.26$$

From R we find

```
> 1-pchisq(11.26, 1)
[1] 0.0007919521
```

so the P value is 0.000791952 so we have very strong evidence against the null hypothesis.

For 2×2 tables where any of the entries are at all small we should really apply Yates' correction. We do this by modifying the formula for X^2 to

$$X^2 = \sum \frac{(|O_i - E_i| - 0.5)^2}{E_i}.$$

Consider the heathland example. The table of expected frequencies is

	Presence	Absence	Total
Area 1	52.5	13.5	66
Area 2	17.5	4.5	22
Total	70	18	88

So using Yates' correction we find $X^2 = 9.312$ and from R we find

```
> 1-pchisq(9.312, 1)
[1] 0.002276578
```

concluding that the p value is now 0.002276578. This is still strong evidence against the null hypothesis but the value of X^2 has reduced considerably and in another example might have a more important effect.