6

Two Samples

There is absolutely no point in carrying out an analysis that is more complicated than it needs to be. Occam's razor applies to the choice of statistical model just as strongly as to anything else: simplest is best. The so-called classical tests deal with some of the most frequently-used kinds of analysis, and they are the models of choice for:

- comparing two variances (Fisher's *F* test, var.test),
- comparing two sample means with normal errors (Student's *t*-test, t.test),
- comparing two sample means with non-normal errors (Wilcoxon's rank test, wilcox.test),
- comparing two proportions (the binomial test, prop.test),
- correlating two variables (Pearson's or Spearman's rank correlation, cor.test),
- testing for independence in contingency tables (chi-square test, chisq.test or Fisher's exact test, fisher.test).

Comparing Two Variances

Before we can carry out a test to compare two sample means, we need to test whether the sample variances are significantly different (see p. 42). The test could not be simpler. It is called Fisher's *F* test after the famous statistician and geneticist R. A. Fisher, who worked at Rothamsted, UK. To compare two variances, all you do is **divide the larger variance by the smaller variance**.

Obviously, if the variances are the same, the ratio will be 1. In order to be significantly different, the ratio will need to be significantly bigger than 1 (because the larger variance goes on top, in the numerator). How will we know a significant value of the variance ratio from a non-significant one? The answer, as always, is to look up the **critical value** of the variance ratio. In this case, we want critical values of Fisher's *F*. The R function for this is qf which stands for 'quantiles of the *F* distribution'. For our example of ozone levels in market gardens (see p. 39) there were ten replicates in each garden, so there were $10 - 1 = 9$ degrees of freedom for each garden. In comparing two gardens, therefore, we have 9 d.f. in the numerator and 9 d.f. in the denominator. Although *F* tests in analysis of

Statistics: An Introduction using R M. J. Crawley
© 2005 John Wiley & Sons, Ltd ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)

variance are typically one-tailed (the treatment variance is expected to be larger than the error variance if the means are significantly different, see p. 41), in this case, we had no expectation as to which garden was likely to have the higher variance, so we carry out a two-tailed test ($p = 1 - \alpha/2$). Suppose we work at the traditional $\alpha = 0.05$, then we find the critical value of F like this:

```
qf(0.975,9,9)
```

```
4.025994
```

This means that a calculated variance ratio will need to be greater than or equal to 4.02 in order for us to conclude that the two variances are significantly different at $\alpha = 0.05$. To see the test in action, we can compare the variances in ozone concentration for market gardens B and C:

```
f.test.data<-read.table("c:\\temp\\f.test.data.txt",header=T)
attach(f.test.data)
names(f.test.data)
```

```
[ 1] "gardenB" "gardenC"
```

First, we compute the two variances:

```
var(gardenB)
```

```
[ 1] 1.333333
```

```
var(gardenC)
```

```
[ 1] 14.22222
```

The larger variance is clearly in garden C, so we compute the F ratio like this:

```
F.ratio<-var(gardenC)/var(gardenB)
F.ratio
```

```
[ 1] 10.66667
```

The variance in garden C is more than ten times as big as the variance in garden B. The critical value of F for this test (with 9 d.f. in both the numerator and the denominator) is 4.026 (see qf, above), so we conclude that **since the calculated value is larger than the critical value we reject the null hypothesis**. The null hypothesis was that the two variances were not significantly different, so we accept the alternative hypothesis that the two variances are significantly different. In fact, it is better practice to present the p value associated with the calculated F ratio rather than just to reject the null hypothesis; to do this we use pf rather than qf. We double the resulting probability to allow for the two-tailed nature of the test:

```
2*(1-pf(F.ratio,9,9))
```

```
[ 1] 0.001624199
```

so the probability that the variances are the same is $p < 0.002$. Because the variances are significantly different, it would be wrong to compare the two sample means using Student's t-test.

There is a built-in function called var.test for speeding up the procedure. All we provide are the names of the two variables containing the raw data whose variances are to be compared (we don't need to work out the variances first):

```
var.test(gardenB,gardenC)
```

```
      F test to compare two variances

data: gardenB and gardenC
F = 0.0938, num df = 9, denom df = 9, p-value = 0.001624
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.02328617 0.37743695
sample estimates:
ratio of variances
           0.09375
```

Note that the variance ratio, F, is given as roughly $\frac{1}{10}$ rather than roughly 10 because var.test put the variable name that came first in the alphabet (garden B) on top (i.e. in the numerator) instead of the bigger of the two variances. However, the p value of 0.0016 is correct, and we reject the null hypothesis. These two variances are highly significantly different.

Comparing Two Means

The question is this: given what we know about the variation from replicate to replicate within each sample (the within-sample variance), how likely is it that our two sample means were drawn from populations with the same average? If the answer is highly likely, then we shall say that our two sample means are not significantly different. If it is rather unlikely, then we shall say that our sample means are significantly different. Perhaps a better way to proceed is to work out the probability that the two samples were indeed drawn from populations with the same mean. If this probability is very low (say, less than 5% or less than 1%) then we can be reasonably certain (95% or 99% in these two examples) that the means really are different from one another. Note, however, that we can never be 100% certain; the apparent difference might just be due to random sampling – we just happened to get a lot of low values in one sample, and a lot of high values in the other.

There are two simple tests for comparing two sample means:

- **Student's *t*-test** when the samples are independent, the variances constant, and the errors are Normally distributed, or
- **Wilcoxon rank sum test** when the samples are independent but the errors are **not** Normally distributed (e.g. they are ranks or scores of some sort).

What you should do when these assumptions are violated (e.g. when the variances are different) is discussed later on.

Student's *t*-Test

Student was the pseudonym of W.S. Gosset who published his influential paper in *Biometrika* in 1908. He was prevented from publishing under his own name by dint of the archaic employment laws in place at the time, which allowed his employer, the Guinness Brewing Company, to prevent him publishing independent work. Student's *t*-distribution, later perfected by R. A. Fisher, revolutionized the study of small sample statistics where inferences need to be made on the basis of the sample variance s^2 with the population variance σ^2 unknown (indeed, usually unknowable). The test statistic is the number of standard errors by which the two sample means are separated:

$$t = \frac{\text{difference between the two means}}{\text{s.e. of the difference}} = \frac{\bar{y}\text{A} - \bar{y}\text{B}}{\text{s.e.}_{\text{diff}}}.$$

Now we know the standard error of the mean (see p. 44) but we have not yet met the standard error of the difference between two means. For two independent (i.e. non-correlated) variables, **the variance of a difference is the sum of the separate variances** (see Box 6.1).

Box 6.1. The variance of a difference between two independent samples

We want to work out the sum of squares of a difference between samples A and B. First we express each *y* variable as a departure from its own mean, μ

$$\sum [(y_\text{A} - \mu_\text{A}) - (y_\text{B} - \mu_\text{B})]^2.$$

If we were to divide by the degrees of freedom, we would get the variance of the difference, $\sigma^2_{\bar{y}_\text{A} - \bar{y}_\text{B}}$. Start by calculating the square of the difference:

$$(y_\text{A} - \mu_\text{A})^2 + (y_\text{B} - \mu_\text{B})^2 - 2(y_\text{A} - \mu_\text{A})(y_\text{B} - \mu_\text{B}),$$

then apply summation

$$\sum (y_\text{A} - \mu_\text{A})^2 + \sum (y_\text{B} - \mu_\text{B})^2 - 2\sum (y_\text{A} - \mu_\text{A})(y_\text{B} - \mu_\text{B}).$$

We already know that the average of $\sum (y_\text{A} - \mu_\text{A})^2$ is the variance of population A and the average of $\sum (y_\text{B} - \mu_\text{B})^2$ is the variance of population B (see Box 4.2). So the variance of the **difference** between the two sample means is the **sum** of the variances of the two samples, minus a term $= 2\sum (y_\text{A} - \mu_\text{A})(y_\text{B} - \mu_\text{B})$, i.e. minus two times the covariance of samples A and B (see Box 6.2). However, because the samples from A and B are independently drawn they are uncorrelated, the covariance is zero, and so $2\sum(y_\text{A} - \mu_\text{A})(y_\text{B} - \mu_\text{B}) = 0$. This important result needs to be stated separately

$$\sigma^2_{\bar{y}_\text{A} - \bar{y}_\text{B}} = \sigma^2_\text{A} + \sigma^2_\text{B}.$$

So if two samples are independent, **the variance of the difference is the sum of the two sample variances**. This is **not** true, of course, if the samples are positively or negatively correlated (see p. 97).

This important result allows us to write down the formula for the **standard error of the difference** between two sample means

$$\text{s.e.}_{\text{difference}} = \sqrt{\frac{s_{\text{A}}^2}{n_{\text{A}}} + \frac{s_{\text{B}}^2}{n_{\text{B}}}}.$$

At this stage we have everything we need to carry out Student's *t*-test. Our null hypothesis is that the two sample means are the same, and we shall accept this unless the value of Student's *t* is so large that it is unlikely that such a difference could have arisen by chance alone. For the ozone example introduced on p. 39, each sample has nine degrees of freedom, so we have 18 d.f. in total. Another way of thinking of this is to reason that the complete sample size is 20, and we have estimated two parameters from the data, $\bar{y}_{\text{A}}$ and $\bar{y}_{\text{B}}$, so we have $20 - 2 = 18$ d.f. We typically use 5% as the chance of rejecting the null hypothesis when it is true (this is the Type I error rate). Since we didn't know in advance which of the two gardens was going to have the higher mean ozone concentration (and we usually don't), this is a two-tailed test, so the **critical value** of Student's *t* is:

```
qt(0.975,18)
```

```
[ 1] 2.100922
```

This means that our test statistic needs to be bigger than 2.1 in order to reject the null hypothesis, and hence to conclude that the two means are significantly different at $\alpha = 0.05$. The dataframe is attached like this:

```
t.test.data<-read.table("c:\\temp\\t.test.data.txt",header=T)
attach(t.test.data)
names(t.test.data)
```

```
[ 1] "gardenA" "gardenB"
```

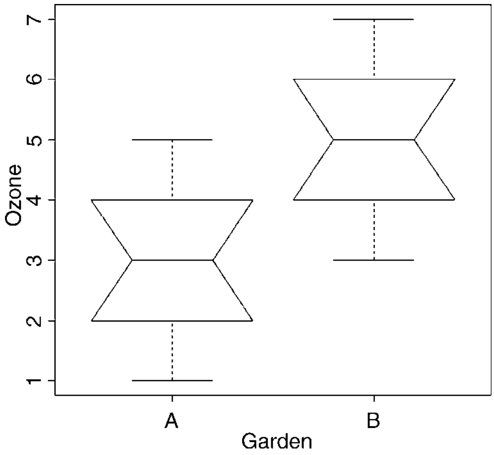
A useful graphical test for two samples employs the 'notches' option of boxplot:

```
ozone<-c(gardenA,gardenB)
label<-factor(c(rep("A",10),rep("B",10)))
boxplot(ozone~label,notch=T,xlab="Garden",ylab="Ozone")
```

Because the notches of two plots do not overlap, we conclude that the medians are significantly different at the 5% level. Note that the variability is similar in both gardens (both in terms of the range – the whiskers – and the inter-quartile range – the boxes).

To carry out a *t*-test longhand, we begin by calculating the variances of the two samples, s2A and s2B:

```
s2A<-var(gardenA)
s2B<-var(gardenB)
```



The value of the test statistic for Student's *t* is: **the difference divided by the standard error of the difference**. The numerator is the difference between the two means, and the denominator is the square root of the sum of the two variances divided by their sample sizes:

```
(mean(gardenA)-mean(gardenB))/sqrt(s2A/10+s2B/10)
```

which gives the value of Student's *t* as

```
[ 1] −3.872983
```

With *t*-tests you can ignore the minus sign; it is only the absolute value of the difference between the two sample means that concerns us. So the calculated value of the test statistic is 3.87 and the critical value is 2.10 (qt(0.975,18), above). This can be written: **since the calculated value is larger than the critical value we reject the null hypothesis**. Notice that the wording is exactly the same as it was for the *F* test (above). Indeed, the wording is always the same for all kinds of tests, and you should try to memorize it. The abbreviated form is easier to remember: **larger reject, smaller accept**. The null hypothesis was that the two means are not significantly different, so we reject this and accept the alternative hypothesis that the two means are significantly different. Again, rather than merely rejecting the null hypothesis, it is better to state the probability that data as extreme as this (or more extreme) would be observed if the mean values were the same. For this we use pt rather than qt, and 2 × pt because we are doing a two-tailed test:

```
2*pt(-3.872983,18)
[ 1] 0.001114540
```

so $p < 0.0015$. You won't be surprised to learn that there is a built-in function to do all the work for us. It is called, helpfully, t.test and is used simply by providing the names of the two vectors containing the samples on which the test is to be carried out (garden A and garden B in our case):

```
t.test(gardenA,gardenB)
```

There is rather a lot of output. You often find this – the simpler the statistical test, the more voluminous the output.

```
       Welch Two Sample t-test

data: gardenA and gardenB
t = -3.873, df = 18, p-value = 0.001115
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -3.0849115 -0.9150885
sample estimates:
mean of x mean of y
        3         5
```

The result is exactly the same as we obtained longhand. The value of t is -3.873 and since the sign is irrelevant in a t test we reject the null hypothesis because the test statistic is larger than the critical value of 2.1. The mean ozone concentration is significantly higher in garden B than in garden A. The computer print-out also gives a p value and a confidence interval. Note that, because the means are significantly different, the confidence interval on the difference does not include zero (in fact, it goes from -3.085 up to -0.915). You might present the result like this: ozone concentration was significantly higher in garden B (mean $=$ 5.0 p.p.h.m.) than in garden A (mean $=$ 3.0 p.p.h.m.; $t = 3.873$, $p = 0.0011$ (two-tailed), d.f. $= 18$).

Wilcoxon Rank Sum Test

This is a non-parametric alternative to Student's t-test, which we could use if the errors were non-Normal. The Wilcoxon rank sum test statistic, W, is calculated as follows. Both samples are put into a single array with their sample names clearly attached (A and B in this case, as explained below). Then the aggregate list is sorted, taking care to keep the sample labels with their respective values. A rank is assigned to each value, with ties getting the appropriate average rank (two-way ties get (rank $i +$ (rank $i + 1$))/2, three-way ties get (rank $i +$ (rank $i + 1$) $+$ (rank $i + 3$))/3, and so on). Finally the ranks are added up for each of the two samples, and significance is assessed on size of the smaller sum of ranks.

First we make a combined vector of the samples

```
ozone < -c(gardenA,gardenB)
ozone
```

```
[ 1] 3 4 4 3 2 3 1 3 5 2 5 5 6 7 4 4 3 5 6 5
```

then make a list of the sample names, A and B

```
label < -c(rep("A",10),rep("B",10))
label
```

```
[ 1] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
```

Now use the built-in function rank to get a vector containing the ranks, smallest to largest, within the combined vector:

```
combined.ranks < -rank(ozone)
combined.ranks
```

```
[ 1]   6.010.510.5 6.0  2.56.01.0 6.015.0 2.5 15.0 15.0 18.5 20.010.5
[ 16] 10.5  6.015.018.5 15.0
```

Notice that the ties have been dealt with by averaging the appropriate ranks. Now all we need to do is calculate the sum of the ranks for each garden. We use tapply with sum as the required operation

```
tapply(combined.ranks,label,sum)
```

```
 A    B
66  144
```

Finally, we compare the smaller of the two values (66) with values in Tables of Wilcoxon rank sums (e.g. Snedecor and Cochran 1980: p. 555), and reject the null hypothesis if our value of 66 is **smaller** than the value in tables. For samples of size ten and ten like ours, the 5% value in tables is 78. Our value is smaller than this, so we reject the null hypothesis. The two sample means are significantly different (in agreement with our earlier *t*-test).

We can carry out the whole procedure automatically, and avoid the need to use tables of critical values of Wilcoxon rank sums, by using the built-in function wilcox.test:

```
wilcox.test(gardenA,gardenB)
```

which produces the following output:

```
        Wilcoxon rank sum test with continuity correction

data: gardenA and gardenB
W = 11, p-value = 0.002988
alternative hypothesis: true mu is not equal to 0

Warning message:
Cannot compute exact p-value with ties in:
wilcox.test.default(gardenA, gardenB)
```

The function uses a normal approximation algorithm to work out a *z* value, and from this a *p* value to assess the hypothesis that the two means are the same. This *p* value of 0.002988 is much less than 0.05, so we reject the null hypothesis, and conclude that the mean ozone concentrations in gardens A and B are significantly different. The warning message at the end draws attention to the fact that there are ties in the data (repeats of the same ozone measurement), and this means that the *p* value cannot be calculated exactly (this is seldom a real worry).

It is interesting to compare the *p* values of the *t*-test and the Wilcoxon test with the same data: $p = 0.001115$ and 0.002988 respectively. The non-parametric test is much more appropriate than the *t*-test when the errors are not normal, and the non-parametric test is about 95% as powerful with normal errors, and can be **more** powerful than the *t*-test if the distribution is strongly skewed by the presence of outliers. Typically, as here, the *t*-test will give the lower *p* value, so the Wilcoxon test is said to be conservative; if a difference is significant under a Wilcoxon test it would have been even more significant under a *t*-test.

Tests on Paired Samples

Sometimes, two-sample data come from paired observations. In this case, we might expect a correlation between the two measurements, either because they were made on the same individual, or were taken from the same location. You might recall that earlier (Box 6.1) we found that the variance of a difference was the average of

$$(y_A - \mu_A)^2 + (y_B - \mu_B)^2 - 2(y_A - \mu_A)(y_B - \mu_B),$$

which is the variance of sample A, plus the variance of sample B, minus two times the covariance of A and B. When the covariance of A and B is **positive**, this is a great help because it reduces the variance of the difference, which makes it easier to detect significant differences between the means. Pairing is not always effective, because the correlation between y_A and y_B may be weak.

The following data are a composite biodiversity score based on a kick sample of aquatic invertebrates.

```
streams<-read.table("c:\\temp\\streams.txt",header=T)
attach(streams)
names(streams)
```

```
[1] "down" "up"
```

The elements are paired because the two samples were taken on the same river, one upstream and one downstream from the same sewage outfall. If we ignore the fact that the samples are paired, it appears that the sewage outfall has no impact on the biodiversity score ($p = 0.6856$):

```
t.test(down,up)
```

```
     Welch Two Sample t-test
data: down and up
t = -0.4088, df = 29.755, p-value = 0.6856
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -5.248256 3.498256
sample estimates:
mean of x mean of y
  12.500    13.375
```

However, if we allow that the samples are paired (simply by specifying the option paired = T), the picture is completely different.

```
t.test(down,up,paired = T)
```

```
      Paired t-test

data: down and up
t = -3.0502, df = 15, p-value = 0.0081
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.4864388 -0.2635612
sample estimates:
mean of the differences
                -0.875
```

Now, the difference between the means is highly significant ($p = 0.0081$). The moral is clear. If you can do a paired *t*-test, then you should always do the paired test. It can never do any harm, and sometimes (as here) it can do a huge amount of good. In general, if you have information on **blocking** or **spatial correlation** (in this case, the fact that the two samples came from the same river), then you should always use it in the analysis.

Here is the same paired test carried out as a one-sample *t*-test on the differences between the pairs:

```
d <- up-down
t.test(d)
```

```
      One Sample t-test

data: d
t = 3.0502, df = 15, p-value = 0.0081
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.2635612 1.4864388
sample estimates:
mean of x
  0.875
```

As you see, the result is identical to the two-sample *t*-test with paired = T ($p = 0.0081$). The upstream values of the biodiversity score were greater by 0.875 on average, and this difference is highly significant. Working with the differences has halved the number of degrees of freedom (from 30 to 15), but it has more than compensated for this by reducing the error variance, because there is such a strong positive correlation between y_A and y_B.

The Sign Test

This is one of the simplest of all statistical tests. Suppose that you cannot **measure** a difference, but you can **see** it (e.g. in judging a diving contest). For example, nine springboard divers were scored as better or worse, having trained under a new regime and under the conventional regime (the regimes were allocated in a randomized sequence to each athlete: new then conventional, or conventional then new). Divers were judged twice – one diver was worse on the new regime, and eight were better. What is the evidence that the new regime produces significantly better scores in competition? The answer comes form a two-tailed binomial test. How likely is a response of 1/9 (or 8/9 or more extreme than this, i.e. 0/9 or 9/9) if the populations are actually the same (i.e. $p = 0.5$) ? We use a binomial test for this, specifying the number of ‘failures’ (1) and the total sample size (9):

```
binom.test(1,9)
```

This produces the output

```
        Exact binomial test

data: 1 out of 9
number of successes = 1, n = 9, p-value = 0.0391
alternative hypothesis: true p is not equal to 0.5
```

from which we would conclude that the new training regime is significantly better than the traditional method, because $p < 0.05$.

It is easy to write a function to carry out a sign test to compare two samples, x and y

```
sign.test <- function(x, y)
{
if(length(x) != length(y)) stop("The two variables must be the same length")
d <- x - y
binom.test(sum(d > 0), length(d))
}
```

The function starts by checking that the two vectors are the same length, then works out the vector of the differences, d. The binomial test is then applied to the number of positive differences (sum(d > 0)) and the total number of numbers (length(d)). If there was no difference between the samples then, on average, the sum would be about half of length(d). Here is the sign test used to compare the ozone levels in gardens A and B:

```
sign.test(gardenA,gardenB)

      Exact binomial test

data: sum(d > 0) and length(d)
number of successes = 0, number of trials = 10, p-value = 0.001953
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.0000000 0.3084971
sample estimates:
probability of success
                     0
```

Note that the p value (0.002) from the sign test is larger than in the equivalent t-test ($p = 0.0011$) that we carried out earlier. This will generally be the case: other things being equal, the parametric test will be more powerful than the non-parametric equivalent.

Binomial Tests to Compare Two Proportions

Suppose that only four females were promoted compared with 196 men. Is this an example of blatant sexism, as it might appear at first glance? Before we can judge, of course, we need to know the number of male and female candidates. It turns out that 196 men were promoted out of 3270 candidates, compared with four promotions out of only 40 candidates for the women. Now, if anything, it looks like the females did better than males in the promotion round (10% success for women versus 6% success for men).

The question then arises as to whether the apparent positive discrimination in favour of women is statistically significant, or whether this sort of difference could arise through chance alone. This is easy in R using the built-in binomial proportions test prop.test in which we specify two vectors, the first containing the number of successes for females and males c(4,196) and second containing the total number of female and male candidates c(40,3270)

```
prop.test(c(4,196),c(40,3270))

2-sample test for equality of proportions with continuity correction

data: c(4, 196) out of c(40, 3270)
X-squared = 0.5229, df = 1, p-value = 0.4696
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06591631 0.14603864
sample estimates:
    prop 1      prop 2
0.10000000 0.05993884
```

There is no evidence in favour of positive discrimination ($p = 0.4696$). A result like this will occur more than 45% of the time by chance alone. Just think what would have happened if one of the successful female candidates had not applied. Then the same promotion system would have produced a female success rate of 3/39 instead of 4/40 (7.7% instead of 10%). In small samples, small changes have big effects.

Chi-square Contingency Tables

A great deal of statistical information comes in the form of **counts** (whole numbers or integers): the number of animals that died, the number of branches on a tree, the number of days of frost, the number of companies that failed, the number of patients that died. With count data, the number 0 is often the value of a response variable (consider, for example, what a 0 would mean in the context of the examples just listed).

The dictionary definition of contingency is ‘a thing dependent on an uncertain event’ (OED 2004). In statistics, however, the contingencies are **all the events that could possibly happen**. A contingency table shows the counts of how many times each of the contingencies actually happened in a particular sample. Consider the following example that has to do with the relationship between hair colour and eye colour in white people. For simplicity, we just chose two contingencies for hair colour: ‘fair’ and ‘dark’. Likewise we just chose two contingencies for eye colour: ‘blue’ and ‘brown’. These two categorical variables, eye colour and hair colour, each has two levels (‘blue’ and ‘brown’, and ‘fair’ and ‘dark’ respectively). Between them, they define four possible outcomes (the contingencies): fair hair and blue eyes, fair hair and brown eyes, dark hair and blue eyes, and dark hair and brown eyes. We take a sample of people and count how many of them fall into each of these four categories. Then we fill in the two-by-two contingency table:

	Blue eyes	Brown eyes
Fair hair	38	11
Dark hair	14	51

These are our observed frequencies (or counts). The next step is very important. In order to make any progress in the analysis of these data we need a **model** which predicts the expected frequencies. What would be a sensible model in a case like this? There are all sorts of complicated models that you might select, but the simplest model (Occam’s razor, or the Principle of Parsimony) is that hair colour and eye colour are **independent**. We may not believe that this is actually true, but the hypothesis has the great virtue of being falsifiable. It is also a very sensible model to choose because it makes it easy to predict the expected frequencies based on the assumption that the model is true. We need to do some simple probability work. What is the probability of getting a random individual from this sample whose hair was fair? A total of 49 people (38 + 11) had fair hair out of a total sample of 114 people. So the probability of fair hair is 49/114 and the probability of dark hair is 65/114. Notice that because we have only two levels of hair colour, these two probabilities add up to one [(49 + 65)/114]. What about eye colour? What is the probability of selecting someone at random from this sample with blue eyes? A total of 52 people had blue eyes (38 + 14) out of the sample of 114, so the probability of blue eyes is 52/114 and the probability of brown eyes is 62/114. As before, these add up to one [(52 + 62)/114]. It helps to add the subtotals to the margins of the contingency table like this:

	Blue eyes	Brown eyes	Row totals
Fair hair	38	11	49
Dark hair	14	51	65
Column totals	52	62	114

Now comes the important bit. We want to know the expected frequency of people with fair hair *and* blue eyes, to compare with our observed frequency of 38. Our model says that the two are independent. This is essential information, because it allows us to calculate the expected probability of fair hair and blue eyes. **If, and only if, the two traits are independent, then the probability of having fair hair and blue eyes is the product of the two probabilities**. So, following our earlier calculations, the probability of fair hair and blue eyes is $49/114 \times 52/114$. We can do exactly equivalent things for the other three cells of the contingency table:

	Blue eyes	Brown eyes	Row totals
Fair hair	$\frac{49}{114} \times \frac{52}{114}$	$\frac{49}{114} \times \frac{62}{114}$	49
Dark hair	$\frac{65}{114} \times \frac{52}{114}$	$\frac{65}{114} \times \frac{62}{114}$	65
Column totals	52	62	114

Now we need to know how to calculate the expected frequency. It couldn't be simpler. It is just the probability multiplied by the total sample ($n = 114$). So the expected frequency of blue eyes and fair hair is $\frac{49}{114} \times \frac{52}{114} \times 114 = 22.35$ which is much less than our observed frequency of 38. It is beginning to look as if our hypothesis of independence of hair and eye colour is false.

You might have noticed something useful in the last calculation: two of the sample sizes cancel out. Therefore, the expected frequency in each cell is just the row total (R) times the column total (C) divided by the grand total (G) like this:

$$E = \frac{R \times C}{G}.$$

We can now work out the four expected frequencies.

	Blue eyes	Brown eyes	Row totals
Fair hair	22.35	26.65	49
Dark hair	29.65	35.35	65
Column totals	52	62	114

Notice that the row and column totals (the so-called 'marginal totals') are retained under the model. It is clear that the observed frequencies and the expected frequencies are different, but in sampling, everything always varies, so this is no surprise. The important question is whether or not the expected frequencies are **significantly** different from the observed frequencies.

We assess the significance of the differences between the observed and expected frequencies using a Chi-square test. We calculate a test statistic χ^2 (Pearson's chi square) as follows:

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

where O is the observed frequency and E is the expected frequency. Capital Greek sigma $\sum$ just means 'add up all the values of'. It makes the calculations easier if we write the observed and expected frequencies in parallel columns, so that we can work out the corrected squared differences more easily.

	O	E	$(O\text{-}E)^2$	$\frac{(O-E)^2}{E}$
Fair hair and blue eyes	38	22.35	244.92	10.96
Fair hair and brown eyes	11	26.65	244.92	9.19
Dark hair and blue eyes	14	29.65	244.92	8.26
Dark hair and brown eyes	51	35.35	244.92	6.93

All we need to do now is to add up the four components of chi square to get $\chi^2 = 35.33$. The question now arises: is this a big value of chi square or not? This is important, because if it **is** a bigger value of chi square than we would expect by chance, then we should reject the null hypothesis. If, on the other hand, it is within the range of values that we would expect by chance alone, then we should accept the null hypothesis.

We always proceed in the same way at this stage. We have a calculated value of the test statistic: $\chi^2 = 35.33$. We compare this value of the test statistic with the relevant critical value. To work out the critical value of chi square we need two things:

- the number of degrees of freedom, and
- the degree of certainty with which to work.

In general, a contingency table has a number of rows (r) and a number of columns (c), and the degrees of freedom are given by

$$\text{d.f.} = (r - 1) \times (c - 1).$$

So we have $(2 - 1) \times (2 - 1) = 1$ degree of freedom for a 2 × 2 contingency table. You can see why there is only one degree of freedom by working through our example. Take the ‘fair hair, brown eyes’ box (the top right in the table) and ask ‘how many values could this possibly take’? The first thing to note is that the count could not be more than 49, otherwise the row total would be wrong but, in principle, the number in this box is free to be any value between 0 and 49. We have one degree of freedom for this box. But when we have fixed this box to be 11

	Blue eyes	Brown eyes	Row totals
Fair hair		11	49
Dark hair			65
Column totals	52	62	114

you will see that we have no freedom at all for any of the other three boxes. The top left box has to be $49 - 11 = 38$ because the row total is fixed at 49. Once the top left box is defined as 38 then the bottom left box has to be $52 - 38 = 14$ because the column total is fixed (the total number of people with blue eyes was 52). This means that the bottom right box has to be $65 - 14 = 51$. Thus, because the marginal totals are constrained, a 2 × 2 contingency table has just one degree of freedom.

The next thing we need to do is say how certain we want to be about the falseness of the null hypothesis. The more certain we want to be, the larger the value of chi square we would need to reject the null hypothesis. It is conventional to work at the 95% level. That is our certainty level, so our uncertainty level is $100 - 95 = 5\%$. Expressed as a fraction, this is called alpha ($\alpha = 0.05$). Technically, alpha is the probability of **rejecting** the null hypothesis when it is **true**. This is called a Type I error. A Type II error is **accepting** the null hypothesis when it is **false**.

Critical values in R are obtained by use of *quantiles* (q) of the appropriate statistical distribution. For the chi-squared distribution, this function is called qchisq. The function has two arguments: the certainty level ($p = 0.95$), and the degrees of freedom (d.f. = 1):

```
qchisq(0.95,1)

[ 1] 3.841459
```

The critical value of chi squared is 3.841.The logic goes like this: since the calculated value of the test statistic is **greater** than the critical value we **reject** the null hypothesis. You should memorize this sentence and put the emphasis on ‘greater’ and ‘reject’.

What have we learned so far? We have rejected the null hypothesis that eye colour and hair colour are independent. However, that's not the end of the story, because we have not established the **way** in which they are related (e.g. is the correlation between them positive or negative?). To do this we need to look carefully at the data and compare the observed and expected frequencies. If fair hair and blue eyes were positively correlated, would the observed frequency be greater or less than the expected frequency? A moment's thought should convince you that the observed frequency will be greater than the expected frequency when the traits are positively correlated (and less when they are negatively correlated). In our case we expected only 22.35 but we observed 38 people (nearly twice as many) to have both fair hair and blue eyes. So it is clear that fair hair and blue eyes are **positively** associated.

In R the procedure is very straightforward. We start by defining the counts as a 2 × 2 matrix like this:

```
count < -matrix(c(38,14,11,51),nrow = 2)
count
```

```
      [ ,1]    [ ,2]
[ 1,]   38       11
[ 2,]   14       51
```

Notice that you enter the data **column-wise** (not row-wise) into the matrix. Then the test uses the chisq.test function, with the matrix of counts as its only argument.

```
chisq.test(count)
```

```
        Pearson' s Chi-squared test with Yates' continuity correction

data: count
X-squared = 33.112, df = 1, p-value = 8.7e-09
```

The calculated value of chi square is slightly different from ours, because Yates' correction has been applied as the default (see Sokal and Rohlf 1995: p. 736). If you switch the correction off (correct = F), you get the value we calculated by hand:

```
chisq.test(count,correct = F)
```

```
        Pearson' s Chi-squared test

data: count
X-squared = 35.3338, df = 1, p-value = 2.778e-09
```

It makes no difference at all to the interpretation that there is a highly significant positive association between fair hair and blue eyes for this group of people.

Fisher's Exact Test

This test is used for the analysis of contingency tables in which **one or more of the expected frequencies is less than 5**. The individual counts are *a*, *b*, *c* and *d*:

2×2 Table	Column 1	Column 2	Row totals
Row 1	a	b	$a+b$
Row 2	c	d	$c+d$
Column totals	$a+c$	$b+d$	n

The probability of any one particular outcome is given by

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!},$$

where n is the grand total, and ! means 'factorial' (the product of all the numbers from n down to 1; zero! is defined as being 1).

Our data concern the distribution of eight ants' nests over ten trees of each of two species (A and B). There are two categorical explanatory variables (ants and trees), and four contingencies, ants (present or absent) and trees (A or B). The response variable (shaded cells) is the vector of four counts c(6,4,2,8).

	Tree A	Tree B	Row totals
With ants	6	2	8
Without ants	4	8	12
Column totals	10	10	20

R does not have a function to calculate factorials, but we can easily write one based on the maximum value of the cumulative product of the numbers from 1 to *x*:

```
factorial<-function(x) max(cumprod(1:x))
```

Now we can calculate the probability for this particular outcome:

```
factorial(8)*factorial(12)*factorial(10)*factorial(10)/(factorial(6)
            *factorial(2)*factorial(4)*factorial(8)*factorial(20))
```

```
[ 1] 0.07501786
```

This is only part of the story. We need to compute the probability of outcomes that are **more extreme** than this. There are two of them. Suppose only one ant colony was found on Tree B. Then the table values would be 7, 1, 3, 9 but the row and column totals would be exactly the same (the marginal totals are constrained). The numerator always stays the same, so this case has probability

```
factorial(8)*factorial(12)*factorial(10)*factorial(10)/
        (factorial(7)*factorial(3)*factorial(1)*factorial(9)*factorial(20))
```

```
[ 1] 0.009526078
```

There is an even more extreme case if no ant colonies at all were found on Tree B. Now the table elements become 8, 0, 2, 10 with probability

```
factorial(8)*factorial(12)*factorial(10)*factorial(10)/
        (factorial(8)*factorial(2)*factorial(0)*factorial(10)*factorial(20))
```

```
[ 1] 0.0003572279
```

and we need to add these three probabilities together

```
0.07501786 + 0.009526078 + 0.000352279
```

```
[ 1] 0.08489622
```

However, there was no *a priori* reason for expecting the result to be in this direction. It might have been Tree A that had relatively few ant colonies. We need to allow for extreme counts in the opposite direction by doubling this probability (all Fisher's Exact Tests are two-tailed).

```
2*(0.07501786 + 0.009526078 + 0.000352279)
```

```
[ 1] 0.1697924
```

This shows that there is no evidence of a correlation between tree and ant colonies. The observed pattern, or a more extreme one, could have arisen by chance alone with probability $p = 0.17$.

There is a built-in function called fisher.test, which saves us all this tedious computation. It takes as its argument a 2×2 matrix containing the counts of the four contingencies. We make the matrix like this (compare with the alternative method of making a matrix, above):

```
x < -as.matrix(c(6,4,2,8))
dim(x) < -c(2,2)
x
```

```
          [ ,1]  [ ,2]
[ 1,]       6      2
[ 2,]       4      8
```

and run the test like this

```
fisher.test(x)
```

```
        Fisher' s Exact Test for Count Data

data: x
p-value = 0.1698
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.6026805 79.8309210
sample estimates:
odds ratio
  5.430473
```

The fisher.test can be used with matrices much bigger than 2 × 2. Alternatively, the function may be provided with two vectors containing factor levels, instead of a two-dimensional matrix of counts, as here; this saves you the trouble of counting up how many combinations of each factor level there are:

```
table <-read.table("c:\\temp\\fisher.txt",header=T)
table
```

```
     tree    nests
1       A     ants
2       B     ants
3       A     none
4       A     ants
5       B     none
6       A     none
7       A     ants
8       B     ants
9       B     none
10      A     none
11      A     none
12      B     none
13      B     none
14      A     ants
15      A     ants
16      B     none
17      A     ants
18      B     none
19      B     none
20      B     none
```

```
attach(table)
fisher.test(tree,nests)
```

```
      Fisher′ s Exact Test for Count Data
data: tree and nests
p-value = 0.1698
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.6026805 79.8309210
sample estimates:
odds ratio
  5.430473
```

Correlation and Covariance

With two continuous variables, *x* and *y*, the question naturally arises as to whether their values are correlated with each other. Correlation is defined in terms of the variance of *x*, the variance of *y*, and the covariance of *x* and *y* (the way the two vary together, or the way they co-vary) on the assumption that both variables are normally distributed. We have symbols already for the two variances; s_x^2 and s_y^2. Now we call the covariance of *x* and *y* cov(*x,y*), after which the correlation coefficient *r* is defined as

$$r = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 . s_y^2}}.$$

We know how to calculate variances, so it remains only to work out the value of the covariance of *x* and *y*. Covariance is defined as **the expectation of the vector product** $x * y$ which sounds difficult, but isn't (Box 6.2). The covariance of *x* and *y* is 'the expectation of the product minus the product of the two expectations'. Note that when *x* and *y* are independent (i.e. they are not correlated) then the covariance between *x* and *y* is 0, so $\mathbf{E}[xy] = \mathbf{E}[x].\mathbf{E}[y]$ (i.e. the product of their mean values).

Box 6.2 Correlation and covariance

The correlation coefficient is defined in terms of the covariance of *x* and *y,* and the geometric mean of the variances of *x* and *y*:

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \times \text{var}(y)}}.$$

We know how to compute var(*x*) and var(*y*), so we need only to find cov(*x,y*). The covariance of *x* and *y* is defined as the **expectation** of the vector product $(x - \bar{x})(y - \bar{y})$

$$\text{cov}(x, y) = \mathbf{E}[(x - \bar{x})(y - \bar{y})].$$

We start by multiplying through the brackets:

$$(x - \bar{x})(y - \bar{y}) = xy - \bar{x}y - x\bar{y} + \overline{xy}.$$

Now applying expectations, and remembering that the expectation of x is $\bar{x}$ and the expectation of y is $\bar{y}$ we get

$$\mathrm{cov}(x,y) = \mathbf{E}(xy) - \bar{x}\mathbf{E}(y) - \mathbf{E}(x)\bar{y} + \overline{xy} = \mathbf{E}(xy) - \overline{xy} - \overline{xy} + \overline{xy}.$$

Then $-\overline{xy} + \overline{xy}$ cancels out, leaving $-\overline{xy}$ which is $-\mathbf{E}(x)\mathbf{E}(y)$ so

$$\mathrm{cov}(x,y) = \mathbf{E}(xy) - \mathbf{E}(x)\mathbf{E}(y).$$

Notice that when x and y are uncorrelated, $\mathbf{E}(xy) = \mathbf{E}(x)\mathbf{E}(y)$ so the covariance is 0 in this case. The corrected sum of products *SSXY* (see p. 133) is given by

$$SSXY = \sum xy - \frac{\sum x \sum y}{n},$$

so covariance is computed as:

$$\mathrm{cov}(x,y) = SSXY\sqrt{\frac{1}{(n-1)^2}}$$

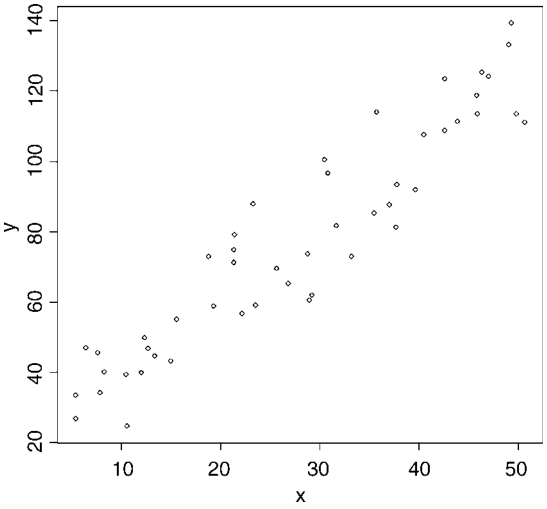
SSXY also provides a shortcut formula for the correlation coefficient

$$r = \frac{SSXY}{\sqrt{SSX.SSY}}$$

because the degrees of freedom $(n-1)$ cancel out. The sign of r takes the sign of *SSXY*: positive for positive correlations and negative for negative correlations.

Let's do a numerical example.

```
data<-read.table("c:\\temp\\twosample.txt",header=T)
attach(data)
plot(x,y)
```



First we need the variance of *x* and the variance of *y*:

```
var(x)
```

```
[ 1] 199.9837
```

```
var(y)
```

```
[ 1] 977.0153
```

The covariance of *x* and *y*, cov(*x*,*y*), is given by the var function when we supply it with two vectors like this:

```
var(x,y)
```

```
[ 1] 414.9603
```

Thus, the correlation coefficient should be $414.96/\sqrt{199.98 \times 977.02}$

```
var(x,y)/sqrt(var(x)*var(y))
```

```
[ 1] 0.9387684
```

Let's see if this checks out:

```
cor(x,y)
```

```
[ 1] 0.9387684
```

Yes it does! So now you know the definition of the correlation coefficient: it is the covariance divided by the geometric mean of the two variances.

Data Dredging

The R function cor returns the correlation matrix of a data matrix, or a single value showing the correlation between one vector and another.

```
pollute < -read.table("c:\\temp\\pollute.txt",header = T)
attach(pollute)
names(pollute)
```

```
[ 1]  "Pollution"  "Temp"     "Industry"  "Population"  "Wind"
[ 6]  "Rain"       "Wet.days"
```

```
cor(pollute)
```

```
              Pollution        Temp     Industry  Population         Wind
Pollution    1.00000000 −0.43360020   0.64516550  0.49377958   0.09509921
Temp        −0.43360020  1.00000000  −0.18788200 −0.06267813  −0.35112340
Industry     0.64516550 −0.18788200   1.00000000  0.95545769   0.23650590
```

```
Population  0.49377958  −0.06267813   0.95545769   1.00000000   0.21177156
Wind        0.09509921  −0.35112340   0.23650590   0.21177156   1.00000000
Rain        0.05428389   0.38628047  −0.03121727  −0.02606884  −0.01246601
Wet.days    0.36956363  −0.43024212   0.13073780   0.04208319   0.16694974
                  Rain     Wet.days
Pollution   0.05428389   0.36956363
Temp        0.38628047  −0.43024212
Industry   −0.03121727   0.13073780
Population −0.02606884   0.04208319
Wind       −0.01246601   0.16694974
Rain        1.00000000   0.49605834
Wet.days    0.49605834   1.00000000
```

The phrase ‘data dredging’ is used disparagingly to describe the act of trawling through a table like this, desperately looking for big values which might suggest relationships that you can publish. This behaviour is not to be encouraged. The correct approach is model simplification (see p. 195). Note that the correlations are identical in opposite halves of the matrix (in contrast to regression, where regression of *y* on *x* would be different from a regression of *x* on *y*). The correlation between two vectors produces a single value:

```
cor(Pollution,Wet.days)
```

```
[ 1] 0.3695636
```

Correlations with single explanatory variables can be highly misleading if (as is typical) there is substantial correlation amongst the explanatory variables (see Chapter 11).

Partial Correlation

With more than two variables, you often want to know the correlation between *x* and *y* when a third variable, say *z*, is held constant. The **partial correlation coefficient** measures this. It enables correlation due to a shared common cause to be distinguished from direct correlation:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}.r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

Suppose we had four variables and we wanted to look at the correlation between *x* and *y* holding the other two, *z* and *w*, constant

$$r_{xy.zw} = \frac{r_{xy.z} - r_{xw.z}.r_{yw.z}}{\sqrt{(1 - r_{xw.z}^2)(1 - r_{yw.z}^2)}}.$$

You will need partial correlation coefficients (p.c.c.) if you want to do **path analysis**. In this book, we prefer to use tree models and various kinds of model simplification following multiple regression. Nevertheless, if you need them, you can use the built-in function lm to get the values of partial correlation coefficients as follows. The sum of squares attributable to a given variable can be determined by deleting it from a model containing all the other variables, using update with anova. Divide this sum of squares by *SSY* and you get what you might call a partial r^2. Take the square root of this to get a partial correlation coefficient.

Correlation and the Variance of Differences Between Variables

Samples often exhibit positive correlations that result from the pairing, as in the upstream and downstream invertebrate biodiversity data that we investigated earlier. There is an important general question about the effect of correlation on the variance of differences between variables. In the extreme, when two variables are so perfectly correlated that they are identical, then the difference between one variable and the other is zero. So it is clear that the variance of a difference will decline as the strength of positive correlation increases.

The following data show the depth of the water table (m below the surface) in winter and summer at nine locations:

```
paired < -read.table("c:\\temp\\paired.txt",header = T)
attach(paired)
names(paired)
```

```
[ 1] "Location" "Summer" "Winter"
```

We begin by asking whether there is a correlation between summer and winter water table depths across locations:

```
cor(Summer, Winter)
```

```
[ 1] 0.8820102
```

There is a strong positive correlation. Not surprisingly, places where the water table is high in summer tend to have a high water table in winter as well. If you want to determine the significance of a correlation (i.e. the *p* value associated with the calculated value of *r*) then use cor.test rather than cor. This test has non-parametric options for Kendall's tau or Spearman's rank depending on the method you specify (method = "k" or method = "s"), but the default method is Pearson's product–moment correlation (method = "p"):

```
cor.test(Summer, Winter)
```

```
      Pearson' s product-moment correlation
data: Summer and Winter
t = 4.9521, df = 7, p-value = 0.001652
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5259984 0.9750087
sample estimates:
      cor
0.8820102
```

The correlation is highly significant ($p = 0.00165$). Now, let's investigate the relationship between the correlation coefficient and the three variances: the summer variance, the winter variance, and **the variance of the differences** (summer–winter)

```
varS = var(Summer)
varW = var(Winter)
varD = var(Summer-Winter)
```

The correlation coefficient ρ is related to these three variances by:

$$\rho = \frac{\sigma_y^2 + \sigma_z^2 - \sigma_{y-z}^2}{2\sigma_y\sigma_z}.$$

So, using the values we have just calculated, we get the correlation coefficient to be

```
(varS + varW-varD)/(2*sqrt(varS)*sqrt(varW))
```

```
[ 1] 0.8820102
```

which checks out. We can also see whether the variance of the difference is equal to the sum of the component variances (see p. 76):

```
varD
```

```
[ 1] 0.01015
```

```
varS + varW
```

```
[ 1] 0.07821389
```

No, it is not. They would be equal only if the two samples were independent. In fact, we know that the two variables are positively correlated, so the variance of the difference should be **less** than the sum of the variances by an amount equal to $2 \times r \times s_1 \times s_2$

```
varS + varW – 2 * 0.8820102 * sqrt(varS) * sqrt(varW)
```

```
[ 1] 0.01015
```

which is a better result.

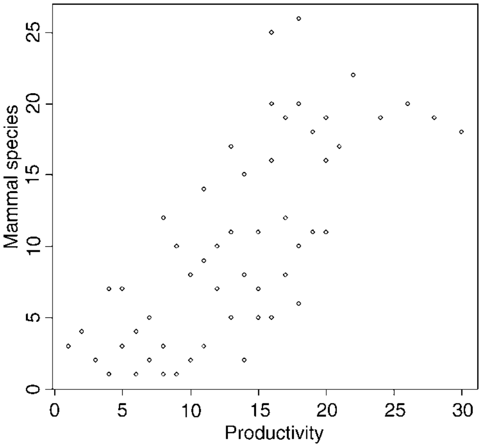
Scale-dependent Correlations

Another major difficulty with correlations is that scatterplots can give a highly misleading impression of what is going on. The moral of this exercise is very important: **things are not always as they seem**. The data show the number of species of mammals in forests of differing productivity:

```
par(mfrow = c(1,1))
rm(x,y)
productivity < -read.table("c:\\temp\\productivity.txt",header = T)
attach(productivity)
names(productivity)
```

```
[ 1] "x" "y" "f"
```

```
plot(x,y,ylab = "Mammal species",xlab = "Productivity")
```

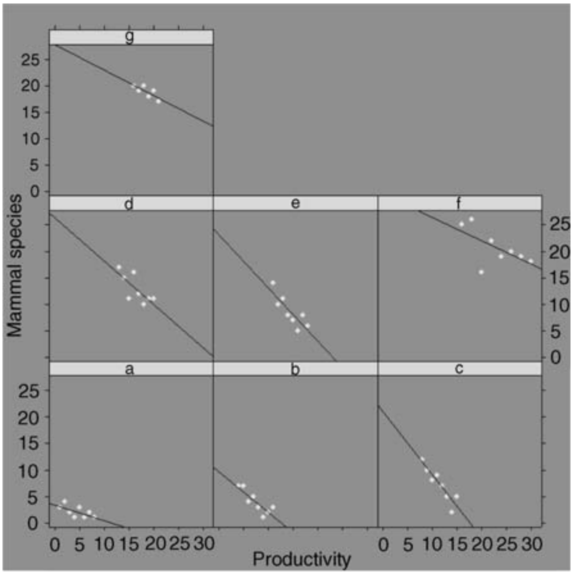


There is a very clear positive correlation: increasing productivity is associated with increasing species richness. The correlation is highly significant:

cor.test(x,y,method = "spearman")

```
        Spearman' s rank correlation rho
data: x and y
S = 6515, p-value = < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7516389
```

However, what if we look at the relationship for each region separately, using xyplot from the library of lattice plots (see the web site)?



I’ve added the regression lines for emphasis, but the pattern is obvious. In every single case, increasing productivity is associated with **reduced** mammal species richness within each region (labelled a–g). The lesson is clear: you need to be extremely careful when looking at **correlations across different scales**. Things that are positively correlated over short time scales may turn out to be negatively correlated in the long term. Things that appear to be positively correlated at large spatial scales may turn out (as in this example) to be negatively correlated at small scales.

Kolmogorov–Smirnov Test

People know this test for its wonderful name, rather than for what it actually does. It is an extremely simple test for asking one of two different questions.

- Are two sample distributions the same, or are they significantly different from one another?
- Does a particular sample distribution arise from a particular hypothesized distribution?

The two-sample problem is the one most often used. The apparently simple question is actually very broad. It is obvious that two distributions could be different because their means were different; but two distributions with exactly the same mean could be significantly different if they differed in variance, or in skew or kurtosis (see p. 69). The Kolmogorov–Smirnov test works on **cumulative distribution functions** (**c**.**d**.**f**.). These give the probability that a randomly selected value of $X \leq x$

$$F(x) = P[X \leq x]$$

This sounds somewhat abstract. Suppose we had insect wing sizes for two geographically separated populations and we wanted to test whether the distribution of wing lengths was the same in the two places.

```
wings < -read.table("c:\\temp\\wings.txt",header = T)
attach(wings)
names(wings)
```

```
[ 1] "size" "location"
```

We need to find out how many specimens there are from each location:

```
table(location)
location
 A   B
50  70
```

So the samples are of unequal size (50 insects from location A, 70 from B). It will be useful, therefore, to create two separate variables containing the wing lengths from sites A and B:

```
A < -size[location = = "A"]
B < -size[location = = "B"]
```

We could begin by comparing mean wing length in the two locations with a *t*-test:

```
t.test(A,B)
```

```
      Welch Two Sample t-test

data: A and B
t = -1.6073, df = 117.996, p-value = 0.1107
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -2.494476 0.259348
sample estimates:
mean of x mean of y
 24.11748 25.23504
```

This shows that mean wing length is not significantly different in the two locations ($p = 0.11$); but what about other attributes of the distribution? This is where Kolmogorov–Smirnov is really useful:

```
ks.test(A,B)
```

```
      Two-sample Kolmogorov-Smirnov test

data: A and B
D = 0.2629, p-value = 0.02911
alternative hypothesis: two.sided
```

The two distributions are, indeed, significantly different from one another ($p < 0.05$); but if not in their means, then in what respect do they differ? Perhaps they have different variances?

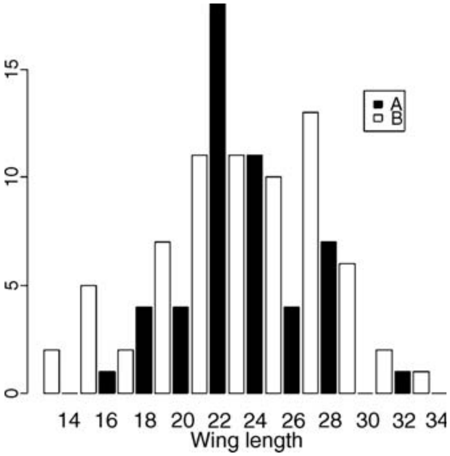
```
var.test(A,B)
```

```
      F test to compare two variances

data: A and B
F = 0.5014, num df = 49, denom df = 69, p-value = 0.01192
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.3006728 0.8559914
sample estimates:
ratio of variances
         0.5014108
```

Indeed they do: the variance of wing length from location B is double that from location A ($p < 0.02$).

We can finish by drawing the two histograms side by side to get a visual impression of the difference in the shape of the two distributions; the open bars show the data from location B, solid bars show location A (see the web site).



The spread of wing lengths is much greater at location B despite the fact that the mean wing length is similar in the two places. Also, the distribution is skew to the left in location B, with the result that modal wing length is greater in location B (26 mm compared with 22 mm).