4

Variance

A measure of variability is perhaps the most important quantity in statistical analysis. The greater the variability in the data, the greater will be our uncertainty in the values of parameters estimated from the data, and the lower will be our ability to distinguish between competing hypotheses about the data.

Consider the following data, *y*, which are plotted simply in the order in which they were measured:

```
y<-c(13,7,5,12,9,15,6,11,9,7,12)
plot(y,ylim=c(0,20))
```

How can we quantify the variation (the scatter) in *y* that we can see here? Perhaps the simplest measure is the `range` of *y* values (p. 34):

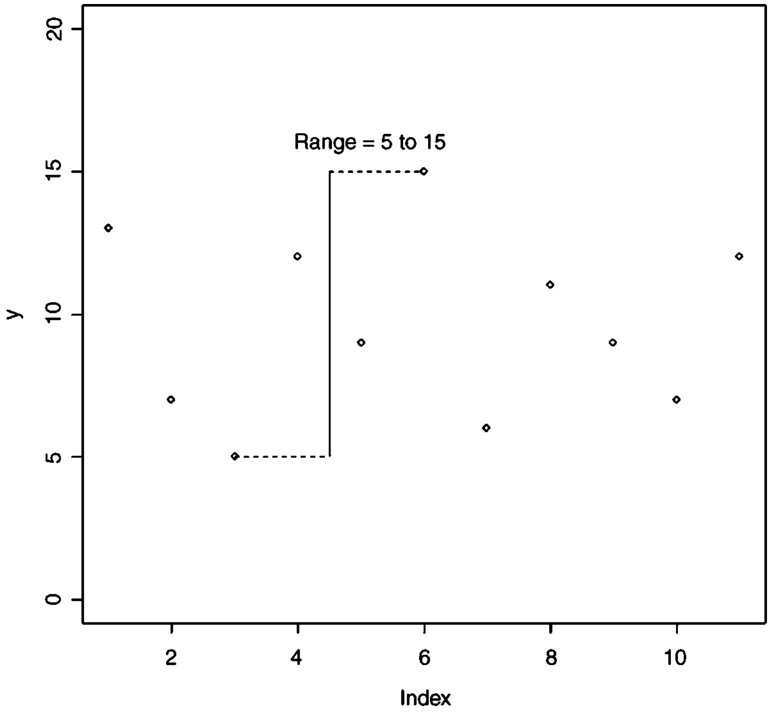
```
range(y)
```

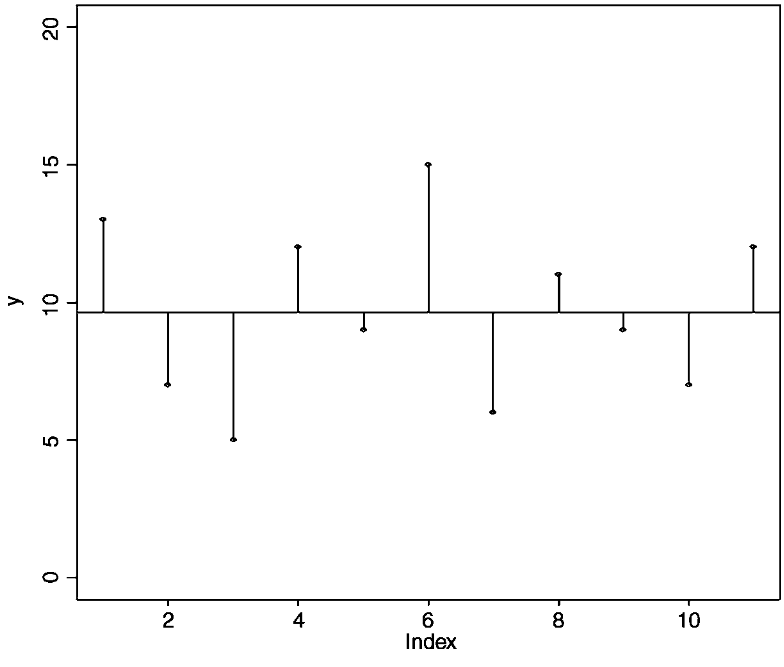
```
[ 1] 5  15
```

This is a reasonable measure of variability, but it is too dependent on outlying values for most purposes. Also, we want all of our data to contribute to the measure of variability, not just the maximum and minimum values. How about estimating the mean value, and looking at the departures from the mean (known as 'residuals' or 'deviations')?

The longer these lines, the more variable the data. So this looks promising. How about adding up the lengths of the lines: $\sum (y - \bar{y})$? A moment's thought will show that this is no good, because the negative residuals (from the points below the mean) will cancel out the positive residuals (from the points above the line). In fact, it is easy to prove that this quantity $\sum (y - \bar{y})$ is zero, no matter what the variation in the data, so that's no good (see Box 4.1).

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


Range = 5 to 15
y
0
5
10
15
20
2
4
6
8
10
Index


y
0
5
10
15
20
2
4
6
8
10
Index

Box 4.1. The sum of the differences $\sum (y - y)$ is zero

Start by writing down the differences explicitly

$$\sum d = \sum (y - \bar{y}).$$

Take $\sum$ through the brackets. The important point is that $\sum \bar{y}$ is the same as $n.\bar{y}$ so

$$\sum d = \sum y - n\bar{y},$$

and we know that $\bar{y} = \sum y/n$ so

$$\sum d = \sum y - \frac{n \sum y}{n}.$$

The n's cancel, leaving

$$\sum d = \sum y - \sum y = 0.$$

The only problem is the minus signs. How about ignoring the minus signs and adding up the absolute values of the residuals: $\sum (|y - \bar{y}|)$. This is a very good measure of variability, and is used in some modern, computationally intensive methods. The problem is that it makes the sums hard, and we don't want that. A much more straightforward way of getting rid of the problem of the minus signs is to square the residuals before we add them up: $\sum (y - \bar{y})^2$. This is perhaps the most important single quantity in all of statistics. It is called, somewhat unimaginatively, the **sum of squares**. So, in our figure (p. 34) imagine squaring the lengths of each of the vertical lines:

```
y-mean(y)
```

```
[ 1] 3.3636364 -2.6363636  -4.6363636  2.3636364  -0.6363636  5.3636364 -3.6363636  1.3636364
[ 9]-0.6363636 -2.6363636   2.3636364
```

```
(y-mean(y))^2
```

```
[ 1]11.3140496  6.9504132  21.4958678  5.5867769  0.4049587 28.7685950 13.2231405  1.8595041
[ 9] 0.4049587  6.9504132   5.5867769
```

then adding up all these squared differences:

```
sum((y-mean(y))^2)
[ 1] 102.5455
```

So the sum of squares for our data is 102.5455. But what are its units? Well that depends on the units in which *y* is measured. Suppose the *y* values were lengths in mm. So the units of the sum of squares are mm^2 (like an area).

Now what would happen to the sum of squares if we added a twelfth data point? It would get bigger, of course. And it would get bigger for every extra data point we added (except in the unlikely event that our new data point was exactly equal to the mean value, in which case we would add zero squared = 0). We don't want our measure of variability to depend on sample size in this way, so the obvious solution is to divide by the number of samples, to get the **mean squared deviation**.

At this point we need to make a brief, but important, diversion. Before we can progress, we need to understand the concept of degrees of freedom.

Degrees of Freedom

Suppose we had a sample of five numbers and their average was 4. What was the sum of the five numbers? It must have been 20, otherwise the mean would not have been 4. So now let's think about each of the five numbers in turn.

We are going to put a number in each of the five boxes. If we allow that the numbers could be positive or negative real numbers, we ask how many values could the first number take. Once you see what I'm doing, you will realize it could take any value. Suppose it was a 2.

2				

How many values could the next number take? It could be anything. Say it was a 7.

2	7			

And the third number could be anything. Suppose it was a 4.

2	7	4		

The fourth number could be anything at all. Say it was 0.

2	7	4	0	

Now, how many values could the last number take? Just one – it **has** to be another 7 because the numbers have to add up to 20 because the mean of the five numbers is 4.

2	7	4	0	7

To recap, we have total freedom in selecting the first number–and the second, third and fourth numbers. However, we have no choice at all in selecting the fifth number. We have four degrees of freedom when we have five numbers. In general we have $(n-1)$ degrees of freedom if we estimated the mean from a sample of size n. More generally still, we can propose a formal definition of degrees of freedom: **degrees of freedom is the sample size, n, minus the number of parameters, p, estimated from the data**. This is so important, you should memorize it. In the example we plotted earlier we had $n = 11$ and we estimated just one parameter from the data, the sample mean, $\bar{y}$. So we have $n - 1 = 10$ d.f.

Variance

We return to developing our quantitative measure of variability. We have come to the conclusion that the sum of squares $\sum (y-\bar{y})^2$ is a good basis for assessing variability, but we have the problem that the sum of squares increases with every new data point we add to the sample. The intuitive thing to do would be to divide by the number of numbers, n, to get the mean squared deviation, but look at the formula for the sum of squares: $\sum (y-\bar{y})^2$. We cannot begin to calculate it until we know the value of the sample mean, $\bar{y}$, and where do we get the value of $\bar{y}$ from? Do we know it in advance? Can we look it up in tables? No, we need to calculate it from the data. The mean value $\bar{y}$ is **a parameter estimated from the data**, so we loose one degree of freedom as a result. Thus, in calculating the mean squared deviation we divide by the degrees of freedom, $n - 1$, rather than by the sample size, n. In the jargon, this provides us with an unbiased estimate of the variance, because we have taken account of the fact that one parameter was estimated from the data prior to computation.

Now we can formalize our definition of the measure that we shall use throughout the book for quantifying variability. It is called **variance** and it is represented conventionally by s^2:

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}.$$

This is one of the most important definitions in the book, and you should commit it to memory. We can put it into a more mathematical form, by spelling out what we mean by each of the phrases in the numerator and the denominator:

$$\text{variance} = s^2 = \frac{\sum (y-\bar{y})^2}{n-1}.$$

Let's write an R function to do this. We have most of the necessary components already (see above); the sum of squares is obtained as `sum((y-mean(y))^2)`. For the degrees of freedom, we need to know the number of numbers in the vector, y. This is obtained by the function `length(y)`. Let's call the function variance and write it like this:

```
variance <- function (x) sum((x-mean(x))^2)/(length(x)-1)
```

Now we can try out the function on our data, like this

```
variance(y)
```

```
[ 1] 10.25455
```

So there we have it. Our quantification of the variation we saw in the first plot is the sample variance, $s^2 = 10.25455$. You will not be surprised that R provides its own, built-in function for calculating variance, and it has an even simpler name than the function we just wrote: var

```
var(y)
```

```
[ 1] 10.25455
```

Variance is used in countless ways in statistical analysis, so this section is probably the most important section in the whole book, and you should re-read it until you are sure that you know exactly what variance is, and precisely what it measures (Box 4.2).

Box 4.2. Short-cut formula for the sum of squares $\sum (y - \bar{y})^2$

The main problem with the formula defining variance is that it involves all those subtractions, $y - \bar{y}$. It would be good to find a simpler way of calculating the sum of squares. Let's expand the bracketed term $(y - \bar{y})^2$ to see if we can make any progress towards a subtraction-free solution:

$$(y - \bar{y})^2 = (y - \bar{y})(y - \bar{y}) = y^2 - 2y\bar{y} + \bar{y}^2.$$

So far, so good. Now we apply the summation

$$\sum y^2 - 2\bar{y}\sum y + n\bar{y}^2 = \sum y^2 - 2\frac{\sum y}{n}\sum y + n\left[\frac{\sum y}{n}\right]^2.$$

Note that only the y's take the summation sign. This is because we can replace $\sum \bar{y}$ by $n\bar{y}$. Now replace $\bar{y}$ with $\sum y/n$ on the right-hand side, then cancel the n's and collect the terms:

$$\sum y^2 - 2\frac{[\sum y]^2}{n} + n\frac{[\sum y]^2}{n^2} = \sum y^2 - \frac{[\sum y]^2}{n}.$$

This is the short-cut formula for computing the sum of squares. It requires only two quantities to be estimated from the data: the sum of the squared y values $\sum y^2$ and the square of the sum of the y values $[\sum y]^2$.

A Worked Example

The data in the following table come from three market gardens. The data show the ozone concentrations in parts per hundred million (pphm) on ten summer days.

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

```
      gardenA   gardenB   gardenC
1           3         5         3
2           4         5         3
3           4         6         2
4           3         7         1
5           2         4        10
6           3         4         4
7           1         3         3
8           3         5        11
9           5         6         3
10          2         5        10
```

The first step in calculating variance is to work out the mean:

```
mean(gardenA)
```

```
[1] 3
```

Now we subtract the mean value (3) from each of the data points:

```
gardenA-mean(gardenA)
```

```
[1] 0  1  1  0 -1  0 -2  0  2 -1
```

This produces a vector of differences (of length = 10). We need to square these differences:

```
(gardenA-mean(gardenA))^2
```

```
[1] 0  1  1  0  1  0  4  0  4  1
```

then add up the squared differences:

```
sum((gardenA-mean(gardenA))^2)
```

```
[1] 12
```

This important quantity is called ‘the sum of squares’. Variance is the sum of squares divided by degrees of freedom. We have ten numbers, and have estimated one parameter from the data (the mean) in calculating the sum of squares, so we have $10 - 1 = 9$ d.f.

```
sum((gardenA-mean(gardenA))^2)/9
```

```
[ 1] 1.333333
```

So the mean ozone concentration in garden A is 3.0 and the variance in ozone concentration is 1.33. We now do the same for garden B:

```
mean(gardenB)
```

```
[ 1] 5
```

It has a much higher mean ozone concentration than garden A, but what about its variance?

```
gardenB-mean(gardenB)
```

```
[ 1] 0  0  1  2 -1 -1 -2  0  1  0
```

```
(gardenB-mean(gardenB))^2
```

```
[ 1] 0  0  1  4  1  1  4  0  1  0
```

```
sum((gardenB-mean(gardenB))^2)
```

```
[ 1] 12
```

```
sum((gardenB-mean(gardenB))^2)/9
```

```
[ 1] 1.333333
```

This is interesting: although the mean values are quite different, the variances are exactly the same (both have $s^2 = 1.33333$). What about garden C?

```
mean(gardenC)
```

```
[ 1] 5
```

Its mean ozone concentration is the same as in garden B.

```
gardenC-mean(gardenC)
```

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

```
(gardenC-mean(gardenC))^2
```

```
[ 1] 4  4  9  16  25  1  4  36  4  25
```

```
sum((gardenC-mean(gardenC))^2)
```

```
[ 1] 128
```

```
sum((gardenC-mean(gardenC))^2)/9
```

```
[ 1] 14.22222
```

So, although the means in gardens B and C are identical, the variances are quite different (1.33 and 14.22 respectively). Are the variances significantly different? We do an *F*-test for this, dividing the larger variance by the smaller variance:

```
var(gardenC)/var(gardenB)
```

```
[ 1] 10.66667
```

Then look up the probability of getting an *F*-ratio as big as this by chance alone if the two variances were really the same. We need the cumulative probability of the *F*-distribution, which is a function called `pf` that we need to supply with three **arguments**: the size of the variance ratio (10.667), the number of degrees of freedom in the numerator (9) and the number of degrees of freedom in the denominator (also 9). We did not know in advance which garden was going to have the higher variance, so we do what's called a two-tail test (we simply multiply the probability by 2):

```
2*(1-pf(10.667,9,9))
```

```
[ 1] 0.001624002
```

This probability is much less than 5%, so we conclude that there is a highly significant difference between these two variances. We could do this even more simply by using the built-in *F* test:

```
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
```

So the two variances are significantly different, but why does this matter?

What follows is one of the most important lessons so far, so keep re-reading it until you are sure that you understand it. Comparing gardens A and B we can see that two samples can have different means, but the same variance. This is assumed to be the case when we carry out standard tests (like Student's *t*-test) to compare two means, or an analysis of variance to compare three or more means.

Comparing gardens B and C we can see that two samples can have the same mean but different variances. Is it right to say samples with the same mean are identical? No! Let's look into the science in a bit more detail. The damage threshold for lettuces is 8 pphm ozone, so looking at the means shows that both gardens are free of ozone damage on their lettuces (the mean of 5 for both B and C is well below the threshold of 8). Let's look at the raw data for garden B. How many of the days had ozone > 8? Look at the dataframe and you will see that none of the days exceeded the threshold. What about garden C?

```
gardenC
```

```
[ 1] 3  3  2  1 10  4  3 11  3 10
```

In garden C ozone reached damaging concentrations on three days out of ten, so 30% of the time the lettuce plants would be suffering ozone damage. This is the key point: when the variances are different, we should not make inferences by comparing the means. When we compare the means, we conclude that garden C is like garden B, and that there will be no ozone damage to the lettuces. When we look at the data, we see that this is completely wrong: there is ozone damage 30% of the time in garden C and none of the time in garden B.

So, **when the variances are different, don't compare the means**. If you do, you run the risk of coming to entirely the wrong conclusion.

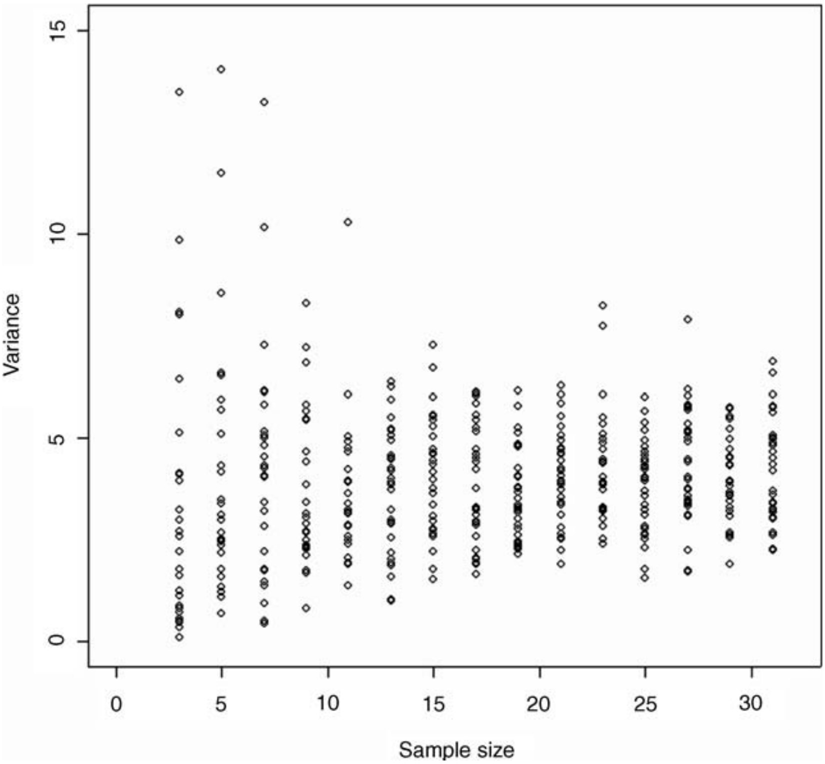
Variance and Sample Size

It is important to understand the relationship between the size of a sample (the replication, *n*) and the value of variance that is estimated. We can do a simple simulation experiment to investigate this:

```
plot(c(0,32),c(0,15),type="n",xlab="Sample size",ylab="Variance")
```

The plan is to select random numbers from a normal distribution using the function rnorm. The distribution is defined as having a mean of 10 and a standard deviation of 2 (this is the square root of the variance, so $s^2 = 4$). We shall work out the variance for sample sizes between $n = 3$ and $n = 31$, and plot 30 independent instances of variance at each of the selected sample sizes:

```
for (df in seq(3,31,2)) {
for( i in 1:30){
x<-rnorm(df,mean=10,sd=2)
points(df,var(x)) }}
```



You see that as sample size declines, the range of the estimates of sample variance increases dramatically (remember that the population variance is constant at $s^2 = 4$ throughout). The problem becomes severe below samples of 13 or so, and is very serious for samples of seven or fewer. Even for reasonably large samples (like $n = 31$) the variance varies more than three-fold in just 30 trials (you can see that the rightmost group of points varies from about 2 to about 6). This means that for small samples, the estimated variance is badly behaved, and this has serious consequences for estimation and hypothesis testing.

When people ask 'how many samples do I need?' a statistician will often answer with another question: 'how many can you afford?' Other things being equal, what we have learned in this chapter is that 30 is a reasonably good sample. Anything less than this is a small sample, and anything less than 10 is a very small sample. Anything more than 30 may be an unnecessary luxury (i.e. a waste of resources). We shall see later when we study Power Analysis how the question of sample size can be addressed more objectively, but for the time being take $n = 30$ samples if you can afford it and you won't go far wrong.

Using Variance

Variance is used in two main ways:

- for establishing measures of unreliability (e.g. confidence intervals), and
- for testing hypotheses (e.g. Student's *t*-test).

A Measure of Unreliability

Consider the properties that you would like a measure of unreliability to possess. As the variance of the data increases, what would happen to unreliability of estimated parameters? Would it go up or down? Unreliability would go up as variance increased, so we would want to have the variance on the top (the numerator) of any divisions in our formula for unreliability:

$$\text{unreliability} \propto s^2.$$

What about sample size? Would you want your estimate of unreliability to go up or down as sample size, n, increased? You would want unreliability to go down as sample size went up, so you would put sample size on the bottom of the formula for unreliability (i.e. in the denominator):

$$\text{unreliability} \propto \frac{s^2}{n}.$$

Finally, consider the units in which unreliability is measured. What are the units in which our current measure is expressed? Sample size is dimensionless, but variance is based on the sum of squared differences, so it has dimensions of mean squared. So if the mean was a length in cm, the variance would be an area in cm^2. This is an unfortunate state of affairs. It would make good sense to have the dimensions of the unreliability measure and the parameter whose unreliability it is measuring to be the same. That is why all unreliability measures are enclosed inside a big square root term. Unreliability measures are called **standard errors**. What we have just worked out is **the standard error of the mean**

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}.$$

This is a very important equation and should be memorized. Let's calculate the standard errors of each of our market garden means:

```
sqrt(s2A/10)
[ 1] 0.3651484
sqrt(s2B/10)
[ 1] 0.3651484
sqrt(s2C/10)
[ 1] 1.19257
```

In written work you should show the unreliability of any estimated parameter in a formal, structured way: ‘the mean ozone concentration in Garden A was 3.0 ± 0.365 pphm (1 s.e., $n = 10$)’. You write plus or minus, then the unreliability measure, the units (parts per hundred million in this case) then, in brackets, tell the reader what the unreliability measure is (in this case one standard error) and the size of the sample on which the parameter estimate was based (in this case, 10). This may seem rather stilted, unnecessary even. But the problem is that unless you do this, the reader will not know what kind of unreliability measure you have used. For example, you might have used a 95% confidence interval or a 99% confidence interval instead of one standard error.

Confidence Intervals

A confidence interval shows the likely range in which the mean would fall if the smpling exercise were to be repeated. It is a very important concept that people always find difficult to grasp at first. It is pretty clear that the confidence interval will get wider as the unreliability goes up, so

$$\text{confidence interval} \propto \text{unreliability measure} \propto \sqrt{\frac{s^2}{n}}.$$

But what do we mean by ‘confidence’? This is the hard thing to grasp. Ask yourself this question. Would the interval be wider or narrower if we wanted to be **more** confident that out repeat sample mean will fall inside the interval? It may take some thought, but you should be able to convince yourself that the more confident you want to be, the **wider** the interval will need to be. You can see this clearly by considering the limiting case of complete and absolute certainty. Nothing is certain in statistical science, so the interval would have to be infinitely wide.

We can produce confidence intervals of different widths by specifying different levels of confidence. The higher the confidence, the wider the interval. How exactly does this work? How do we turn the proportionality ($\propto$) in the equation above into equality? The answer is by resorting to an appropriate theoretical distribution. Suppose our sample size is too small to use the normal distribution ($n < 30$, as here), then we traditionally use Student’s t-distribution. The values of Student’s t associated with different levels of confidence are available in the function qt, which gives the quantiles of the t-distribution. Confidence intervals are always two-tailed because the parameter may be larger or smaller than our estimate of it. Thus, if we want to establish a 95% confidence interval we need to calculate the value of Student’s t associated with $\alpha = 0.025$ (i.e. with 0.01*(100%-95%)/2). The value is found like this for the left (0.025) and right-hand (0.975) tails:

```
qt(.025,9)
```

```
[ 1] –2.262157
```

```
qt(.975,9)
```

```
[ 1] 2.262157
```

The first argument in qt is the probability and the second is the degrees of freedom. This says that values as small as -2.262 standard errors below the mean are to be expected in 2.5% of cases ($p = 0.025$), and values as large as $+2.262$ standard errors above the mean with similar probability ($p = 0.975$). Values of Student's t are **numbers of standard errors** to be expected with specified probability and for a given number of degrees of freedom. The values of t for 99% are bigger than these (0.005 in each tail):

```
qt(.995,9)
```

```
[ 1] 3.249836
```

and the value for 99.5% confidence are bigger still (0.0025 in each tail):

```
qt(.9975,9)
```

```
[ 1] 3.689662
```

Values of Student's t like these appear in the formula for calculating the width of the confidence interval, and their inclusion is the reason why the width of the confidence interval goes up as our degree of confidence is increased. The other component of the formula, the standard error, is not affected by our choice of confidence level. So, finally, we can write down the formula for the confidence interval of a mean based on a small sample ($n < 30$):

$$\text{confidence interval} = t\text{-value} \times \text{standard error},$$

$$CI_{95\%} = t_{(\alpha=0.025,\text{d.f.}=9)}\sqrt{\frac{s^2}{n}}.$$

For Garden B, therefore, we calculate

```
qt(.975,9)*sqrt(1.33333/10)
```

```
[ 1] 0.826022
```

and we would present the result in written work: 'the mean ozone concentration in Garden B was $5.0 \pm 0.826(95\%\ \text{C.I.}, n = 10)$.'

Bootstrap

A completely different way of calculating confidence intervals is called bootstrapping. You have probably heard the old phrase about 'pulling yourself up by your own bootlaces'. That is where the term comes from. It is used in the sense of getting 'something for nothing'. The idea is very simple. You have a single sample of n measurements, but you can sample from this in very many ways, so long as you allow some values to appear more than once, and other samples to be left out (i.e. sampling with replacement). All you do is calculate the sample mean lots of times, once for each sampling from your data, then obtain the confidence interval by looking at the extreme highs and lows of the estimated means using a function called `quantile` to extract the interval you want (e.g. a 95% interval is specified using `c(0.0275, 0.975)` to locate the lower and upper bounds). Here are the data:

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

```
[ 1] "values"
```

We shall simulate sample sizes (k) between 5 and 30, and for each sample size we shall take 10 000 independent samples from our data (the vector called values), using the function called `sample` with replacement (`replace=T`):

```
plot(c(0,30),c(0,60),type="n",xlab="Sample size",ylab="Confidence interval")
for (k in seq(5,30,3)){
a<-numeric(10000)
for (i in 1:10000){
  a[i]<-mean(sample(values,k,replace=T))
}
points(c(k,k),quantile(a,c(.025,.975)),type="b")
}
```

The confidence interval narrows rapidly over the range of sample sizes up to about 20, but more slowly thereafter. At $n = 30$, the bootstrapped CI based on 10 000 simulations was

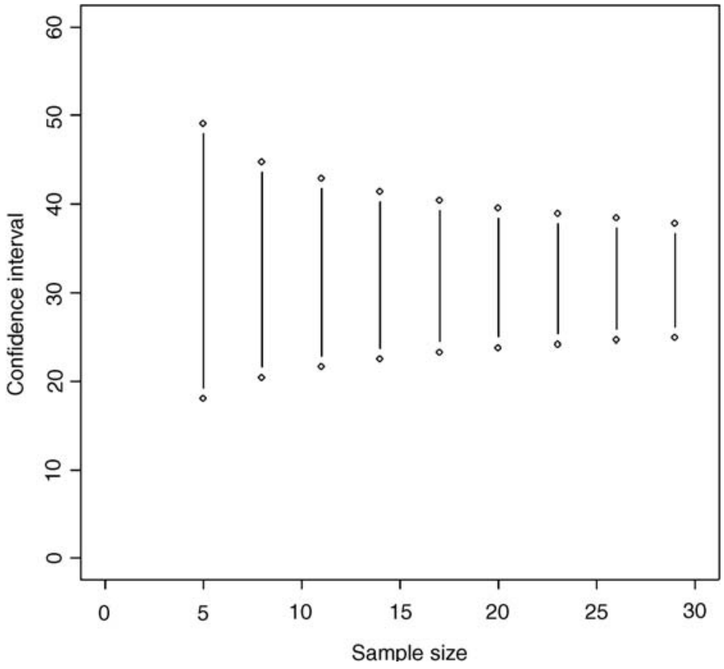
```
quantile(a,c(.025,.975))
```

```
   2.5%     97.5%
24.86843  37.68985
```

(you will get slightly different values because of the randomization). It is interesting to see how this compares with the Normal theory confidence interval:

$$1.96\sqrt{\frac{s^2}{n}} = 1.96\sqrt{\frac{337.065}{30}} = 6.5698,$$

implying that the sample mean lies in the range 24.39885 to 37.53846. As you see, the estimates from the bootstrap and Normal theory are reassuringly close, but they are not identical.



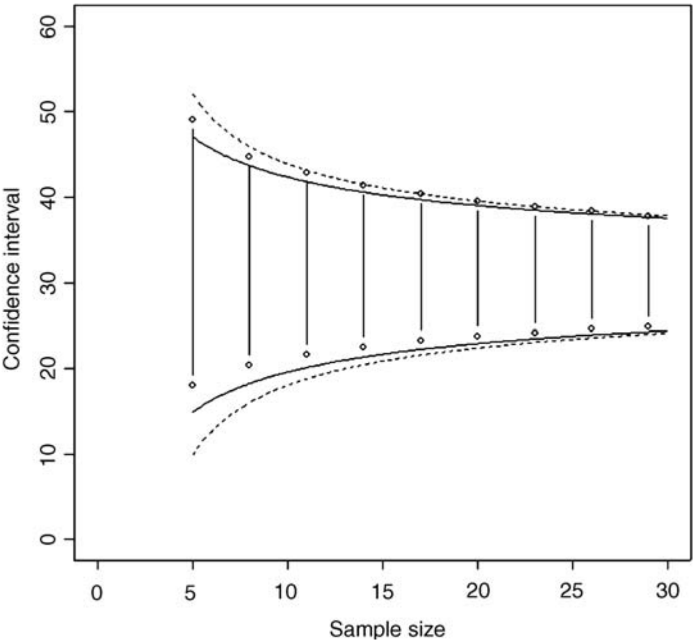
Here are the bootstrapped intervals compared with the intervals calculated from the Normal (solid line) on p. 49:

```
xv<-seq(5,30,0.1)
yv<-mean(values)+1.96*sqrt(var(values)/xv)
lines(xv,yv)
yv<-mean(values)-1.96*sqrt(var(values)/xv)
lines(xv,yv)
```

and Student's *t*-distribution (dotted line):

```
yv<-mean(values)-qt(.975,xv)*sqrt(var(values)/xv)
lines(xv,yv,lty=2)
yv<-mean(values)+qt(.975,xv)*sqrt(var(values)/xv)
lines(xv,yv,lty=2)
```

For the upper interval, you see that the bootstrapped intervals (vertical lines and open symbols, `type="b"`) fall between the Normal (the lower, solid line) and the Student's *t* distribution (the greater, dotted line). For the lower interval, however, the bootstrapped intervals are quite different. This is because of the skewness exhibited by these data (see p. 70). Very small values of the response are substantially less likely than predicted by the symmetrical Normal (solid line) or Student's *t*-distributions (dotted line). Recall that for small-sample confidence intervals using Student's *t*-distribution, the sample size, *n*,



enters the equation twice: once as the denominator in the formula for the standard error of the mean, then again as a determinant of the quantile of the *t*-distribution `qt(0.975,n)`. That is why the difference between the Normal and the Student's *t* confidence intervals gets bigger as sample size gets smaller.

So which kind of confidence interval should you choose? I prefer the bootstrapped estimate because it makes fewer assumptions. If, as in our example, the data are skewed, then this is reflected in the asymmetry of the confidence intervals above and below the mean (6.7 above the mean, and 6.1 below it, at $n = 30$). Both Normal and Student's *t* assume that there is no skew, and so their confidence intervals are symmetrical, whatever the data actually show.