Simulation

Contents

1
One-sample problems

Inference is the process of reaching conclusions based on evidence and reasoning. In *statistical inference* we use data, and statistical models, to try and say something about a population, about a system or about an underlying process. What makes statistical inference harder is the variability in data; data as an imperfect sample or estimate from the real population/system/process. More formally, statistical inference is mostly concerned with *estimating* population parameters, and *estimating the uncertainty* involved in these estimates. The typical topics are estimation of parameters, uncertainty estimates (e.g. confidence intervals), and hypothesis tests.

1.1 Using simulation as a tool for solving inferential problems

In previous courses you would have been introduced to quite a large number of procedures for testing hypotheses and constructing associated confidence intervals (z-tests, t-tests, F-tests, Chi-squared tests, non-parametric tests, analysis of variance, . . .). Often in first statistics courses these are taught by pairing data "types" with a particular tests, often in a fairly formulaic way[1]. A complete understanding of theoretical foundations which support the procedures and tests introduced in earlier courses requires a relatively high level of mathematics, beyond that with which most students in this course will be familiar. Our approach will be to develop a degree of understanding and insight by *simulating* sampling processes, hypothesis tests and estimation procedures on a computer (largely within a spreadsheet framework, such as Microsoft Excel). Some mathematical representations of the simulated phenomena will be introduced, without rigorous proofs, but with demonstration of consistency with empirical results. These mathematical representations then underpin more efficient and rigorous statistical methods.

It is useful at this stage to illustrate our *modus operandi* in the context of one of the simplest statistical problems, namely that of drawing inferences about a single population mean.

A mining company has been concerned about its cash flow position, and one of the problem areas seems to be the delay in payments for bulk shipments of copper to industrial clients. A sample of 30 recent deliveries has been carefully followed up, and the days

[1] For example, if you wish to compare two group means and the population variance is unknown, you use a t-test. But then you need to know "which" t-test to use – if the data is paired, you use a paired t-test; if there's no pairing you use an unpaired t-test, but the degrees of freedom depend on whether the sample variances of the two groups are significantly different or not (which you test with an F-test!).

between invoicing and receipt of payment recorded in each case as follows:

37	42	38	44	39	35
43	41	42	38	36	34
37	42	36	38	41	39
39	37	34	34	41	41
40	38	38	46	38	42

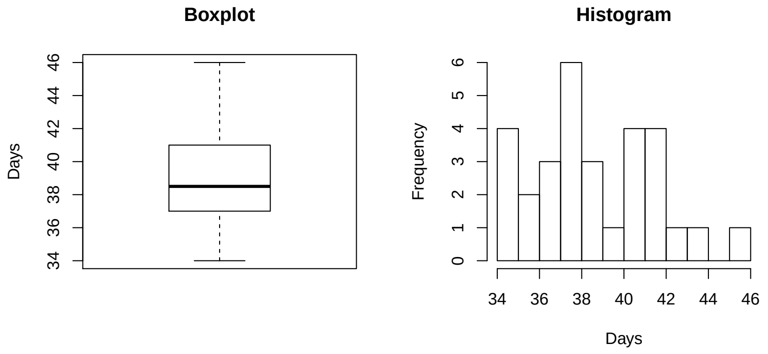
The sample data can be summarized in any of the familiar ways, for example by a box-and-whisker plot, or by histogram. We can do this by loading the data into R using

```
pdelays <- c(37,42,38,44,39,35,43,41,42,38,36,34,37,42,
       36,38,41,39,39,37,34,34,41,41,40,38,38,46,38,42)
```

and creating the plots

```
boxplot(pdelays,ylab="Days",main="Boxplot")
hist(pdelays,breaks=10,xlab="Days",main="Histogram")
```

Figure 1.1: Summary plots of sample data copper payment delays



(a)

(b)

We can also compute a few summary statistics:

```
mean(pdelays)

[1] 39

sd(pdelays)

[1] 3.051286

median(pdelays)

[1] 38.5
```

The first question to be asked might relate to the mean time to payment, as in the long run this would impact on the cash reserves tied up. It is easily calculated that the mean and median of the sample data are 39 and 38.5 days respectively, but how good are these estimates? By how much could they change had some other data point been included in the sample? Or, in other words, what differences might arise between the *sample mean* and the *population mean*? The traditional approach to answering this question, also called sampling theory approach to statistical inference, focusses on the following fundamental questions:

> *What would happen if we were able to repeat the sampling process many times? How different would the sample mean and standard deviation be each time we re-sampled? Would that lead to different conclusions? How sure does this make us of our current conclusions?*

It is precisely these questions which lead us to the familiar concepts of hypothesis tests (e.g. are the data consistent with a claim that the long-run mean time does not exceed 40 days?) and of confidence intervals for the true population mean.

The mathematics behind the standard tests can become quite intricate (e.g. deriving the theoretical sampling distribution for the sample mean is not easy[2]). But there is another approach which does not rely on the central limit theorem, or any other distributional assumptions: we can mimic, or *simulate* the process of re-sampling on a computer (for example within a spreadsheet). One simple procedure for performing such a simulation is the method which has been termed "*bootstrapping*". In essence, this procedure is as follows:

(1) Start with the observed set of observations (a "random sample"), denoted by $x_1, x_2, \ldots, x_n$, and calculate any relevant summary statistics (e.g. sample mean and variance).

(2) Place all the observations "in a hat" and "shuffle" them. Then draw n items *with replacement* to form a new pseudo-sample (a "bootstrap sample").

(3) Recalculate the summary statistics for this new sample.

(4) Repeat the previous two steps a large number of times.

The above procedure would be equivalent to the process of repeatedly re-sampling from the population, if the original sample were an exact representation of the total population, i.e. with precisely $1/n$ of the population taking on each of the values $x_1, x_2, \ldots, x_n$. The reason for sampling with replacement is that the proportions in the population corresponding to each x_i should not change during the sampling process. Of course, the sample can never be an exact representation of the population, but it is usually close enough for purposes of assessing the ranges of variation which can arise through re-sampling.

The "shuffling" and "sampling with replacement" step of the bootstrap procedure is easily achieved algorithmically (i.e. within a computer program) by the following mechanism:

[2] By using the central limit theorem, it can be shown that for large enough n

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

This is an example of sampling theory: the above distribution gives the *sampling distribution* for the sample mean, which describes the behaviour of the sample mean *if* we would have a large number of samples from the population. The underlying theory (for sample mean and related statistics such as z, t, F, χ^2) is also referred to as *normal theory*, as it depends strongly on the properties of the normal distribution.

```
For $i = 1,...,n$ do:

• Draw a uniformly distributed random number $u$
  (e.g. by using Excel's RAND() function);

• Define $k$ as 1+INT($n \times u$), where the function
  'INT()' is the Excel spreadsheet function
  which returns the integer part of a real
  number;

• Set the $i$-th observation in the bootstrap
  sample to $x_k$.
```

The process is illustrated in the following table, in which the sample is composed of the first eight observations only from the payment delay data above.

Original sample		**1st bootstrap sample**			**2nd bootstrap sample**		
index	**value**	**random number**	**index**	**value**	**random number**	**index**	**value**
(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
1	37	0.4488	4	44	0.5033	5	39
2	42	0.7895	7	43	0.7191	6	35
3	38	0.5253	5	39	0.6484	6	35
4	44	0.3476	3	38	0.3332	3	38
5	39	0.6935	6	35	0.4328	4	44
6	35	0.9859	8	41	0.5852	5	39
7	43	0.4037	4	44	0.5374	5	39
8	41	0.4019	4	44	0.3565	3	38
mean	*39.88*			*41.00*			*38.38*
std dev	*3.14*			*3.38*			*2.83*

In column (3) is shown a set of 8 random numbers generated using the Excel RAND() function, which are used as described in the box above to give the index numbers in column (4). The corresponding values from the original data set are given in column (5). This process is repeated in columns (6)–(8), in order to give a second bootstrap sample.

Note that in the first bootstrap sample, the 4th observation from the original data appears three times, while the first two do not appear at all. In the second bootstrap sample, only observations 3–6 of the original sample appear at all. If this process is repeated a large number of times, many different combinations of the original data will appear.

For each bootstrap sample, the sample mean and standard deviation can be calculated, giving an immediate impression of how variable these estimates are. (This is even evident from the two repetitions above!)

Such a process would be tedious to repeat for very large numbers of resamplings, so that some form of automated procedure is needed, i.e. a computer programme. We'll first implement this using the *Visual Basic for Applications (VBA)* facility in Excel, and then write our own R program. VBA is an extremely useful feature for many statistical and management science applications, and students are advised to become familiar with this package. However, we do make available for students a simple spreadsheet package (BootStrap.xls), in which such a macro has already been coded. [*Students interested in the coding can examine the code by clicking Tools|Macro|Visual Basic Editor; feel free to experiment with changing the code!*] If this file is opened under Excel, the user will see rows 1-6 and columns A-H of the spreadsheet illustrated in Figure 1.2.

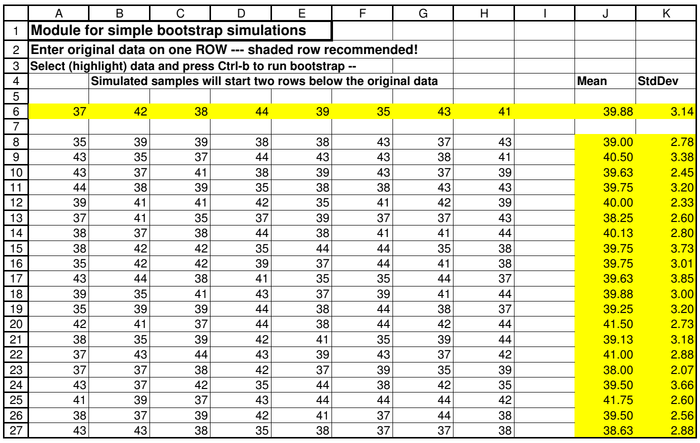


Figure 1.2: Illustration of the BootStrap.xls package

	A	B	C	D	E	F	G	H	I	J	K
1	**Module for simple bootstrap simulations**										
2	**Enter original data on one ROW --- shaded row recommended!**										
3	**Select (highlight) data and press Ctrl-b to run bootstrap --**										
4		**Simulated samples will start two rows below the original data**								**Mean**	**StdDev**
5											
6	37	42	38	44	39	35	43	41		39.88	3.14
7											
8	35	39	39	38	38	43	37	43		39.00	2.78
9	43	35	37	44	43	43	38	41		40.50	3.38
10	43	37	41	38	39	43	37	39		39.63	2.45
11	44	38	39	35	38	38	43	43		39.75	3.20
12	39	41	41	42	35	41	42	39		40.00	2.33
13	37	41	35	37	39	37	37	43		38.25	2.60
14	38	37	38	44	38	41	41	44		40.13	2.80
15	38	42	42	35	44	44	35	38		39.75	3.73
16	35	42	42	39	37	44	41	38		39.75	3.01
17	43	44	38	41	35	35	44	37		39.63	3.85
18	39	35	41	43	37	39	41	44		39.88	3.00
19	35	39	39	44	38	44	38	37		39.25	3.20
20	42	41	37	44	38	44	42	44		41.50	2.73
21	38	35	39	42	41	35	39	44		39.13	3.18
22	37	43	44	43	39	43	37	42		41.00	2.88
23	37	37	38	42	37	39	35	39		38.00	2.07
24	43	37	42	35	44	38	42	35		39.50	3.66
25	41	39	37	43	44	44	44	42		41.75	2.60
26	38	37	39	42	41	37	44	38		39.50	2.56
27	43	43	38	35	38	37	37	38		38.63	2.88

In order to use the BootStrap.xls package, the user needs to do the following:

- Enter the sample data as a long row, in the spreadsheet row 6;
- Create the formulae to calculate any summary statistics (e.g. mean and standard deviation) to the right of the data, in the same row (row 6);
- Select (highlight) the sample data (not the summary statistics) in row 6, and press Ctrl-b; you will be prompted to enter the number of times the re-sampling of the data is to be carried out;
- The simulated bootstrap samples will appear row-wise, starting in row 8; copy the formulae for the summary statistics down to all the re-sampled rows;
- Analyze the variation in the summary statistics from sample to sample.

For example, in Figure 1.2, the first 20 bootstrap samples (based on a sample of size 8) are shown, together with the means and standard deviations in each case. Of course, in order to obtain a meaningful understanding of the extent of variation in sample estimates (such as of the mean and standard deviation), you need to carry out much more than 20 repeated samples.

To do bootstrap sampling in R, we'll first load the full sample of 30 payment delays previously described:

```
pdelays <- c(37,42,38,44,39,35,43,41,42,38,36,34,37,42,
       36,38,41,39,39,37,34,34,41,41,40,38,38,46,38,42)
```

We then use the `sample()` function to create a single bootstrap sample, which we store in a new variable called `bootx`.[3]

```
bootx <- sample(pdelays, size = 30, replace = TRUE)
```

To create many bootstrap samples we put the code above in a `for` loop, being careful to store the results.

```
# set up a matrix to store 5000 bootstrap samples in rows
all_boots <- matrix(NA, nrow = 5000, ncol = 30)
for(i in 1:5000){
  # draw a single bootstrap sample from pdelays
  boot <- sample(pdelays, size = 30, replace = TRUE)
  # store that bootstrap in row i
  all_boots[i,] <- boot
}
```

Having created the bootstrap samples[4], we can now extract the bootstrap means using the `apply` function[5].

```
bs_means <- apply(all_boots, MARGIN = 1, FUN = mean)
```

However we construct the bootstrap sample means, we can now summarize these in a number of ways. Figure 1.3 shows a boxplot and histogram of the bootstrapped sample means. This distribution is also called the *bootstrap distribution for the sample mean*, as opposed to the *sampling distribution* $\bar{X} \sim N(\mu, \sigma^2)$ (see earlier).

```
boxplot(bs_means,ylab="Days",main="Boxplot")
hist(bs_means,breaks=10,xlab="Days",main="Histogram")
```

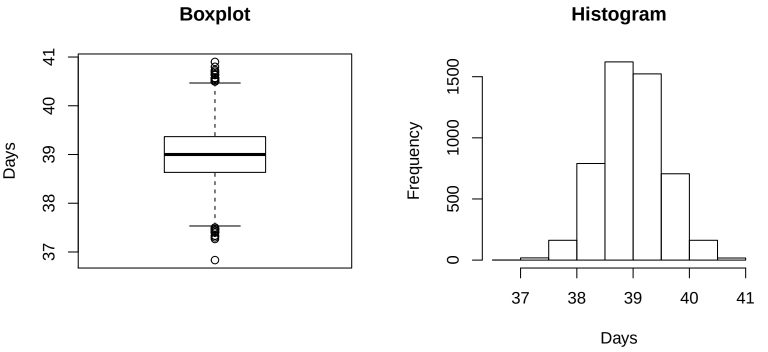
We can also compute a few summary statistics:

```
mean(bs_means)

[1] 38.99818

sd(bs_means)

[1] 0.5539776

min(bs_means)

[1] 36.83333

max(bs_means)

[1] 40.9
```

[3] The parts inside the brackets are called the *arguments* of the `sample` function. The `sample` function takes 3 arguments: the original data (which we have called `pdelays`, the size of the new sample, and a logical variable (`TRUE` or `FALSE`) indicating whether sampling should be with replacement or not. In R you can get help on any function (say `sample`) by `?sample` or `help(sample)`

[4] In general there are many ways to program a particular task. Another way of constructing a matrix containing the bootstrap samples is `all_boots=matrix(sample(pdelays, size=5000*30,replace=TRUE),nrow= 5000,ncol=30)`.This way is slightly better because we avoid the for loop, which saves time, but the code is a bit trickier to understand.

[5] `apply` applies a function (e.g. mean) to every row (MARGIN = 1) or every column (MARGIN = 2) of a matrix. We obtain a vector with 5000 bootstrapped sample means.



(a)

(b)

Figure 1.3: Summary plots of bootstrap distribution for sample mean copper payment delays. B = 5000.

Figure 1.3 shows the variation in sample means based on 5000 bootstrap replicate samples drawn from the full sample of 30 payment delays previously described. Note that the re-sampled sample means range between 36.83 and 40.9 (in comparison with the original sample mean of 39). Thus we see that *sampling errors*[6] in the estimation of the mean are at least of the order of ±2 days, when these estimates are based on a sample of size 30.

Let's take a step back and consider: We now have the boostrap distribution for the sample mean. But what does this tell us, and how can we use it for statistical inference?

A few notes on bootstrapping:

- The size of each bootstrap sample must be the same as that of the original sample. This is because we want to know the behaviour of our original statistic, which was calculated from the original sample with sample size n. Statistics based on larger samples are more precise, and less precise when based on smaller samples, but our statistic is based on sample size n. We need to work with the information we have!

- The bootstrap distribution replaces the sampling distribution (which may be very difficult to derive, or unknown). This is useful! Even the distribution of the sample mean may not be well approximated by the normal approximation in small samples.

- We are resampling from the sample, not from the population. The sample is centered at the observed statistic, whereas the population is centered at the population parameter. There are two important implications of this (**?**):

  - The original parameter estimate is still the best estimate. We cannot use the mean of the bootstrap distribution to improve on this.

  - We can use the bootstrap distribution only to estimate things like the standard deviation of $\hat{\theta}$ (standard error), the expected value of $\hat{\theta} - \theta$ (bias), and the CDF and quantiles of $\hat{\theta} - \theta$ or $(\hat{\theta} - \theta)/SE$.

[6] sampling error: difference between observed statistic and true parameter value because of sampling, i.e. each sample will be slightly different, and different from the population.

- Bootstrapping relies on an important assumption/principle, called the bootstrap principle: $\hat{\theta}^* - \hat{\theta} \sim \hat{\theta} - \theta$, i.e. the behaviour of the bootstrapped statistics around the original sample statistic reflects the behaviour of the sample statistic around the parameter value.

  Here we call $\hat{\theta} - \theta$ the sampling error, and we estimate the size of this using $\hat{\theta}^* - \hat{\theta}$

θ = population parameter
$\hat{\theta}$ = original sample statistic
$\hat{\theta}^*$ = bootstrapped statistic

1.1.1 *Using bootstrapping to construct confidence intervals*

Continuing with the copper payment example, we start by sorting all the bootstrap sample means from smallest to largest. In R we would do this by

```
# sort the elements in bs_means from smallest to largest
sorted_bs_means = sort(bs_means, decreasing = FALSE)
# show the first six
sorted_bs_means[1:6]

[1] 36.83333 37.26667 37.30000 37.33333 37.33333 37.40000
```

For our particular simulation of 5000 repetitions of the sampling we can find, for example, the 125-th and 4875-th (sorted) mean, corresponding to the 0.025 and 0.975 quantiles.

```
sorted_bs_means[125]

[1] 37.93333

sorted_bs_means[4875]

[1] 40.1
```

These are perhaps more interesting than the absolute extremes, and tell us that 4750 out of 5000 (i.e. 95%) of the bootstrapped means lay between 37.93 and 40.1. This is a useful observation, but has to be interpreted with some caution (e.g. see **?** Figure 9, for a nice demonstration of what can go wrong.). Nevertheless, quantiles from the bootstrap distribution are sometimes directly used to construct a 95% confidence interval, and this is then called a *percentile bootstrap interval*.

Alternatively, if we recognize that the bootstrap samples have been generated from a hypothetical population with mean = the original sample mean, i.e. 39.0, the bootstrap distribution really tells us that in 95% of samples (and by implication in hypothetical resampling from the population), the deviation between the sampled mean and the true mean will lie between -1.07 (=37.93-39) and 1.1 (=40.1-39). IF (and this is the critical assumption) the same errors apply to the real sample, we may be 95% confident that the error of estimation lies between -1.07 and 1.1. Based on our observed mean of 39.0, we argue that:

- If the error is as small as -1.07, then the population mean would actually be 39 + 1.07 = 40.07;

- If the error is as large as 1.1, then the population mean would actually be 39 - 1.1 = 37.9.

We would thus claim a 95% confidence interval (the "bootstrap confidence interval") for the mean as [37.9; 40.07]

Suppose that the true (population) mean is μ, while the sample mean is $\bar{X}$. The above ideas can be expressed in algebraic terms by noting that the measurement error is $\bar{X} - \mu$, so that the bootstrap results imply:

$$\Pr[-1.07 \leq \bar{X} - \mu \leq 1.1] = \Pr[\bar{X} - 1.1 \leq \mu \leq \bar{X} + 1.07] = 0.95$$

Although this looks like a probability statement about μ, in a confidence interval context the population parameter, μ here, is fixed (although unknown). The probability refers to the different realizations of the sample mean when the sampling process is repeated many times.

The confidence interval described in introductory statistics courses is in fact derived on precisely the same type of re-sampling argument. The only difference is that the assumption is made that the errors themselves are *normally distributed* with zero mean, i.e. that $\bar{X} - \mu$ has a normal distribution with mean 0 and standard deviation $\sigma/\sqrt{n}$, where σ is the population standard deviation and n the sample size. When σ is known, the standard confidence interval is written in the form:

$$\Pr[\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}] = 1 - 2\alpha$$

where z_α is the α critical value of the normal distribution distribution (e.g. $z_{0.025} = 1.960$ for the usual 95% interval).

When σ is unknown and is replaced by its corresponding sample estimate s, we calculate the confidence interval as:

$$\Pr[\bar{X} - t_{n-1,\alpha} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha} \frac{s}{\sqrt{n}}] = 1 - 2\alpha$$

where $t_{n-1,\alpha}$ is the α critical value of the t-distribution with $n-1$ degrees of freedom. We shall return later to the precise reasons for this result, but record for now that the 95% confidence interval (based on normal sampling theory) for the (population) mean payment delay in our copper example is:

$$39 \pm \frac{2.045 \times 3.05}{\sqrt{30}} \ = \ [37.86; 40.14]$$

where the factor 2.045 is the 0.025 critical value for the t-distribution with 29 degrees of freedom. Note how close the bootstrapped and t-based confidence intervals are; this would suggest that the assumptions of the normal theory hold quite well in this case.

1.1.2 *Bootstrap hypothesis testing*

In a similar manner we can perform hypothesis tests on the basis of the bootstrap samples. Suppose, for example, that we wish to test the null hypothesis $H_0 : \mu \leq 38$ against the alternative $H_1 : \mu > 38$. If the (population) mean does not actually exceed 38, then the observed sample mean corresponds to a sampling error of at least +1.0 days ($\hat{\theta} - \theta$). Since the bootstrap sample means were based on a "population" with mean 39, errors of 1.0 or more correspond to bootstrap sample means of 40.0 or more. We can find the number of values exceeding 40.0 by

```
# create a vector with elements TRUE if >40, else FALSE
mean_gt40 <- (bs_means > 40)
# count up the number of TRUEs (TRUE = 1, FALSE = 0)
sum(mean_gt40)

[1] 179
```

So in our bootstrap sample 179 of the 5000 values exceeded 40.0 (3.58%). This may be interpreted as a *p*-value of 0.036 and evidence that the true mean delay time exceeds 38 days. Formally, we calculate the p-value as the probability of observing a sample mean $>= 39$ under $H_0 : \theta = \theta_0 = 38$:

$$p = P((\hat{\theta} - \theta_0) >= 39 - 38)$$

and estimate this probability from the corresponding sampling errors in the bootstrap distribution:

$$p = P((\hat{\theta}^* - \hat{\theta}) >= 39 - 38) = P(\hat{\theta}^* >= 39 + 39 - 38) = P(\hat{\theta}^* >= 40)$$

Comparing this to the the standard *t*-test based on normal errors, recall that the *t*-statistic for this test would be:

$$\frac{39 - 38}{3.05/\sqrt{30}} = 1.795$$

We can use R to find that the corresponding p-value is 0.042 (one-sided, 29 degrees of freedom).[7,8]

```
# calculate parts of the t-statistic
mpd <- mean(pdelays)
sdpd <- sd(pdelays)
npd <- length(pdelays)
# compute the t-statistic
tstat <- (mpd - 38) / (sdpd/sqrt(npd))
# enter the degrees of freedom
dof <- npd - 1
# get the p-value from a t-distribution
p <- 1 - pt(tstat,dof)
# display p
p
```

[7] The crucial part of the code below is `pt`, which calculates the value of the CDF of a t-distribution at the point 1.8. This gives us the probability of obtaining a value less than 1.8 (think of the definition of a CDF). For our (one-sided) hypothesis test we want the probability of getting more than 1.8, which is why we take `1-` the value returned by `dt`. R has similar functions for lots of other distributions, see for example `dnorm`, `dgamma`, `dbeta`, `dbinom`. Note that `dt` is specifically for the *t*-distribution!

[8] Note that again you could do this in a single line of code: `1-pt((mean(pdelays)-38)/(sd(pdelays)/sqrt(length(pdelays))),length(pdelays-1))`. Its just easier to understand, and harder to make a mistake, if you break it up a little. In fact, you can run the *t*-test directly using `t.test(pdelays,mu=38, alternative="greater")`. Try this, and note that you get the same *p*-value. For help type `help(t.test)`. Using R's built-in `t.test` function is the way you would normally run a *t*-test in practice, but it doesn't give you any insight into sampling theory, which is the whole point of this section!

```
[1] 0.04153733
```

The conclusion is similar as before, there is some evidence that the true mean delay is longer than 38 days. This p-value is somewhat larger than (but of much the same order of magnitude as) the bootstrap estimate.

It is worth emphasizing two critical insights derived from the above arguments:

- A 95% confidence interval *does not* mean a probability of 0.95 that the population parameter is within the stated bounds . . . it is the frequency (in repeated sampling) of stating valid bounds (confidence limits);
- The p-value *does not* give the probability that the null hypothesis is correct . . . it is the probability of the observed data/test statistic (or more extreme) under H_0, and a measure of how incompatible the data are with the model.
- Scientific or management decisions should never be based on a p-value alone, but should be evaluated in the context of other evidence and knowledge.

While the "bootstrap" concept does in principle allow us to simulate the effects of re-sampling, so as to generate confidence intervals or p-values, this can become quite tedious, especially for complicated sampling situations involving two or more different populations. For this reason, we seek means of solving essentially the same problems, at least approximately, by a more analytical approach. In the following chapter, we extend both the bootstrap and analytical approaches into more complex sampling situations involving two or more samples.

1.2 Some References on the Bootstrap

- Davison AC, Diego Kuonen. An introduction to the bootstrap with applications in R. Statistical Computing and Statistical Graphics Newsletter. 13(1):6-11.
- Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statist Med. 2000 May 15;19(9):1141-64.
- Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge University Press; 1997.
- Hesterberg TC. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. The American Statistician. 2015 Oct 2;69(4):371-86.

1.3 Simulating Random Variables

Recall that a *random variable* is defined such that its value is only determined by performing some form of experiment, measurement or observation. Examples in earlier courses would have included the number of heads in a fixed number of spins of a coin, the time to occurrence of some specified event (e.g. the breakdown of a machine), or rainfall at a particular point over a fixed period of time. The critical point is that prior to making the necessary observations, the value of the random variable is unknown, and can only be described probabilistically.

A convenient notation is to use upper case letters (e.g. $X, Y, \ldots$) to denote the random variable itself (prior to any observation), and to use lower case letters (e.g. $a, b, \ldots, x, y, \ldots$) to represent particular real values that might be taken on by the random variable. An expression such as $\Pr[X = x]$ would then denote the probability that when the random variable X is observed, it is found to take on the value x. Clearly, expressions such as $\Pr[X = y]$ or $\Pr[X = 10]$ are equally meaningful.

The *distribution function* (sometimes termed the cumulative distribution function) of the random variable X is a function $F(x)$, such that for each real number x: $F(x) = \Pr[X \leq x]$. Clearly, as x increases, the probability cannot decrease, although it could remain constant over some range of x (since if X cannot take on values between a and b, say, then $F(a) = \Pr[X \leq a] = \Pr[X \leq b] = F(b)$). In other words, $F(x)$ is a non-decreasing function of x. Furthermore, by the properties of probabilities, $0 \leq F(x) \leq 1$. Figure 1.4 illustrates two possible forms of distribution function.

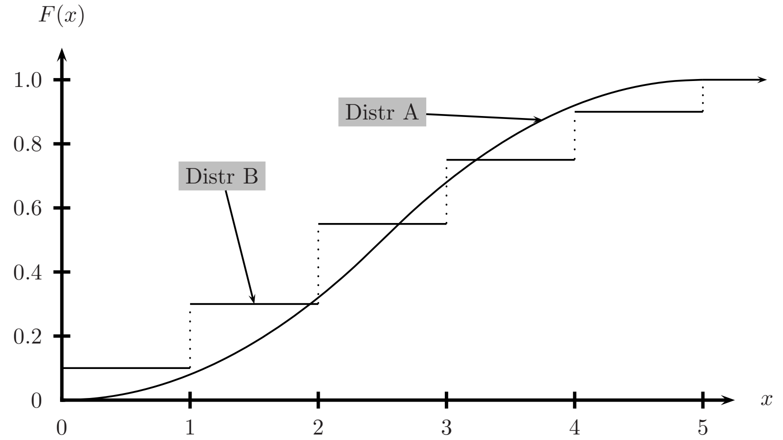


Figure 1.4: Examples of distribution functions

Both distributions illustrated in Figure 1.4 relate to random variables taking on values between 0 and 5 only. The distribution function for Distribution A is continuous, which implies that any real number between 0 and 5 is at least possible. On the other hand, the function for Distribution B has a discrete number of jumps (at integer values of x), and is otherwise flat (horizontal). Thus, for example, $F(1.8) = F(1.1)$, so that $\Pr[X \le 1.8] = \Pr[X \le 1.1]$, implying that the probability of $1.1 < X \le 1.8$ is zero. In fact, X can only take on the integer values $0, 1, \ldots, 5$, where the probabilities are given by the magnitudes of the jumps. This leads us to the concept of the *probability function*, or *probability mass function* $p(x)$ defined by $p(x) = \Pr[X = x]$.

Often, but not always, discrete random variables take on non-negative integer values (e.g. X = number of events). Then $p(x) = 0$ if x is not a non-negative integer, while for $k = 0, 1, 2, \ldots$:

$$p(x) = \Pr[X = k] = \begin{cases} F(0) & \text{for } k = 0 \\ F(k) - F(k-1) & \text{for } k = 1, 2, ... \end{cases}$$

Conversely,

$$F(k) = \sum_{i=0}^{k} p(i).$$

The probability function is often intuitively easier to understand than the more fundamental distribution function, as it can be interpreted as showing relative frequencies for each possible value for X. Unfortunately, it does not carry over to continuous distributions such as Distribution A in Figure 1.4. The problem is that the probability of any precise value is always zero (for example, $\Pr[X = \pi]$ in Figure 1.4), even though it is possible. We can always evaluate the probability associated with a range of values, i.e. $\Pr[a < X \le b] = F(b) - F(a)$, but of course the magnitude of this probability depends on the length of the interval. The *probability density* at any point x on the real line is then defined by the probability of X belonging to a small interval on the real line containing x, divided by the length of the interval, in the limit as this length tends to zero. Formally we express this in terms of the *probability density function*:

$$f(x) = \lim_{h \to 0} \frac{\Pr[x < X \le x + h]}{h} = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \frac{dF(x)}{dx}.$$

Clearly this also implies that:

$$F(x) = \int_{-\infty}^{x} f(x) dx$$

so that $F(x)$ is simply the area under the probability density function curve up to the point x.

It is conventional to express the lower bound as $-\infty$ in general expressions for continuous random variables, although $f(x)$ may be zero for some range of values (such as for $x < 0$ in many cases).

It also follows that

$$\Pr[a < X \le b] = F(b) - F(a) = \int_{a}^{b} f(x) dx$$

i.e. the area under the density function curve between the points a and b on the x-axis.

A notational convention: Sometimes it becomes necessary to identify to which random variable a particular distribution, probability or density function applies. In such cases we will label the functions by a subscript denoting the name of the random variable, e.g. $F_X(x)$, p_X or $f_X(x)$. The subscripts will, however, be omitted if no confusion can arise.

You should be familiar with the following distributional forms:

- Binomial (discrete): $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \ldots, n$
- Poisson (discrete): $p(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0, 1, \ldots$
- Exponential (continuous): $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, so that $F(x) = \int_{u=0}^{x} f(u)\,du = 1 - e^{-\lambda x}$
- Normal: $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

It is worth noting here that a non-negative random variable X is said to have the *log-normal* distribution if $Y = \log X$ follows a normal distribution. The base to which the logarithm is taken is irrelevant, but it is conventional to use natural logarithms (i.e. to base e).

As we have seen in Section 1.1, it is possible to explore the behaviour of statistical sampling processes by means of numerical experimentation in computer simulations (often then termed a *Monte Carlo* approach). In order to implement such an approach, we often need to simulate realizations of random variables drawn from some specified distribution (such as the normal or Poisson with specific parameter values). Some computer systems allow us to do this directly, but it is useful to master the general principles which are quite simple.

Suppose that we observe a sequence of random variables, say $X_1, X_2, \ldots$, drawn independently from the same probability distribution $f(x)$, and let $x_1, x_2, \ldots$ be the actual values observed. Suppose now that we actually report the cumulative probabilities corresponding to each observation, i.e. $u_1 = F(x_1);\ u_2 = F(x_2);\ \ldots$. It can be shown that the sequence of values $u_1, u_2, \ldots$ arise in fact from the uniform distribution on [0,1]. We can reverse this process to generate numbers from any desired distribution (where the distribution function $F(x)$ is given) as follows[9]:

- Generate a sequence of numbers from the uniform distribution, say $u_1, u_2, \ldots$. Most computer software systems provide some facility for doing this. For example, in R the function `runif(n)` returns n values drawn from $U[0,1]$

```
runif(4)

[1] 0.64219599 0.09792858 0.48317477 0.07788739
```

[9] This method of generating random variables is called the *probability integral transform*.

while in Excel the spreadsheet function RAND() does the same.

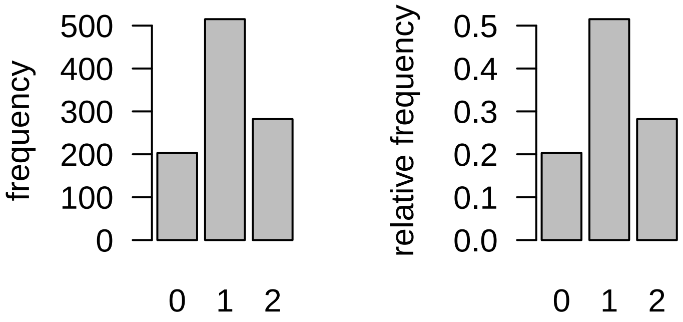
- For each u_i find the value x_i such that $F(x_i) = u_i$; the resulting sequence $x_1, x_2, \ldots$ arises from the desired distribution.

The only complication may arise from having to find the x such that $F(x) = u$. For discrete distributions, the search for a solution can be carried out systematically: Set X to the smallest non-negative integer k such that $u \leq \sum_{i=0}^{k} p(i)$. This is easily found by a set of nested IF functions[10]. For example, suppose we want to draw from a discrete distribution giving $X = 0$ with probability 0.2, $X = 1$ with probability 0.5, and $X = 2$ with probability 0.3. We could simulate 1000 values from this distribution as follows:

```
# set up an empty vector to store the values
x <- c()
# generate 1000 U[0,1] variates
u <- runif(1000)
# use u to pick appropriate value of x
for(i in 1:length(u)){
  if(u[i] < 0.2){
    x[i] <- 0
  } else if(u[i] < 0.7){
    x[i] <- 1} else {
      x[i] <- 2
    }
}
# print out first 5 u values
round(u[1:5],2)

[1] 0.51 0.59 0.81 0.73 0.15

# print out first 5 x values
x[1:5]

[1] 1 1 2 2 0
```

Two ‘histograms’ of the generated values are shown below:

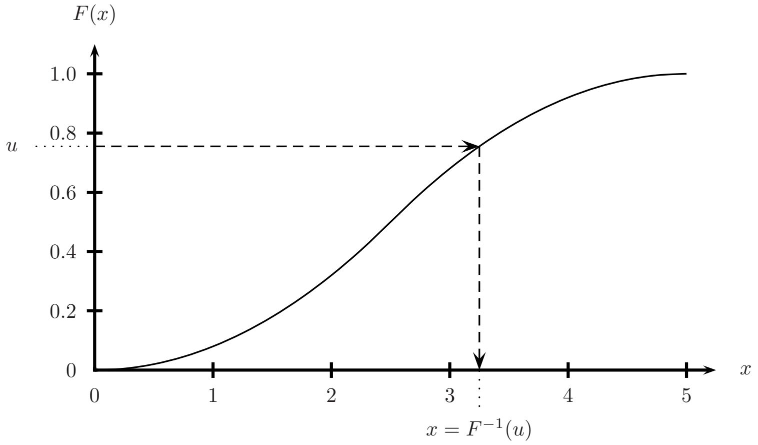
[10] Again there are different ways to program this. A better way, once you have generated your uniform random numbers, is to use `x<-ifelse(u<0.2,0,ifelse(u<0.7,1,2))`. This does the same thing as in the main text, but avoids the use of the `for` loop. Try `help(ifelse)` to get more information on the useful `ifelse` function.



For continuous distributions, we have to solve the non-linear equation $F(x) = u$ for x in terms of u. Formally we often represent this by an *"inverse"* function, i.e. as $x = F^{-1}(u)$. The idea is illustrated in Figure 1.5.

$x = F^{-1}(u)$ means: return the value of x which would result in u in the function $u = F(x)$.

Figure 1.5: Finding $x = F^{-1}(u)$



In some cases the solution is relatively simple. For example, the exponential distribution has a distribution function given by $F(x) = 1 - e^{-\lambda x}$. Solving $1 - e^{-\lambda x} = u$ for x gives $x = -[\ln(1-u)]/\lambda$. In order to generate a sequence of exponentially distributed random variables in Excel, we need therefore only to generate an array of uniform numbers (using the RAND() function), and then to convert these to the desired x's using the above function. In fact there is a little simplifying trick! If the random variable U has a uniform distribution on [0,1], then so has $1 - U$. The expression $x = -[\ln(u)]/\lambda$ will therefore also produce a number from the desired exponential distribution.

More generally, it may be difficult to solve the equation $F(x) = u$ in a closed form. However, R provides inverse functions for many standard distributions, i.e. giving values for $F^{-1}(u)$ directly. For example, the function `qnorm(u,m,s)` gives $F^{-1}(u)$ for the normal distribution with mean m and standard deviation s. Another function, `rnorm(n,m,s)` directly generates n values from $N(m,s)$.

```
# a few examples of qnorm and rnorm
qnorm(p = 0.025, mean = 0, sd = 1) # value of x ~ N(0, 1) with cumulative prob of 0.025

[1] -1.959964

qnorm(0.975,0,1)

[1] 1.959964

qnorm(runif(1), 1, 3)

[1] -0.1832144

# generate 4 values from X ~ N(10, 1), and round to 2 decimal places
round(rnorm(4, 10, 1), 2)

[1]  9.78  8.96 10.34 10.65
```

Excel also provides inverse functions for many distributions. For example, the spreadsheet function NORMINV(u, m, s) gives $F^{-1}(u)$ for the normal distribution with mean m and standard deviation s.

Simulating the central limit theorem

As an illustration of the use of simulation in understanding statistical sampling, let us consider an extremely simple situation. *Students are advised to repeat this exercise for themselves!*

We start with the Bernoulli distribution, which is the Binomial distribution with $n = 1$, so that $X = 1$ with probability p, and $X = 0$ otherwise. The random variable is easily generated; simply obtain a uniformly distributed random number u, and set $X = 1$ if $u < p$, and $X = 0$ otherwise. For the experiments reported below, we arbitrarily chose $p = 0.25$.

A set of 200000 Bernoulli random variables was generated as described above i.e.

```
u <- runif(200000)
x <- ifelse(u < 0.25, 1, 0)
```

`ifelse` operates on vectors, i.e. it executes the instruction for each value in the given vector and returns a vector of the same length. In this example, for each value in the vector u, it tests whether the value is less than 0.25, if TRUE it returns 1, else it returns 0. x will be a vector of 0's and 1's.

Two sets of analyses were carried out using the generated sequence of 0's and 1's:

- The 200000 values were grouped into 20000 sets of 10 observations each, and the mean of each set recorded.

  ```
  # arrange x into matrix of 20000 x 10
  xmat1 <- matrix(x,20000,10)
  # calculate mean of each row
  mean1 <- apply(xmat1,1,mean)
  ```

Clearly these means must take on one of the values $0, 0.1, \ldots, 1.0$, so that the distribution of the sample means is still very much discrete. The observed frequencies of the means are displayed in the histogram below

```r
hist(mean1,xlab="mean values",main="Histogram of means")
```

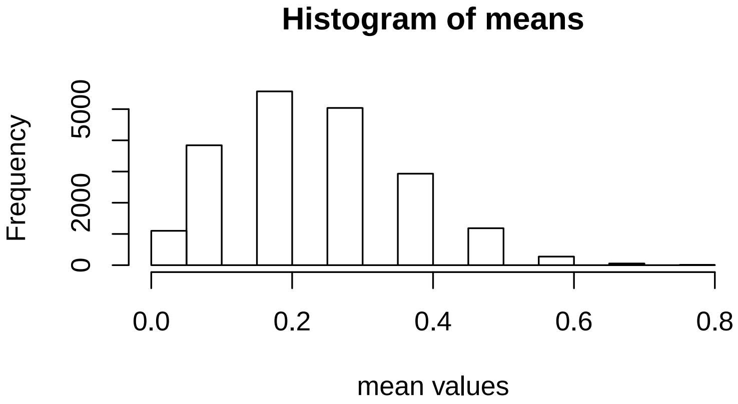


Figure 1.6: Distributions of means of samples of size 10

Even for samples of size 10, the distribution of the mean is starting to look quite smooth, although somewhat skewed to the right rather than normally distributed.

- The same 200000 values were then grouped into 2000 sets of 100 observations each, once again calculating the means in each set.

```r
# arrange x into matrix of 2000 x 100
xmat2 <- matrix(x,2000,100)
# calculate mean of each row
mean2 <- apply(xmat2,1,mean)
```

These means may now take on values $0, 0.01, 0.02, \ldots$ which are still not continuous, but are very nearly so to a reasonable approximation. Again, we can show the distribution of mean values in the form of a histogram:

```r
hist(mean2,xlab="mean values",main="Histogram of means")
```

Clearly the distribution of means is starting to take on a distinctly normal shape, even if still a little ragged.

What this type of numerical experiment shows is that even for random variables which are far from normal (the Bernoulli in this case), sample means tend to exhibit increasingly normal-like behaviour. This is the basic principle of the *central limit theorem*. We shall return to this theorem later, but it is useful to summarize the key concepts here:

Central limit theorem: the sum (or mean) of a large number of independent random variables tends towards a normal distribution.

This is very useful in statistical inference for deriving theoretical sampling distributions, as many of our statistics are sums or means, even if the original data were not normally

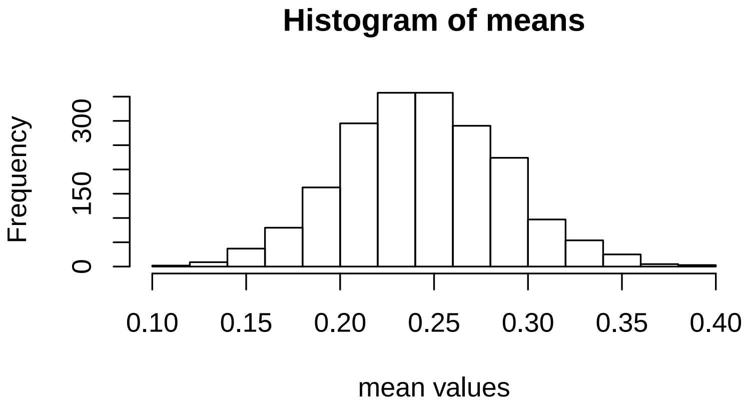


Figure 1.7: Distributions of means of samples of size 100

- A *random sample* is a set of independent random variables drawn from the same probability distribution.
- The sample mean and sample variance from a random sample are also random variables, and thus have their own probability distributions (sampling distributions).
- For almost all distributions of practical interest (discrete and continuous), the distribution of the sample mean approaches a normal distribution for large enough sample sizes (as in the above example).
- The central limit theorem thus justifies using normal distribution theory for any inference involving averages (which explains why the "bootstrap" and normal theory results in the introductory chapter were so similar).

1.4 *Order statistics*

In Section 1.1, we have used simulation to re-examine some fairly straightforward inference problems involving the mean of a probability distribution function, of the kind that you would have encountered in a first-year statistics course. In this section, we use the same approach to look at similar inference problems involving statistics other than the mean, and in particular those called *order statistics*.

Order statistics are obtained by simply ordering a random sample from smallest to largest. Suppose that we have a random sample $X_1, X_2, \ldots, X_n$. We can order the observed values of the random sample from smallest to largest, and denote the sorted values by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where:

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

Prior to observing the random sample, we won't know the values of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, and we won't even know which observation will turn out to be the smallest, second smallest, etc. But for any given set of observations we can calculate the corresponding realizations of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Thus each of these quantities satisfies the definition of a statistic: they are termed the *order statistics* of the sample.

Example: Suppose that we observe the following four numbers: $5, 8, 3, 10$. These would usually be denoted $x_1 = 5$, $x_2 = 8$, $x_3 = 3$, $x_4 = 10$. That is, the subscript i in x_i just denotes the order in which the observations were recorded and does not indicate any ranking. The *order statistics*, however, would be denoted $x_{(1)} = 3$, $x_{(2)} = 5$, $x_{(3)} = 8$, $x_{(4)} = 10$. Here, the subscript (i) indicates that that observation is the i^{th} largest in the sample. The first order statistic $X_{(1)}$ is always the minimum of the sample, that is

$$X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$$

which is why, here, $x_{(1)} = 3$. For a sample of size n the n^{th} order statistic $X_{(n)}$ is always the maximum of the sample, that is

$$X_{(n)} = \min\{X_1, X_2, \ldots, X_n\}$$

so that here $x_{(n)} = 10$. The sample range is the difference between the maximum and minimum, and so is expressed as a function of the order statistics:

$$\text{Range} = X_{(n)} - X_{(1)}$$

Here, the range is clearly the $10 - 3 = 7$.

The "five-number summary" introduced at the start of the first year course consisted of certain order statistics, or averages of pairs of order statistics. Other useful summaries can also be derived from the order statistics, such as range (which we have just seen), or inter-quartile range, which are alternatives to standard deviation as a measure of spread. As with other statistics, such as the sample mean and variance, we need to derive the distributions of the order statistics, if we are to use them for statistical inference. One way of doing this is to use a bootstrap approach, while another way is to use a mathematical analysis. We first consider the bootstrap approach.

Let us return to our earlier example examining the times to payment for the customers of a mining company. The original data, which was shown on page 3, has been sorted from smallest to largest in the table below (note that this is not necessary for any of the bootstrap calculations, but makes it easier to see that the sample median is in fact 38.5 days).

Suppose that there is some concern from management that a few very late customers may be unfairly skewing the average payment times. In such a case, it might be a better approach to examine the *median* payment time, which is resistant to such outliers. We can approach the construction of a bootstrap confidence interval around

34	34	34	35	36	36
37	37	37	38	38	38
38	38	38	39	39	39
40	41	41	41	41	42
42	42	42	43	44	46

the median in much the same way that we did for the mean. That is, we can generate a large number (5000 or whatever) of bootstrap samples as before, and for each bootstrap sample compute the median. We can then sort the bootstrap sample medians from smallest to largest. For one particular run of 5000 bootstrap replications, the following selection of sorted bootstrap medians was obtained

Rank	1	25	125	250	500	1000	2500
Median	36	37	38	38	38	38	38.5
Rank	4000	4500	4750	4875	4975	5000	
Median	39	40	40.5	41	41	42	

We have used these values to graphically show the distribution of bootstrap medians in Figure 1.8.

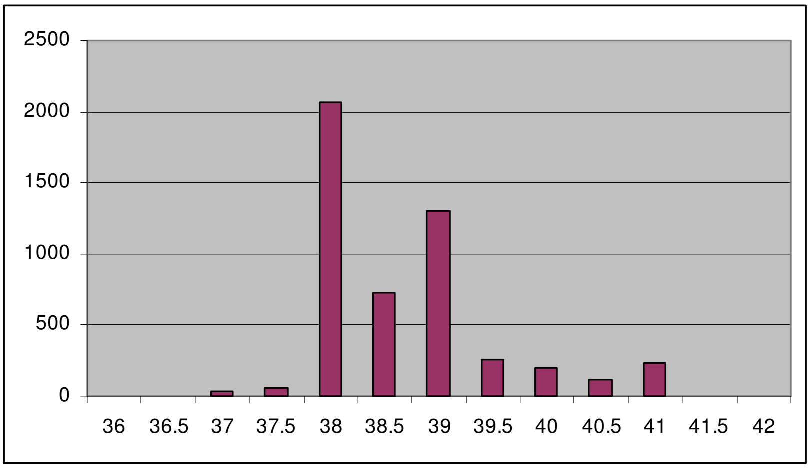


Figure 1.8: Distribution of bootstrapped medians

Once again, we need to remember that these bootstrapped medians in fact refer to sampling errors that might be made in the estimation of the median. Our bootstrap samples were drawn from a population with known median of 38.5. Therefore, for a 95% confidence interval:

- the 125th sorted median of 38 represents an *underestimate* of 0.5 days i.e. a sampling error of $38 - 38.5 = -0.5$.
- the 4875th sorted median of 41 represents an *overestimate* of 2.5 days i.e. a sampling error of $41 - 38.5 = 2.5$.

Under the key assumption that the same sampling errors apply to the originally-taken sample, we can use the above statements about sampling errors to construct a bootstrap confidence interval around the median.

- If the sampling error is as small as -0.5 (i.e. if the true median can be underestimated by 0.5 days), then the population median would be $38.5 + 0.5 = 39$.
- If the sampling error is as large as +2.5 (i.e. if the true median can be overestimated by 2.5 days), then the population median would be $38.5 - 2.5 = 36$.

The 95% (bootstrap) confidence interval for the median is therefore given by 36–39. Note that this is precisely the approach we took in building a bootstrap confidence interval for the mean. All we do here is to use the median in place of the mean. In fact, the 95% confidence interval for the median is slightly wider than the confidence interval for the mean (calculated earlier to be 37.9–40.3). This will generally be the case. It is also quite straightforward to use our bootstrap approach to perform hypothesis tests on the median, although we will not do that here – the details are left to the interested student.

Can we work out a confidence interval for the median without using a bootstrap approach? In fact, it turns out that we can, although the mathematics are somewhat more complicated. The p-th quantile (or, equivalently, the $100p$-th percentile) of the distribution of X defined by $F(x)$ is simply the quantity ξ_p such that:

$$\Pr[X \leq \xi_p] = F(\xi_p) = p.$$

ξ (xi) will be used for quantiles.

In other words, $100p$% of the population of X falls below ξ_p. Special cases are $\xi_{0.5}$, which is the median, $\xi_{0.25}$, which is the lower quartile, and $\xi_{0.75}$, which is the upper quartile. The order statistics can be viewed as estimates of the $1/n$-th, $2/n$-th, etc. quantiles. Our concern here is to obtain a confidence interval for ξ_p for any arbitrary p, based on the sample observations, but not using any assumed distributional properties (apart from assuming that the distribution is continuous). For this purpose, the order statistics are useful. Our aim will be to find two integers ($1 \leq r < s \leq n$) such that:

$$\Pr[X_{(r)} \leq \xi_p < X_{(s)}] \geq 1 - \alpha$$

for some given α. We can't be sure that we can ever find r and s such that the above probability is exactly $1 - \alpha$, which is why we use the $\geq$. We seek the shortest interval for which the above applies.

For any given p, let us define the random variable Y as the number of observations in the sample which do not exceed ξ_p. Since we don't know ξ_p, we can never actually observe Y, but we can still state its probability distribution. For example, we know that the probability of a single observation drawn from the sample being less than the median $\xi_{0.5}$ is by definition 0.5, even if we don't know the value of $\xi_{0.5}$. Therefore the *number of observations* in the sample which are less than $\xi_{0.5}$ is a random variable which follows the binomial distribution (since there are multiple independent trials) with parameters n and $p = 0.5$. Generally, since by definition

$F(\xi_p) = p$, Y has the binomial distribution with parameters n and p (i.e. the p defining the required quantile). Now $\{X_{(r)} \leq \xi_p\}$ is equivalent to $Y \geq r$, while $\{\xi_p < X_{(s)}\}$ is equivalent to $Y < s$. We then have:

$$\begin{aligned} \Pr[X_{(r)} \leq \xi_p < X_{(s)}] &= \Pr[r \leq Y < s] \\ &= \sum_{y=r}^{s-1} (\begin{matrix} n \\ y \end{matrix}) p^y (1-p)^{n-y} \end{aligned}$$

Generally, we have to use trial and error to find values of r and s, as close together as possible, but for which the above expression evaluates to at least $1 - \alpha$. Fortunately, this trial and error is facilitated either by the use of binomial tables or even more easily by using a spreadsheet package like Microsoft Excel, as is illustrated in the next example.

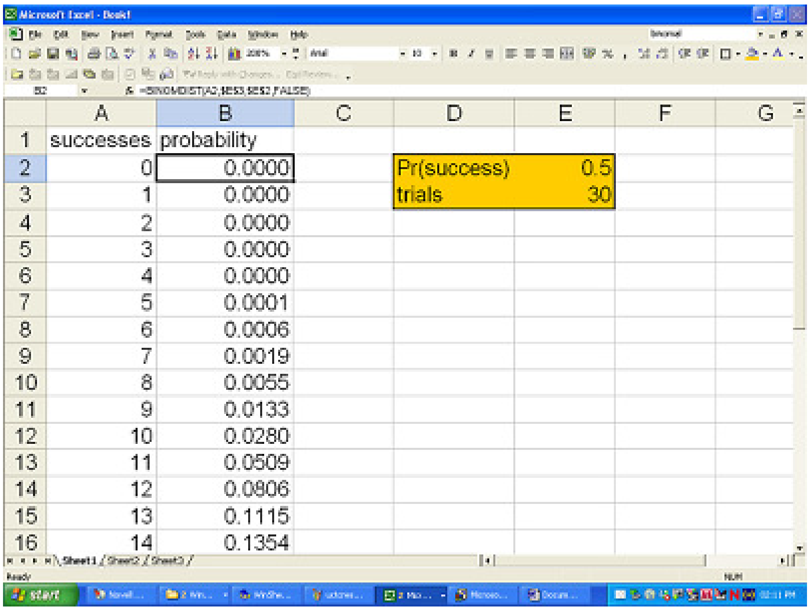
Example: Returning to our example, suppose we wish to find the 95% confidence interval for the median $\xi_{0.5}$ based on our sample, which is of size 30. We set up the following calculations in a spreadsheet environment

- In column A, set up all possible values of x. Here, there can be anywhere between 0 and 30 "successes", so we write the values 0 to 30 in the rows of column A.
- In some cell (we have used cell E2), set up the desired value of p. This is the "probability of success" in the usual calculation of a binomial probability, and here $p = 0.5$ since we are interested in the median.
- In another cell (we have used cell E3), set up the sample size n. This is the "number of trials" in the usual calculation of a binomial probability, and here $n = 30$.
- In column B, calculate the probability of achieving x out of n successes when the probability of an individual success is p by entering the following formula:
  `=BINOMDIST(x,n,p,false)`
  NOTE: the n, p and x's should refer to appropriate cells on the spreadsheet e.g. in cell B2, type `=BINOMDIST(A2,$E$3,$E$2,false)`. Then drag the formula down to compute the other values.

Following the setup of your spreadsheet calculations, the screen should look something like the following.

The entries in column B give the probability of achieving x out of 30 successes with $p = 0.5$. We can now quite easily find values of r and s for which the sum of all the probabilities between r and $s - 1$ evaluates to at least $1 - \alpha$. Here, we find that the entries between 11 ($\Pr(X = 11) = 0.0519$) and 24 ($\Pr(X = 24) = 0.0006$) sum to 0.9505, which is just above 0.95. Thus we have found that $r = 11$ and $s = 25$ and that

$$\Pr[11 \leq Y < 25] = 0.9505$$



B2 =BINOMDIST(A2,E3,E2,FALSE)

	A	B	C	D	E
1	successes	probability			
2	0	0.0000		Pr(success)	0.5
3	1	0.0000		trials	30
4	2	0.0000			
5	3	0.0000			
6	4	0.0000			
7	5	0.0001			
8	6	0.0006			
9	7	0.0019			
10	8	0.0055			
11	9	0.0133			
12	10	0.0280			
13	11	0.0509			
14	12	0.0806			
15	13	0.1115			
16	14	0.1354			

Figure 1.9: Excel: calculating a confidence interval for the population median.

and therefore that $[X_{(11)}\ ;\ X_{(25)}]$ is approximately a 95% confidence interval for $\xi_{0.5}$. This means that we can form a 95% confidence interval for the median by taking the 11th-ranked observation (which turns out to be 38) and the 25th ranked observation (which turns out to be 42) i.e. 38–42. This interval is quite different to the results of our bootstrap experiments (which gave a confidence interval of 36–39), perhaps because of the bimodality of the original sample (see Figure 1.9).

For large values of n, it becomes easier to use the normal approximation to the binomial distribution for Y. In other words, we approximate the distribution of Y by the normal distribution with mean np and variance $np(1-p)$. In moving in this way from a discrete to a continuous distribution, we need to apply the "*continuity correction*". What this in effect means is that we approximate the probability that $Y = i$ (for any integer i) by the probability implied by the normal distribution for the interval $i - \frac{1}{2} < Y < i + \frac{1}{2}$. Thus the event $\{r \leq Y < s\}$ is replaced by $\{r - \frac{1}{2} < Y < s - \frac{1}{2}\}$. In other words, by standardizing the normal distribution in the usual way, we approximate $\Pr[r \leq Y < s]$ by:

$$\Pr\left[\frac{r - \frac{1}{2} - np}{\sqrt{np(1-p)}} < Z < \frac{s - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right]$$

where Z has the standard normal distribution. From normal tables, we can look up the critical value $z_{\alpha/2}$ such that:

$$\Pr[-z_{\alpha/2} < Z < +z_{\alpha/2}] = 1 - \alpha$$

and thus by equating corresponding terms above, we can solve for r and s. These will normally turn out to be fractions, which means that we have to widen the interval by moving out to the next integer values.

Example: Suppose we wish to find a 95% confidence interval for the median (i.e. $\xi_{0.5}$) of a distribution, based on a sample of size 30 (note: a sample size of 30 – as in our payment time example – is probably large enough to start using large sample approximations), but using the normal approximation to the binomial. Thus $np = 15$, $np(1-p) = 7.5$ and $\sqrt{np(1-p)} = 2.74$. We thus need to find r and s such that:

$$\Pr\left[\frac{r-15.5}{2.74} < Z < \frac{s-15.5}{2.74}\right] = 0.95$$

The 2.5% critical value of the standard normal distribution is 1.96. We therefore require:

$$\frac{r-15.5}{2.74} = -1.96$$

which gives $r = 10.1$, and:

$$\frac{s-15.5}{2.74} = 1.96$$

which gives $s = 20.8$. Moving out to the next integers gives $r = 10$ and $s = 21$, and thus the desired confidence interval for the median is $[X_{(10)} \ ; \ X_{(21)}]$. This is reasonably similar to the confidence interval arrived at using the full binomial calculations, which gave $[X_{(11)} \ ; \ X_{(25)}]$. Checking back with the original data, we state that the 95% confidence interval using the normal approximation is given by 38–41.

An important point in both these examples is that we have succeeded in the construction of *distribution free* (i.e. applying no matter what the underlying form of the probability distribution function $F(x)$) confidence intervals for population *quantiles* or *percentiles*. Using a bootstrap approach, we even get the full sampling distribution of these order statistics. However, we can also work out the sampling distributions of order statistics using a more analytical approach, which is useful for some inferential problems: the bootstrap method does not work very well for the extreme quantiles. We will only consider the cases of the smallest and largest order statistics, $X_{(1)}$ and $X_{(n)}$ respectively, although the results can be extended to the intermediate order statistics too.

For ease of notation, let us denote the distribution function of $X_{(r)}$ by $F_{(r)}(x) = \Pr[X_{(r)} \leq x]$, with p.d.f. $f_{(r)}(x)$. The case $r = n$ (the largest order statistic) is particularly simple as the event $\{X_{(n)} \leq x\}$ is simply the event that *all* n observations do not exceed x. Because the initial observations are independent, we can therefore write the distribution function of $X_{(n)}$ as:

$$F_{(n)}(x) = \Pr[X_{(n)} \leq x] = [F(x)]^n \, .$$

The p.d.f. can easily be obtained by differentiation,

$$f_{(n)}(x) = n[F(x)]^{n-1} f(x)$$

The case $r = 1$ (the smallest order statistic) is almost as easy. In this case, we note that the event $\{X_{(1)} > x\}$ is the event that *all n* observations are greater than x. Thus, once again by independence:

$$1 - F_{(1)}(x) = \Pr[X_{(1)} > x] = [1 - F(x)]^n$$

i.e.:

$$F_{(1)}(x) = 1 - [1 - F(x)]^n .$$

Once again the distribution function can be differentiated with respect to x to obtain the p.d.f.

$$f_{(1)}(x) = n[1 - F(x)]^{n-1} f(x)$$

Example: Let X be the lifetime of a single light bulb, which is exponentially distributed with a mean of 2000 hours. Six light bulbs are put into operation together (in some sort of bank of lights). What is:

- The probability that the time until the last bulb fails is greater than 8000 hours, assuming that no bulbs are replaced in the interim?
- The p.d.f. of the time until the last bulb fails?

We are thus concerned with properties of the distribution of $X_{(6)}$. The distribution function is thus:

$$F_{(6)}(x) = [1 - e^{x/2000}]^6$$

Substituting $x = 8000$ gives $F_{(6)}(8000) = 0.8950$, and thus $\Pr[X_{(6)} > 8000] = 1 - 0.8950 = 0.1050$. Thus, the probability that any of the lightbulbs will be working after 8000 hours (remembering that they *each* have an average lifetime of 2000 hours) is 10.5%.

The p.d.f. is:

$$f_{(6)}(x) = \frac{6}{2000} e^{-x/2000} [1 - e^{-x/2000}]^5$$

Tutorial Exercises

1. For each of the two data sets given below:

   (a) Perform 1000 replications of a "bootstrap" simulation of resampling from the population

   (b) Use the simulation to find 95% confidence limits on the sampling error in determining the mean

   (c) Convert the error limit into a confidence interval for the population mean, and compare this with the usual limits based on the t-statistics

   Data Set A: The following data refer to numbers of vehicles arriving at a service station during each 5-minute interval over a total period of one-and-a-half hours:

11	0	2	0	3	4
1	3	10	2	8	6
4	5	0	0	6	1

   Data Set B: Data were collected by a building contractor, in order to improve his tendering strategy. For each of the 25 contracts, the table on the next page shows the contractors own cost estimates and the lowest (winning) bid for the contract. Also shown is the ratio of winning bid to estimated cost. Perform the analysis on this set of ratios only.

2. A random sample of size 10 from an unknown distribution with a mean μ has yielded the following observations:

8.54	2.11	6.78	8.06	9.57
5.52	0.05	5.40	21.89	2.08

   The sample mean and standard deviation were calculated as 7.00 and 6.08 respectively.

   (a) A total of 1000 "bootstrap samples" were generated from this data. Explain what is meant by this assertion.

   (b) The sample averages for each of the 1000 bootstrap samples were calculated, and then ordered from smallest to largest. The following are some of the ordered results obtained:

Sample No.	1	10	25	50	100	250	500
Sample Ave.	2.39	3.30	3.93	4.42	4.86	5.71	6.82
Sample No.	750	900	950	975	990	1000	
Sample Ave.	8.13	9.27	9.96	10.68	11.50	12.78	

   i. Estimate the p-value corresponding to a test of the null hypothesis $\mu \leq 5$ versus $\mu > 5$.

   ii. Construct a 95% confidence interval for μ based on the bootstrap data

Table 1.1: Data Set B

Estimated	Lowest	Ratio
Cost	Bid	
74600	90500	1.213
170600	195100	1.144
94500	101100	1.070
57800	73400	1.270
65400	76300	1.167
127200	133300	1.048
56200	77400	1.377
65400	61500	0.940
50600	67500	1.334
112500	117600	1.045
135200	160200	1.185
77800	92000	1.183
65000	71400	1.098
61900	68500	1.107
148800	196900	1.323
69600	76700	1.102
135200	114700	0.848
77000	79900	1.038
65800	59000	0.897
148700	138200	0.929
127100	109100	0.858
122900	122500	0.997
99900	98700	0.988
70300	85200	1.212

(c) Compare the results from the previous section with that obtained from standard normal theory.

3. A random sample of size 8 from an unknown distribution with a mean μ has yielded the following observations:

0.326	1.463	0.421	0.060
0.038	0.203	0.125	0.182

The sample mean and standard deviation were calculated as 0.352 and 0.437 respectively. The sample averages for each of the 2000 bootstrap samples were calculated, and then ordered from smallest to largest. The following are some of the ordered results obtained:

Sample No.	1	20	50	100	200	500	1000
Sample Ave.	0.093	0.109	0.126	0.141	0.168	0.223	0.343
Sample No.	1500	1800	1900	1950	1980	2000	
Sample Ave.	0.456	0.553	0.665	0.704	0.733	0.950	

(a) What is meant by a "bootstrap" sample?

(b) Estimate the p-value corresponding to a test of the null hypothesis $\mu \leq 0.15$ versus $\mu > 0.15$.

(c) Construct a 95% confidence interval for μ based on the bootstrap data

(d) Compare the results from part (b) and (c) (i.e. both hypothesis test *and* confidence interval) with the results that would be obtained from standard normal theory.

(e) What would you expect to happen to the following quantities as the number of bootstrap samples increases?

i. The minimum bootstrap sample average?

ii. The median bootstrap sample average?

iii. The maximum bootstrap sample average?

4. The probability density function for the random variable X is expressed in the following form:

$$f(x) = \frac{c}{x^2} \quad \text{for } x > 2$$

for some constant c.

(a) Determine the value of c.

(b) The first three of a sequence of uniformly distributed random variables has been generated as follows: 0.883; 0.167; 0.545. Use these to simulate the generation of three corresponding values for X.

5. The distribution function of X is given by:

$$F(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } 0 < x < 1 \\ 1 - \frac{1}{2}(2-x)^2 & \text{for } 1 < x < 2 \end{cases}$$

(a) Calculate and sketch the pdf for X

(b) Suppose that you need to generate a simulated random sample of values from X. Calculate such simulated values corresponding to the following three random numbers generated by the RAND() function in Excel: 0.5924; 0.7374; 0.1504.

6. The probability density function of a random variable X has been stated as follows:

$$f(x) = \begin{cases} x & \text{for } 0 < x < k \\ 2 - x & \text{for } k \leq x < 2 \end{cases}$$

(a) Show that $k = 1$ (3)

(b) Explain how you would generate numbers from the above probability distribution, and illustrate your answer by calculating values of x corresponding to the following three random numbers generated by the RAND() function in Excel: 0.543; 0.344; 0.054

7. Derive the distribution functions of the distributions with the following probability density functions (pdf's):

(a) $f(x) = 6x(1 - x)$ for $0 \leq x \leq 1$

(b) $f(x) = \dfrac{5}{x^6}$ for $x > 1$

8. For the exponential distribution with a mean of 10 (i.e. $\lambda = 0.1$), and for the second distribution defined in the previous question: Simulate the occurrence of 1000 sample values from the distribution, group the results of the simulation into 100 sets of 10 values each, and calculate the corresponding sample means in each set. Plot a histogram of these means. Comment on the results.

9. The probability density function (pdf) of a random variable X is given by: $f(x) = c(1 - x^2)$ for $-1 < x < 1$.

(a) Determine the value of c

(b) Determine the distribution function for X

10. A random sample of size 10 from an unknown distribution with a mean μ has yielded the following observations:

8.54	2.11	6.78	8.06	9.57
5.52	0.05	5.40	21.89	2.08

The sample *medians* for each of the 1000 bootstrap samples were calculated, and then ordered from smallest to largest. The following are some of the ordered results obtained:

Sample No.	1	10	25	50	100	250	500
Sample Ave.	1.08	2.09	2.11	3.74	3.75	5.46	6.15
Sample No.	750	900	950	975	990	1000	
Sample Ave.	7.42	8.06	8.54	8.81	9.05	15.73	

(a) Construct a 95% confidence interval for the median based on the bootstrap data

(b) Use the analytical approach (and binomial tables) to construct a 90% confidence interval for $\xi_{0.6}$

(c) Compare the two intervals obtained in the previous two questions with those obtained using the normal approximation to the binomial distribution (which is not really appropriate, since our n here is not at all large)

11. Let $X_1, X_2, \ldots, X_{15}$ be a random sample from the density function $f(x) = 9xe^{-3x}$ for $x > 0$. What is

(a) The probability that the $X_{(15)} > 2$?

(b) The p.d.f. of the random variable representing the smallest of the 15 observations?

12. Let $T_1, T_2, \ldots, T_8$ be a random sample from the density function $f(x) = 0.75e^{-x} + 0.05e^{-0.2x}$ for $x > 0$. What is

(a) The probability that the $X_{(1)} \leq 3$?

(b) The p.d.f. of the random variable representing the largest observation?

13. For each of the two data sets in Question 1:

(a) Perform 5000 replications of a "bootstrap" simulation of resampling from the population

(b) Use the simulation to find 95% confidence limits on the sampling error in determining the *median*

(c) Convert the error limit into a confidence interval for the population median, and compare this with the limits based on the statistics making use of (i) probabilities obtained using the binomial distribution (ii) the normal approximation to the binomial

(d) Repeat parts (b) and (c), this time for the 35th percentile.

2
Two-sample and three-sample problems

2.1 Distributions of Sample Statistics from Normal Theory

We saw in Chapter 1 how important it is to develop an understanding of sampling variations when, for example, using the sample mean to represent the population mean. We saw further that some such understanding can be obtained by simulations such as the "bootstrap". The Central Limit Theorem enables us to examine the same problems more simply and elegantly, by making use of the properties of the normal distribution.

Suppose again that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution (or population) with mean μ and variance σ^2, and let $\bar{X}$ be the sample mean. We know that if we repeated the sampling many times, we would obtain different values for $\bar{X}$. We want to know how variable these sample means are, in order to be able to draw conclusions from the currently observed mean. The CLT tells us that for "large enough" n, the distribution of $\bar{X}$ under such repeated sampling is "approximately" normal with mean μ and variance σ^2/n. In fact, for practical purposes, the approximation can be pretty good for quite moderate n (perhaps as low as 20), except for very heavy-tailed distributions.

Define z_α to be upper 100α percentile of the standard normal distribution, i.e. such that $\Pr[Z > z_\alpha] = \alpha$ when Z has the standard normal distribution.

Provided that the population variance is known, we can use the CLT and normal tables to address the following questions regarding the population mean:

Confidence Intervals: Since

$$\Pr\left[-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right] = 1-\alpha$$

we are $100(1-\alpha)\%$ "confident" that

$$-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2},$$

or (re-arranging terms) that

$$\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}.$$

Hypothesis Tests: Suppose we wish to test a conjecture (alternative hypothesis) $H_1 : \mu > \mu_0$ (for some specified number μ_0) against the null hypothesis $H_0 : \mu \leq \mu_0$. We would regard a large $\bar{X}$ as evidence against H_0. Now if H_0 is true, then for any $0 < \alpha < 1$ it follows that:

$$\Pr\left[\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha\right] \leq \alpha$$

for the stated μ_0. Recall that we may either set the desired significance level *a priori* (e.g. something like the conventional 0.05, giving $z_{0.05} = 1.645$) and reject H_0 if $\bar{X} > \mu_0 + z_\alpha \sigma/\sqrt{n}$; or we could look up the p-value: $P(\bar{X}-\mu_0)/(\sigma/\sqrt{n} \geq t_{observed})$.

In this course (at least in these first 2 chapters) we will stay away from significance levels and dichotomous decisions (reject/not reject the null hypothesis) as far as possible, and we will rather see the p-value as only a part of the information which needs to be interpreted in the wider context of the problem (what else do we know? what did we expect? how large was the sample? how many other tests did we conduct? what did other studies show?).

We note again that the above probabilities refer to sampling variability in $\bar{X}$ for given μ and σ. They are not probabilities on μ or on the truth of H_0.

But what do we do if the population variance is unknown (which would seem to be the rule rather than the exception)? To get some sense of the problem, we carry out the following exercise:

- Generate a large number of random values (say 45000, as we shall want the number to be divisible by 9), drawn from a normal distribution with mean 0 and variance $\sigma^2 = 9$.

```
x <- rnorm(45000,0,3)
hist(x)
```

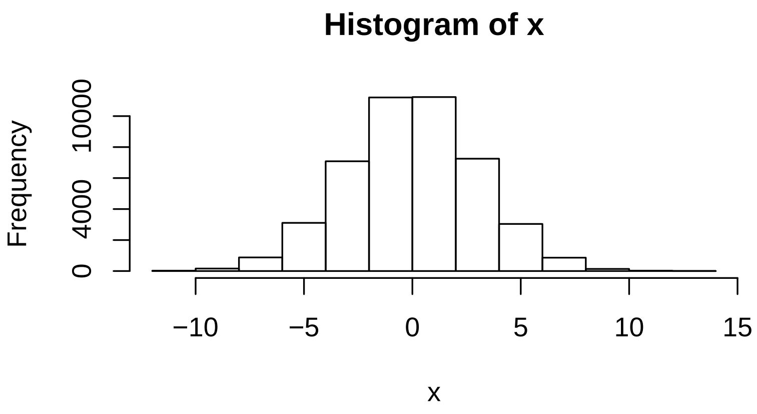


Figure 2.1: Distibution of simulated values

- Cluster the results into groups of 9 each, and obtain the sample mean and sample variance within each group

```
xmat <- matrix(x,nrow=5000,ncol=9)
mymean <- apply(xmat,1,mean)
myvar <- apply(xmat,1,var)
```

- The group sample means should have the standard normal distribution (why?). We check this by plotting a histogram of the sample means. This shows how sample means do vary from sample to sample, as we can view each group as a sample of size 9.

```
hist(mymean,main="Histogram of sample means")
```

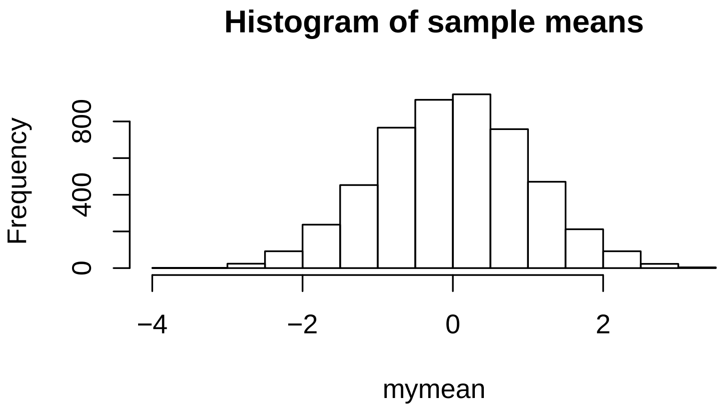


Figure 2.2: Distibution of means from constructed samples of size 9. We know that these means should be distributed around the value of zero that we used to generate the random numbers.

- If we had not known the population variance (σ^2), we would have had to estimate it by the sample variance S^2. We would expect, of course, that S^2 would vary around the true value of 9. Below we plot a histogram of the sample standard deviations S (which should vary around 3)

```
hist(sqrt(myvar),main="Histogram of sample std devs")
```

- If we used data from any one group to draw inferences about the true mean, then our analyses would need to be based on the standardized form: $T = \bar{X}/(S/3)$. We calculate these values for each group, and plot the corresponding histogram.

```
myt <- mymean/(sqrt(myvar)/3)
hist(myt,main="Histogram of sample t-stats")
# overlay a scaled N(0,3) dbn for comparison
xn <- seq(-8,8,length=1000)
yn <- 5000*dnorm(xn,0,3)
lines(xn,yn,col="red")
```

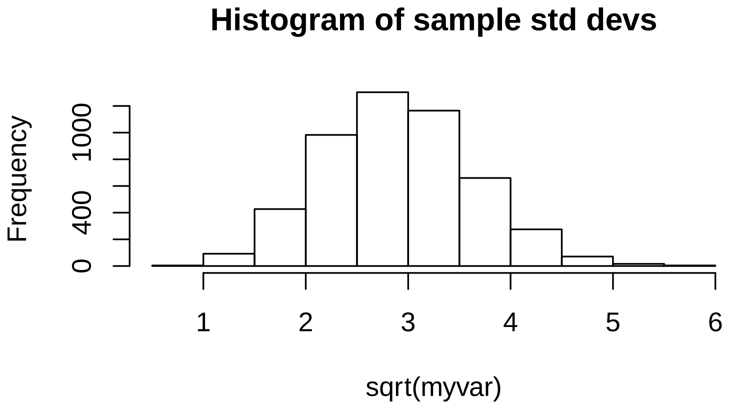


Figure 2.3: Distibution of standard deviations from constructed samples of size 9. We see that the sample standard deviation estimates range from considerably below to considerably above the true value of 3.

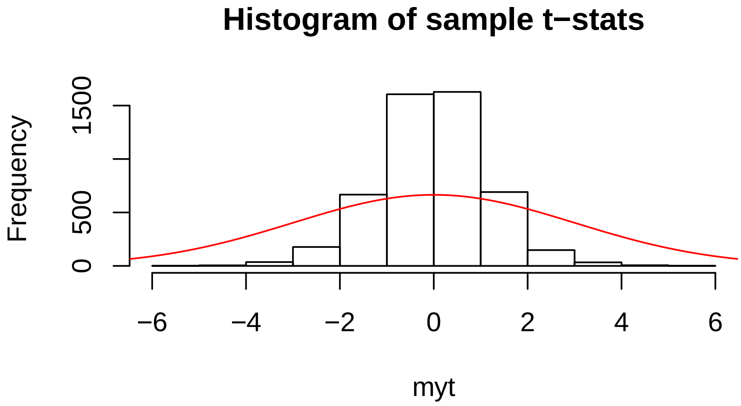


Figure 2.4: Distibution of t-statistics from constructed samples of size 9 (histogram), with a scaled N(0,3) distribution shown for comparison. Note that to get the N(0,3) onto the same scale as the histogram, we have to multiply the p.d.f. at each point by 5000 (the number of values used to create the histogram).

- Note how the histogram of T differs from a normal distribution; it has much higher kurtosis. The reason for this is that large values of T can derive from two sources, viz. large values of $\bar{X}$ *or* small values of S

At first sight, the distribution of the T-values in Figure 2.4 may appear fairly normal. Closer examination, however, reveals the following features:

- Only 27 out of 5000 sample means (5.4%) exceeded an absolute value of 2.5 (compared to the theoretical probability of 1.24% for the standard normal distribution), while 76 of the 5000 T-values (15.2%) exceeded an absolute value of 2.5 (compared to a theoretical probability of 3.70% for the t-distribution with 8 degrees of freedom).

- Only 4 out of 5000 sample means (0.8%) exceeded an absolute value of 3 (compared to the theoretical probability of 0.27% for the standard normal distribution), while 41 of of the 5000 T-values (1.72%) exceeded an absolute value of 3 (compared to a theoretical probability of 1.71% for the t-distribution with 8 degrees of freedom).

- 0 out of 5000 sample means exceeded 3.5 in absolute value (the largest value being 3.25). On the other hand, 41 of the 5000 T-values exceeded 3.5 in absolute value, with a maximum absolute value of 5.23.

- The sample estimate of the kurtosis for the T's was 4.0.

The above points indicate that the distribution of the T's has a distinct tendency towards having heavier tails than the normal distribution. This can seriously bias estimates of p-values or confidence intervals, if the sample estimate is used in place of the true population mean, but critical values from the normal distribution are still used.

We can deal with the additional variation due to using sample estimates of the variance in quite a simple manner. We simply express the "t-statistic" in the following way:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right] \left[\frac{\sigma}{S}\right] = Z\frac{\sigma}{S}.$$

We know that $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has the standard normal distribution. The usual sample variance estimator is given by

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

so that U defined by:

$$U = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left[\frac{X_i - \bar{X}}{\sigma}\right]^2$$

has the χ^2 distribution with $n-1$ degrees of freedom. We can thus express T in the form:

$$T = \frac{Z}{\sqrt{U/(n-1)}}$$

where we know the distributions of Z (standard normal) and U (χ^2 with $n-1$ degrees of freedom).

The precise derivation of the distribution of T from the distributions of Z and U requires a bit of intricate mathematics, which will be omitted here. We state, however, the following theorem which is expressed in a slightly more general form.

Theorem 1. *Let Z and U be independent random variables, having the standard normal distribution and the χ^2 distribution with d degrees of freedom respectively. Define:*

$$T = \frac{Z}{\sqrt{U/d}}.$$

Then the probability density function for T is given by:

$$f(t) = \frac{\Gamma((d+1)/2)}{\sqrt{\pi d}\Gamma(d/2)} \cdot \frac{1}{[1+t^2/d]^{(d+1)/2}}$$

The distribution defined by the p.d.f. defined in Theorem 1 is called the t-distribution with d degrees of freedom, tables for which are widely available. Values for the cumulative distribution can be obtained from Excel's TDIST() spreadsheet function. Excel also provides the TINV() function for the inverse function.

The value of Theorem 1 is that it allows the statistician to discover many sampling situations in which the distribution of relevant statistics follows a t-distribution. One can then derive p-values or confidence intervals just by using tables of the t-distribution. The trick is to recognize which is the normally distributed and which the χ^2 distributed random variables, as illustrated in the following examples.

Example: Suppose that X, Y and Z are independent random variables, all normally distributed with zero means and with standard deviations 3, 4 and 5 respectively. Thus $X/3$, $Y/4$ and $Z/5$ have the standard normal distribution. Furthermore:

$$\frac{Y^2}{16} + \frac{Z^2}{25}$$

has the χ^2 distribution with 2 degrees of freedom. The theorem thus tells us that:

$$\frac{X}{3\sqrt{\left(\frac{Y^2}{16} + \frac{Z^2}{25}\right)/2}} = \frac{\sqrt{2}X}{3\sqrt{\frac{Y^2}{16} + \frac{Z^2}{25}}}$$

has the t-distribution with 2 degrees of freedom. From tables, therefore, we could conclude that:

$$\Pr\left[\frac{\sqrt{2}X}{3\sqrt{\frac{Y^2}{16} + \frac{Z^2}{25}}} \geq 2.920\right] = 0.05$$

i.e.:

$$\Pr\left[X \geq 6.194\sqrt{\frac{Y^2}{16}+\frac{Z^2}{25}}\right] = 0.05.$$

Example: Suppose that $X_1, X_2, \ldots, X_8$ is a random sample from $N(0, 4\sigma^2)$, while $Y_1, \ldots, Y_5$ is a random sample from $N(0, \sigma^2)$. The two samples are independent of each other. What is the probability that $\bar{X}$ exceeds $0.5\sqrt{\sum_{i=1}^{5} Y_i^2}$?

We know that $\bar{X}$ has the normal distribution with mean 0 and a variance of $4\sigma^2/8 = 0.5\sigma^2$. Thus $\sqrt{2}\bar{X}/\sigma$ has the standard normal distribution.

Furthermore, $\sum_{i=1}^{5} Y_i^2/\sigma^2$ has the χ^2 distribution with 5 (not 4!) degrees of freedom, and is independent of $\bar{X}$ (as it is independent of all the X_i's). We therefore conclude that:

$$W = \frac{\sqrt{2}\bar{X}/\sigma}{\sqrt{\sum_{i=1}^{5} Y_i^2/5\sigma^2}} = \sqrt{10}\frac{\bar{X}}{\sqrt{\sum_{i=1}^{5} Y_i^2}}$$

has the t-distribution with 5 degrees of freedom.

The probability that $\bar{X}$ exceeds $0.5\sqrt{\sum_{i=1}^{5} Y_i^2}$ is given by:

$$\Pr\left[\frac{\bar{X}}{\sqrt{\sum_{i=1}^{5} Y_i^2}} > 0.5\right] = \Pr[W > 0.5\sqrt{10} = 1.581].$$

This last probability we can look up in t-tables, and it turns out to be 8.74%.

It is useful to record here another probability distribution which arises naturally from functions of random variables having the χ^2 distribution, especially (for example) in problems involving the comparison of variances.

Theorem 2. *Let U and V be independent random variables, having χ^2 distributions with p and q degrees of freedom respectively. Define:*

$$Y = \frac{U/p}{V/q}.$$

Then the probability density function for Y is given by:

$$f(y) = \frac{\Gamma((p+q)/2)\,(p/q)^{p/2}}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{y^{(p/2)-1}}{[1+py/q]^{(p+q)/2}}$$

The distribution defined by the p.d.f. defined in Theorem 2 is called the F-distribution with p and q degrees of freedom (sometimes referred to as the numerator and denominator degrees of freedom respectively). You have already met the F-distribution in tests for equality of variances in a number of contexts. Values for the cumulative distribution can be obtained from Excel's FDIST() spreadsheet function. Excel also provides the FINV() function for the inverse function.

In the remainder of this chapter, we return to a number of classical sampling situations, in order to examine how both bootstrapping and the theoretical results above contribute to our understanding of sampling variation.

2.2 *Two-Sample Problems*

Let us return to the cash flow problem illustrated in Chapter 1, but now suppose that we wish to compare delays in invoice payments in two different markets, say the copper and zinc markets. Once again, suppose that 30 recent deliveries in each market have been analyzed to determine numbers of days between invoicing and receipt. (At this stage, equal sample sizes for the two populations is not essential, but we shall keep to these for illustration). The data were recorded as follows:

Copper	39	35	44	33	19	6	27	24	40	13
	35	34	34	33	61	56	43	19	34	40
	34	28	29	45	46	28	41	46	25	17
Zinc	39	32	22	32	55	34	46	37	31	45
	41	40	46	66	36	42	37	43	35	62
	48	47	34	42	43	33	47	41	34	41

We note that the average delay time[1] in the Zinc market sample is 7.43 days larger than that in the Copper market. The questions which arise are (1) how large is the true difference in population means? (2) Is the evidence for a larger delay in the Zinc market convincing in the light of sampling variation.

The first question can be structured in terms of a confidence interval for the true difference in means. The second can be formulated in hypothesis testing terms, viz. to test the "null hypothesis" H_0 that the means are equal.

The hypothesis test is easily addressed in a bootstrapping manner. Suppose that we re-formulate the null hypothesis simply as: "The distributions of times in the two populations are identical". If this H_0 is true, then both sets of 30 observations come from the same population; in other words all 60 observations arise from the same population. We may then simulate the two-sample results as follows:

- Place all 60 observations in a "hat" and "shuffle"
- Draw a sample of size 60 with replacement, and split arbitrarily into two sets of 30 each
- Calculate the sample mean in each set, and the difference between the two sample means
- Repeat as often as needed, to obtain a distribution of the differences in means

The above process is easily implemented by using the macro in the BootStrap.xls spreadsheet package. Simply enter and highlight all 60 values in row 6, and press Ctrl-b. The two sample means can be calculated from the first and last 30 columns of the results. The process is also easily programmed in R, using a small extension to the code we used in Chapter 1.

[1] The mean of the numbers for copper is 33.60; the standard deviation is 12.09. The mean of the numbers for zinc is 41.03; the standard deviation is 9.08

```
# load the payment delay data
cop_pds <- c(39,35,44,33,19,6,27,24,40,13,35,34,34,
33,61,56,43,19,34,40,34,28,29,45,46,28,41,46,25,17)
zinc_pds <- c(39,32,22,32,55,34,46,37,31,45,41,40,
46,66,36,42,37,43,35,62,48,47,34,42,43,33,47,41,34,41)
# put all data together (in a hat)
pdelays2 <- c(cop_pds,zinc_pds)
# shuffle - this step is not needed (why?)
pdelays2 <- sample(pdelays2, size=60, replace=FALSE)
# set up a variable to store the bootstrap samples
all_boots <- matrix(NA,nrow=5000,ncol=60)
for(i in 1:5000){
  # draw a single bootstrap sample from pdelays2
  boot <- sample(pdelays2,size=60,replace=TRUE)
  # store that bootstrap
  all_boots[i,] <- boot
}
```

We can now extract the bootstrap means for "group 1"[2] using the `apply` function, but now *only applied to the first 30 columns* of `all_boots`.

```
bs_means1 <- apply(all_boots[,1:30],1,mean)
```

Similarly we extract the bootstrap means for "group 2" by using the `apply` function on the last 30 columns of `all_boots`.

```
bs_means2 <- apply(all_boots[,31:60],1,mean)
```

Finally, we can compute the difference between the two group means in each of our 5000 bootstraps samples, and plot the histogram of these differences:

```
# difference in means (note the order is arbitrary)
bs_diffs <- bs_means1 - bs_means2
hist(bs_diffs)
```

In the 5000 repetitions reported here, the difference in means exceeded +7.43 on 34 occasions[3], and was less than -7.43 on 21 occasions[4]. In other words the absolute magnitude of the difference exceeded the originally observed difference on 55 out of 5000 occasions, around 1%. This would give a p-value appropriate to a two sided test of approximately 0.01, so we would conclude that the difference in means is significant.

The same bootstrap results can, with a little thought, also be used to construct a confidence interval for the true difference in population means. We sort the 5000 differences and find the 125th and 4875th smallest difference (corresponding to the 2.5 and 97.5 percentile respectively of the distribution shown in Figure 2.5)

[2] Although note that referring to group 1 and group 2 in the context of the bootstrap is rather meaningless, since we have already randomly shuffled all 60 observations.

[3] In R, `sum(bs_diffs > 7.43)`

[4] In R, `sum(bs_diffs < -7.43)`

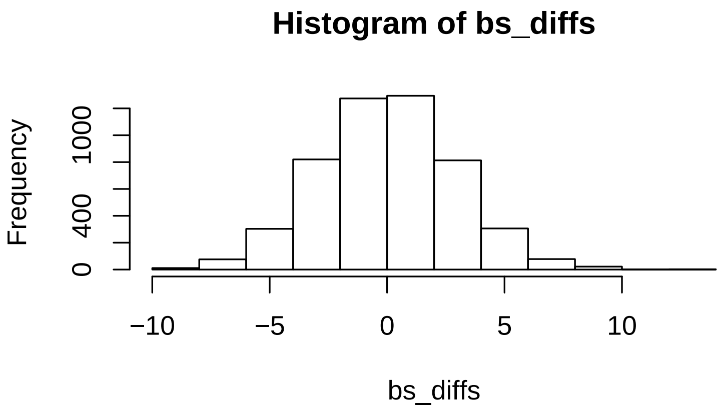


Figure 2.5: Distribution of bootstrapped difference in sample means

```
sorted_bs_diffs <- sort(bs_diffs)
# 2.5 percentile
sorted_bs_diffs[125]

[1] -5.6

# 97.5 percentile
sorted_bs_diffs[4875]

[1] 5.7
```

Since by construction, the bootstrap sampling was from a distribution having a true difference in means of zero, we can conclude as follows:

- With 95% "confidence", the errors (defined as the deviation between the observed and true differences in means) will lie between -5.6 and +5.7;
- Since the actually observed difference in means was -7.43, an error of -5.6 must mean that the true population difference is -1.83 ($= -7.43 + 5.6$) – Note the direction of the signs!;
- Similarly, an error of 5.7 means that the true population difference must be -13.13 ($= -7.43 - 5.7$).

A 95% confidence interval could thus be stated as [-13.13 ; -1.83], based on the bootstrap results.

Now, how may we use the central limit theorem and the theoretical results for sampling from a normal distribution to obtain corresponding solutions to the same problems, perhaps more easily?

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be random samples from populations with means and variances μ_X, σ_X^2, μ_Y and σ_Y^2 respectively. By the CLT, $\bar{X}$ and $\bar{Y}$ are approximately normally distributed with means μ_X and μ_Y, and variances σ_X^2/m and σ_Y^2/n respectively.

Our usual assumption is that the samples are independent, so that $\bar{X} - \bar{Y}$ is also approximately normal with mean $\mu_X - \mu_Y$, and variance $\sigma_X^2/m + \sigma_Y^2/n$ (Note the sum!).

If z_α is the critical point of the normal distribution such that $\Pr[Z > z_\alpha] = \alpha$, then:

- Under a null hypothesis that $\mu_X = \mu_Y$, the appropriate two-sided test would be based on:

$$\Pr\left[\frac{|\bar{X} - \bar{Y}|}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} > z_\alpha\right] = 2\alpha$$

  (so that for a significance level of 5% we would need $\alpha = 0.025$);

- Under a null hypothesis that $\mu_X \leq \mu_Y$ (implying *a priori* information or judgement that only situations in which $\mu_X > \mu_Y$ are important or interesting), the appropriate one-sided test would be based on:

$$\Pr\left[\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} > z_\alpha\right] = \alpha$$

  (since any observations in which $\bar{X} < \bar{Y}$, no matter how large the difference, are consistent with the null hypothesis).

This immediately solves the problems of both hypothesis tests and confidence intervals, provided that the two variances σ_X^2 and σ_Y^2 are known. But again we have the problem of unknown variances. For the two sample problem, it is useful to distinguish at least three separate cases.

Equal variances: Suppose that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, say. The standardized difference can thus be written as:

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

which would approximately normally distributed with mean 0 and variance 1.

To the same levels of approximation:

$$\frac{\sum_{i=1}^m (X_i - \bar{X})^2}{\sigma^2} \quad \text{and} \quad \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2}$$

have χ^2 distributions with degrees of freedom $m - 1$ and $n - 1$ respectively, so that their sum has the χ^2 distribution with $m + n - 2$ degrees of freedom. This sum can be expressed as:

$$\frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2} = \frac{(m + n - 2)S_{pool}^2}{\sigma^2}$$

where S_{pool}^2 is the usual "pooled" variance estimate.

It therefore follows that:

$$\frac{\bar{X} - \bar{Y}}{S_{pool}\sqrt{\frac{1}{m} + \frac{1}{n}}} = Z \cdot \frac{\sigma}{S_{pool}}$$

must have (approximately) the t-distribution with $m+n-2$ degrees of freedom, so that we can use t-tables in place of normal tables for carrying out hypothesis tests or constructing confidence intervals.

Recall that under the same assumption of equal variances, the ratio:

$$\frac{\sum_{i=1}^{m}(X_i-\bar{X})^2/[(m-1)\sigma^2]}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2/[(n-1)\sigma^2]}=\frac{n-1}{m-1}\frac{\sum_{i=1}^{m}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}=\frac{S_X^2}{S_Y^2}$$

has the F-distribution with $m-1$ and $n-1$ degrees of freedom. Similarly:

$$\frac{m-1}{n-1}\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sum_{i=1}^{m}(X_i-\bar{X})^2}=\frac{S_Y^2}{S_X^2}$$

has the F-distribution with $n-1$ and $m-1$ degrees of freedom. These two facts can be used to test the hypothesis of equal variances using the appropriate F-test. In our current example, $s_Y^2=12.09^2$ and $s_X^2=9.08^2$, so that the ratio of the two is $F=1.77$. We can compare this to an F-distribution with $m-1=29$ and $n-1=29$ degrees of freedom, and obtain a p-value = 0.065 (`=FDIST(1.77,29,29)` in Excel, `pf(1.77, 29, 29, lower.tail = FALSE)` in R). This p-value should leave us a bit undecided, there is perhaps some evidence that the variances are not equal, but this is not very strong.

A little further thought will make it clear that we could also use a bootstrap approach to test the hypothesis of equal variances. Suppose we again re-formulate the null hypothesis simply as: "The distributions of times in the two populations are identical". We may then perform the same bootstrap sampling as for the two-sample test of means, except that we compute the ratio of variances after each replication:

- Place all 60 observations in a "hat" and "shuffle"
- Draw a sample of size 60 with replacement, and split arbitrarily into two sets of 30 each
- Calculate the sample *variance* in each set, and the ratio of the two variances
- Repeat as often as needed, to obtain a distribution of the ratio of variances.

We can again use either the R code of a few pages above, or the BootStrap.xls spreadsheet package (this is left to you as an exercise!). A histogram of ratios of sample variances obtained using this procedure, based on 5000 repetitions, is shown in Figure 2.6. In fact, you can see for yourself that this (empirical) distribution closely resembles the F-distributions of your earlier courses.

In this particular experiment, the originally observed F-ratio of 1.77 was exceeded 500 times out of 5000 repetitions. Under H_0, such an occurrence would therefore appear to be fairly likely

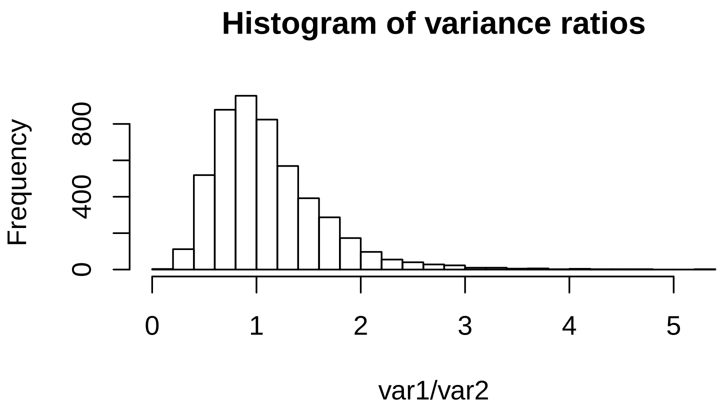


Figure 2.6: Bootstrap distribution of ratios of two sample variances

(p=0.1), and we might say that we don't have strong evidence for a difference in variance. Note that we cannot say that we have evidence that the variances are equal!! This is consistent with the results obtained using standard normal theory, although the p-value is somewhat larger (0.1 vs. 0.065)

Paired observations: Suppose that we are able to pair up the X_i and Y_j observations in some way. This does of course mean that $m = n$. In some cases the pairing will be natural, or even forced upon us, for example when X_i and Y_i are results of two different treatments applied to the same subject. In other cases, where the X_i and Y_j observations are entirely independent of each other, but sample sizes are the same, they may be paired in random fashion.

Now define $W_i = X_i - Y_i$. The W_i are independent, with zero mean under the null hypothesis, but with unknown variance ($=$ $\sigma_X^2 + \sigma_Y^2$, but we never really need to know the individual variances). We can then simply apply the single sample procedure to test the hypothesis that $\mu_W = 0$. This approach brings with it an added bonus!. As long as the $X_1, X_2, \ldots, X_n$ are independent of each other, and the $Y_1, Y_2, \ldots, Y_n$ are independent of each other, *it does not matter if X_i and Y_i are associated for the same i*! The unknown variance of W will be given by $\sigma_W^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$, but once again we do not need separately to identify the components. In order to apply the one sample analysis to the W_i, we simply need to estimate the unknown variance σ_W^2 by the sample variance of the W_i.

The Behrens-Fisher Problem: We now turn to the general case in which $\sigma_X \neq \sigma_Y$ and $m \neq n$. We could of course pair up what we can and throw away any unmatched observations. But discarding sample information does not seem to be good statistics.

The obvious alternative would be to calculate separate sample variances S_X^2 and S_Y^2, and to base inferences on an expression of the form:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/m + S_Y^2/n}}.$$

But what is the probability distribution of this ratio? Struggle as we may, it turns out to be impossible to massage the above expression into anything to which Theorem 1 may be applied, so that no exact t-distribution result can be derived (a problem first identified by Behrens and Fisher). Nevertheless, from many simulation studies it has emerged that the required distribution can be approximated by a t-distribution with fractional "degrees of freedom" given by

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}\right]} - 2$$

Note that the definition of the p.d.f. of the t-distribution in Theorem 1 does not mathematically require d to be integer.

2.3 *One-Way Analysis of Variance*

Analysis of Variance (ANOVA) problems can be seen as a generalization of the two sample problem to many samples. Consider for example the following data on traffic counts (vehicles per hour) at each of five intersections (which may, for example, be suggested sites for a new petrol station). Figure 2.7 gives a box-and-whisker plot for the same data, comparing the five samples.

Place	Hourly Traffic Counts									
I	344	382	353	395	207	312	407	421	366	222
II	365	391	538	471	431	450	299	371	442	343
III	261	429	402	391	239	295	129	301	317	386
IV	422	408	470	523	398	387	433	440		
V	367	445	480	323	366	325	316	381	407	339

The question at this stage is whether the population means (the true long-run means at each intersection) do differ significantly (or could the differences displayed in Figure 2.7 simply be due to chance?). We could apply the two-sample approach to all pairs of intersections, but this quickly becomes messy, and it sometimes happens that no single pairwise difference is significant even though other tests (see below) indicate that *some* difference occurs *somewhere*. It is useful, therefore, to base our analysis on some more aggregated summary statistics.

Define Y_{ij} as the j-th observation from set (or "treatment" or "population") i. Suppose that there are k different populations (so that $k = 5$ in the above example), and that there are n_i observations from population i. (In the above example, $n_1 = n_2 = n_3 = n_5 = 10$, while $n_4 = 8$.) Further define $N = \sum_{i=1}^{k} n_i$, i.e. the total number

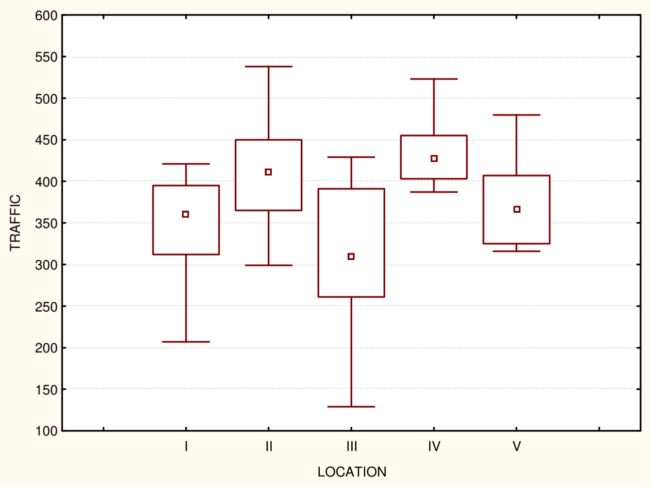


Figure 2.7: Box and whisker plots of 5 samples of traffic counts

of observations. Let μ_i denote the population mean for population i, and suppose that the sampling variances are the same in each population ($= \sigma^2$).

We now define the following summary statistics:

- Sample mean for population i: $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$
- Overall sample mean: $Y_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}/N$, which can also be written in the form:

  $$\sum_{i=1}^{k} \frac{n_i}{N} Y_{i\cdot}$$

  i.e. as a weighted average of the sample means for each population.
- Error sum of squares: $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i\cdot})^2$
- "Treatment" sum of squares: $SST = \sum_{i=1}^{k} n_i (Y_{i\cdot} - Y_{\cdot\cdot})^2$.

We note that $SSE/(N-k)$ is an unbiased estimator of the sampling variance σ^2 (which is a general result, not dependent upon any assumptions of normality). Under the null hypothesis (H_0) that all means are equal, say $\mu_1 = \cdots = \mu_k = \mu$, $Y_{i\cdot}$ has mean μ and variance σ^2/n_i, so that $\sqrt{n_i}(Y_{i\cdot} - \mu)$ has mean 0 and variance σ^2. It thus follows that:

$$\mathrm{E}\left[\frac{\sum_{i=1}^{k} n_i (Y_{i\cdot} - \mu)^2}{k}\right] = \sigma^2.$$

It can once again be shown that replacement of the population mean μ by the sample mean $Y_{\cdot\cdot}$ "loses" one degree of freedom, so that:

$$\mathrm{E}\left[\frac{\sum_{i=1}^{k} n_i (Y_{i\cdot} - Y_{\cdot\cdot})^2}{k-1}\right] = \mathrm{E}\left[\frac{SST}{k-1}\right] = \sigma^2$$

so that if the null hypothesis is true, then $SST/(k-1)$ is *also* an unbiased estimator of σ^2.

Overall, this implies that if H_0 is true, then the ratio:

$$F = \frac{SST/(k-1)}{SSE/(N-k)}$$

should not significantly deviate from 1; but if there are any deviations from H_0 (i.e. if one or more means differ from others), then the above ratio will tend to be larger than 1. In the case of the traffic count data above, the F-ratio turns out to be 4.56.

For any given set of sample data, we can easily calculate the observed value for F. If this is less than 1, then there is clearly no evidence to support any difference between the means of the populations. If $F > 1$, we need (as usual) to ask whether the deviations can be due to chance.

Once again, we can get an answer to this last question by means of a bootstrap simulation. In this case, the procedure would be as follows:

- Place all N observations in a "box" and "shuffle"
- Draw (with replacement) a sample of size N, and divide these arbitrarily into k samples of sizes $n_1, n_2, \ldots, n_k$ respectively.
- Calculate the SSE and SST, and the F-ratio as defined above.
- Repeat as many times as desired, to obtain a distribution of the F-ratio under H_0
- Compare the originally observed ratio with the empirical distribution

This process can again be implemented using R and/or the bootstrap macro in the BootStrap.xls spreadsheet package[5]. Results from 5000 repetitions for the traffic count data are displayed in Figure 2.8.

In this experiment, the originally observed F-ratio of 4.56 was exceeded only 15 times out of 5000 repetitions. Under H_0, such an occurrence would appear to be highly unlikely (p=0.003), evidence that the means differ.

Once again, we ask whether a similar conclusion can be reached without such heavy computation. We now need to make the stronger assumption that the individual Y_{ij} values are normally distributed. If this is (at least approximately) true, then the $Y_{i.}$ must also be normally distributed. The following properties then apply:

- For each population i

$$\sum_{j=1}^{n_i} \frac{(Y_{ij} - Y_{i.})^2}{\sigma^2}$$

has the χ^2 distribution with $n_i - 1$ degrees of freedom.

[5] This is left to you as one final exercise for the chapter!

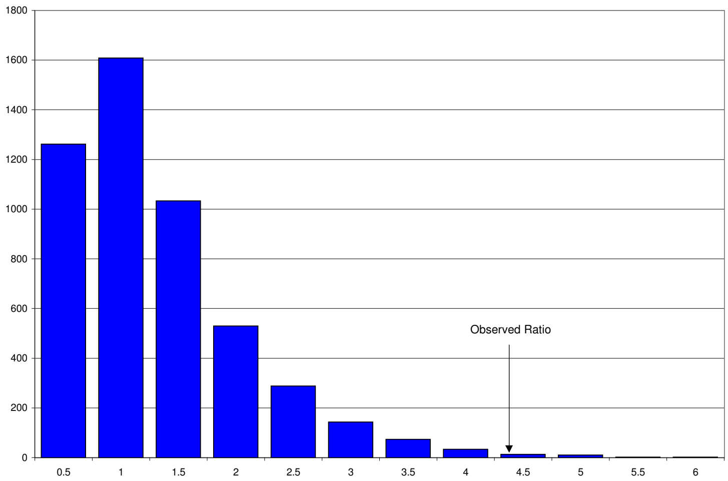


Figure 2.8: Bootstrap distribution of F-ratios for traffic count data

- Since observations from the different populations are assumed to be independent, it follows that

$$U = \frac{SSE}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(Y_{ij} - Y_{i\cdot})^2}{\sigma^2}$$

  has the χ^2 distribution with $\sum_{i=1}^{k}(n_i - 1) = N - k$ degrees of freedom.

- If H_0 is true, then the $Y_{i\cdot}$ have the same mean and variances σ^2/n_i. By a slight variation of previous results, it can be shown that

$$V = \frac{SST}{\sigma^2} = \sum_{i=1}^{k} \frac{(Y_{i\cdot} - Y_{\cdot\cdot})^2}{\sigma^2/n_i}$$

  has the χ^2 distribution with $k-1$ degrees of freedom.

- It can also be shown that U and V are independent.

- The F-ratio can thus be expressed as

$$\frac{V/(k-1)}{U/(N-k)}$$

  which by Theorem 2 has the F-distribution with $k-1$ and $N-k$ degrees of freedom.

We may thus use F-tables to reach a conclusion as to whether the observed F-ratio is too large to be due to chance under H_0. In our traffic count example, we had $k = 5$ and $N = 48$. From F-tables we can find that the 1% critical value with 4 and 43 degrees of freedom is 3.79. Since the observed $4.56 > 3.79$, we know that the p-value < 0.01. Even better, we can find the exact p-value using R or Excel (pf(), FDIST(), repectively): 0.0037, which is close to our Bootstrap approximation.

Tutorial Exercises

1. For the example discussed at the start of Section 2.2 (for the problem of comparing payment delays in two markets), compare the results based on the two-sample t-test with pooled variance, with those reported in the notes based on bootstrapping approach. Remember to first check the hypothesis of equal variances, using the F-test, before applying the above t-test.

2. Relationships between normal, χ^2 and t distributions:

   (a) X has the normal distribution with mean 0 and variance 4; V has the χ^2 distribution with 5 degrees of freedom. X and V are independent random variables. Making use of t-tables, obtain an approximate value for the probability that $X^2 > V$.

   (b) $X_1, X_2, \ldots, X_{10}$ is a random sample of values from a normal distribution with mean 0 and variance 9. $Y_1, Y_2, \ldots, Y_{16}$ is a random sample of values from a normal distribution with mean 20 and variance 1. Let $\bar{X}$ and $\bar{Y}$ be the corresponding sample means. Find the value z such that:

   $$\Pr\left[\frac{\bar{X}^2}{\sum_{i=1}^{16}(Y_i - \bar{Y})^2} \geq z\right] = 0.05$$

3. The following are independent random variables:

   - X: Normal with mean 0 and variance 10;
   - Y: Normal with mean 5 and standard deviation 3
   - U: χ^2 with 10 degrees of freedom

   Answer the following:

   (a) Identify the distribution of $X - Y$

   (b) For what values of k and c does $U + k(X - Y + c)^2$ have a χ^2 distribution? What degrees of freedom does it have?

   (c) Find the relevant constants which would make the following probability statements true:

   i. $\Pr\left[\frac{Y-5}{\sqrt{0.1X^2 + U}} > k\right] = 0.05$

   ii. $\Pr\left[\frac{X+Y+c}{\sqrt{U}} > k\right] = 0.95$

   iii. $\Pr\left[\frac{(X-Y-c)^2}{U} \leq k\right] = 0.9$

   iv. $\Pr\left[\frac{X^2/10 + (Y-5)^2/9}{U} \leq k\right] = 0.99$

4. In a study of fuel consumption with two different fuel additives, 20 new cars were selected at random. Tests were conducted under identically controlled conditions, with 10 of the cars using petrol containing additive A, and 10 using petrol containing additive B. The sample means and standard deviations of consumption expressed as litres per 100km for the two groups were recorded as follows:

	Additive	
	A	B
mean	6.89	7.19
std.dev.	0.374	0.475

A bootstrap simulation was performed in which 1000 data sets of size 20 were regenerated from the original data, and split into two groups of size 10. The resultant differences in means (A-B) from the two groups were sorted from smallest to largest. A summary of the differences observed at various positions in the sorted list are given as follows:

Position No.:	1	10	25	50	75
Value	-0.776	-0.477	-0.385	-0.308	-0.260
Position No.:	100	250	500	750	900
Value	-0.234	-0.127	0.008	0.131	0.249
Position No.:	925	950	975	990	1000
Value	0.274	0.319	0.387	0.469	0.585

(a) Use the bootstrapped data to estimate the p-value for the test of differences between the means. Compare this with the corresponding value obtained on the assumption that the data are normally distributed, and that the group variances are the same.

(b) Use the bootstrapped data to construct a 95% confidence interval for the difference between the means.

5. For each of the data sets attached to the end of this chapter, use bootstrapping and standard normal theory to test the hypothesis of no differences between the groups.

6. The following data represent numbers of defects observed in product produced from three assembly lines for electronic equipment. Each count represents the number of defects in one hour of operation, where the hours selected for inspection were chosen randomly. The question of interest is whether defect rates differ between assembly lines.

Line 1	Line 2	Line 3
6	34	13
38	28	35
3	42	19
17	13	4
11	40	29
30	31	0
15	9	7
16	32	33
25	39	18
5	27	24

Using a simulation ("bootstrap") approach, estimate the relevant significance level, and compare this with that obtained from the standard analysis of variance (F) test, and the non-parametric (Kruskal-Wallis) test, both of which can be obtained through *Statistica*. Since the data are counts, they are unlikely to be normally distributed. From the results of this exercise can you comment on the robustness of the usual ANOVA F-test?

7. Independent random variables X, Z and W have the following distributions:

   - X is normally distributed with a mean of 10 and a standard deviation of 5;
   - Z has the standard normal distribution;
   - W has the χ^2 distribution with 9 degrees of freedom.

   (a) For what values of a and b will $\frac{a(X+Z-b)}{\sqrt{W}}$ have a t-distribution? How many degrees of freedom will it have?

   (b) Use tables to find the value of β such that

   $$\Pr\left[\frac{(X-10)^2+25Z^2}{W} \geq \beta\right] = 0.05.$$

8. As part of an investigation into differences between fuel consumptions on different trucks in a transport fleet, total fuel consumptions (in litres) over the same fixed route were measured a number of times for each of three trucks. Results obtained were as follows:

Truck	Consumptions (Litres)						$Y_{i\cdot}$	$\sum_{j=1}^{n_i}(Y_{ij} - Y_{i\cdot})^2$
A	35.6	37.1	32.6	31.3	32.4		33.80	23.780
B	34.5	34.2	32.5	30.5			32.93	10.168
C	36.6	33.9	32.5	35.5	35.6	37.5	35.27	16.453

   (a) Motivate and explain how a bootstrapping approach might be used to test the hypothesis of no differences between mean fuel consumptions of the three trucks.

   (b) Such a bootstrapping approach was applied, based on 5000 bootstrap replications. For each replication, the SSE, SST and the ratio SST/SSE were calculated. (NOTE: The ratio has not been adjusted for degrees of freedom.) The SST/SSE ratios were sorted from smallest to largest, and the following are a selection of the observed values:

Sample No.	2500	4000	4500	4750	4875	4950
SST/SSE:	0.078	0.181	0.263	0.352	0.474	0.672

   What conclusion should be drawn about any differences between the trucks?

   (c) Compare the above answer with that obtained from the standard normal theory approach to ANOVA.

Data Sets for Exercises

Data Set C: A press used to remove water from copper-bearing materials is being tested using two different types of filter plates. These data are obtained on the percentage of moisture remaining in the material after treatment.

Regular chamber (I)			Diaphragm chamber (II)		
8.10	8.16	8.16	7.58	7.65	7.69
7.96	7.98	7.93	7.66	7.67	7.67
7.97	8.08	8.06	7.58	7.62	7.65
8.02	7.87	7.94	7.65	7.58	7.71
7.82	8.11	7.92	7.63	7.54	
8.15	7.91	8.00	7.46	7.40	

Data Set D: It is thought that the gas mileage obtained by a particular model of automobile will be higher if unleaded premium gasoline is used in the vehicle rather than regular unleaded gasoline. To gather evidence to support this contention 10 cars are randomly selected from the assembly line and tested using a specified brand of premium gasoline; 10 others are randomly selected and tested using the brand's regular gasoline. Tests are conducted under identical controlled conditions. These data result:

Premium		Regular	
35.4	31.7	29.7	34.8
34.5	35.4	29.6	34.6
31.6	35.3	32.1	34.8
32.4	36.6	35.4	32.6
34.8	36.0	34.0	32.2

Data Set E: A study of visual and auditory reaction time is conducted for a group of college basketball players. Visual reaction time is measured by time needed to respond to a light signal and auditory reaction time is measured by time needed to respond to the sound of an electric switch. Fifteen subjects were measured with time recorded to the nearest millisecond.

Subject	Visual	Auditory
1	161	157
2	203	207
3	235	198
4	176	161
5	201	234
6	188	197
7	228	180
8	211	165
9	191	202
10	178	193
11	159	173
12	227	137
13	193	182
14	192	159
15	212	156

Is there evidence that the visual reaction time tends to be slower than the auditory reaction time?

Data Set F: A firm has two possible sources for its computer hardware. It is thought that supplier X tends to charge more than supplier Y for comparable items. Do these data support this contention at the $\alpha = 0.05$ level?

Item	X price ($)	Y price ($)
1	6 000	5900
2	575	580
3	15000	15000
4	150000	145000
5	76000	75000
6	5650	5600
7	10000	9975
8	850	870
9	900	890
10	3000	2900

Data Set G: Twenty randomly selected cars of the same make and model were split into two groups of ten each. Premium grade petrol was used in cars from the first group and regular grade in the other group. Petrol consumptions over a standard set of identically controlled conditions were measured as follows:

Premium		Regular	
6.71	7.49	8.00	6.82
6.88	6.71	8.02	6.86
7.52	6.73	7.40	6.82
7.33	6.49	6.71	7.29
6.82	6.60	6.99	7.38

Data Set H: These are the running times in minutes of films produced by two different directors. Is there a difference?

Director I	103	94	110	87	98		
Director II	97	82	123	92	175	88	118

Data Set J: Ten samples of dried milk produced by Company A were analyzed for fat content by the company's own laboratory, and by the laboratory of their main customer (Company B). As each pair of analyses relate to the same original sample, they are not independent. We must therefore use a paired test in both the simulation and the *t*-test.

Sample Number	Analysis by Company A	Analysis by Company B
1	0.50	0.79
2	0.58	0.71
3	0.90	0.82
4	1.17	0.82
5	1.14	0.73
6	1.25	0.77
7	0.75	0.72
8	1.22	0.79
9	0.74	0.72
10	0.80	0.91

Data Set K: A study on the tensile strength of aluminium rods is conducted. Forty identical rods are randomly divided into four groups each of size 10. Each group is subjected to a different heat treatment and the tensile strength, in thousands of pounds per square inch, of each rod is determined. The following data result.

Treatment			
1	2	3	4
18.9	18.3	21.3	15.9
20.0	19.2	21.5	16.0
20.5	17.8	19.9	17.2
20.6	18.4	20.2	17.5
19.3	18.8	21.9	17.9
19.5	18.6	21.8	16.8
21.0	19.9	23.0	17.7
22.1	17.5	22.5	18.1
20.8	16.9	21.7	17.4
20.7	18.0	21.9	19.0

Data Set L: Following a major accidental spill from a chemical manufacturing plant near a river, a study was conducted to determine whether certain species of fish caught from the river differ in terms of the amounts of the chemical absorbed. If differences are found, regulations on human consumption may be recommended. Samples from catches of three major species were measured in parts per million. The resulting data are given below.

Species		
A	B	C
18.1	29.1	26.6
16.5	15.8	16.1
21.0	20.4	18.8
18.7	23.5	25.0
7.4	18.5	21.8
12.4	21.3	15.4
16.1	23.1	19.9
17.9	23.8	15.5
	20.1	21.1
	11.9	25.5

Data Set M: Four brands of tyres are tested for tread wear. Since different cars may lead to different amounts of wear, cars are considered as blocks to reduce the effect of differences among cars. An experiment is conducted with cars considered as blocks, and brands of tyres randomly assigned to the four positions of tyres on the cars. After a predetermined number of miles driven, the amount of tread wear (in millimetres) is measured for each tyre. The resulting data are given below.

Car	Tyre brand			
	A	B	C	D
1	8.9	6.6	5.6	4.2
2	7.2	6.9	7.3	6.9
3	3.1	6.2	7.2	4.1
4	7.1	8.3	6.3	5.8
5	6.7	6.4	5.9	9.4
6	5.3	6.7	8.0	7.9
7	2.4	5.5	6.1	3.1
8	5.7	9.2	9.6	4.2