Chapter II
Simulation methods in for classical statistics

Introduction

Classical statistical theory contains many methods for testing hypotheses in numerous different situations.

Derivation of these tests can be difficult or impossible in some cases and often relies on asymptotic results or approximations.

If the test we wish to perform is non-standard then deriving a suitable test procedure may not be possible (or we may have forgotten the correct test!).

In this Chapter we consider what can be done using simulation methods.

2.1 Monte Carlo tests

Recap of hypothesis testing framework

Suppose that we have a null hypothesis H_0 represented by a completely specified model and that we wish to test this hypothesis using data $X_1, \ldots, X_n$. We proceed as follows

1. Assume H_0 is true.
2. Find a test statistic $T(X_1, \ldots, X_n)$ for which large values indicate departure from H_0.
3. Calculate the theoretical sampling distribution of T under H_0.
4. The observed value $T_{obs} = T(x_1, \ldots, x_n)$ of the test statistic is compared with the distribution of T under H_0. Either
   - (Neyman-Pearson) reject H_0 if $T_{obs} > c$. Here c is chosen so that $\mathbb{P}(T \geq c | H_0) = \alpha$ where α is the **size** of the test, i.e., $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) = \alpha$.
   - (Fisher) compute the p-value $p = \mathbb{P}(T \geq T_{obs} | H_0)$ and report it. This represents the strength of evidence against H_0.

Example 1: normal parametric test

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Suppose that $\sigma^2 = 1$ is known. Consider the null hypothesis

$$H_0 : \mu = 0.$$

1.
2.
3.
4.

It may not be possible to derive the sampling distribution of T under H_0.

- T is not some fairly simple function,
- or if $X_1, \ldots, X_n$ are not independent samples from the population of interest (dependent data are common in real problems).

Moreover, in deriving the distribution of T, we assume that n is large, equal variances, normality etc. If these assumptions don't hold then our distribution for T will be incorrect.

Monte Carlo Tests

We may not know the distribution of T under H_0, but often it is possible to simulate from the model to produce sample data sets

$$\{X_1^{(i)}, \ldots, X_n^{(i)}\}$$

for $i = 1, \ldots, m-1$.

From these we can calculate $m-1$ sample values of the statistic under H_0,

$$\{T_1, \ldots, T_{m-1}\}$$

We can then estimate the distribution of T under H_0 from this sample and can estimate the critical value c or the p-value by a Monte Carlo approximation, i.e., estimate $\mathbb{P}(T > T_{obs}|H_0)$ by

$$\frac{1}{m-1} \sum_{i=1}^{m-1} \mathbb{I}_{T_i > t_{obs}}.$$

Monte Carlo Testing Algorithm

1. Generate $m-1$ sample test statistics $t_1, \ldots, t_{m-1}$ according to H_0.
2. For a test of size α, define $k = m\alpha$. If t_{obs} is one of the k^{th} largest values in $\{T_1, \ldots, T_{m-1}, t_{obs}\}$ then reject H_0.

i.e. reject H_0 if $t_{obs} > T_{(m-k)}$

where $T_{(1)}, \ldots, T_{(m)}$ are the order statistics of $T_1, \ldots, T_{m-1}, t_{obs}$.

Example: normal parametric test revisited

In this simple case we know the distribution of T under H_0, but it is informative to consider the Monte Carlo test.

1. Generate 1000 samples of size n from a $N(0, \sigma^2)$ distribution and calculate T.

```
t.sample <- c()
for(i in 1:999){
    temp <- rnorm(n=n, mean=0, sd=sigma)
    t.sample[i] <- mean(temp)
}
z.sample <- t.sample*sqrt(n)/sigma
z <- c(z.sample, 2)
```

add observation $Z_{obs} = 2$ to simulated data

For $\alpha = 0.05$, we find the 95th percentile of the sampling distribution

```
c<-quantile(z, 0.95)
```

Then we compare c with the observed value of 2. I found $c = 1.67$ so we would reject H_0 at the 5% level.

If instead, we wanted to estimate the p-value $\mathbb{P}(T \geq T_{obs}|H_0)$ we could estimate it using the R command

```
sum(z>2)/1000
```

For my implementation I found a p-value of 0.028 which again suggests we should reject H_0 at the 5% level.

Note that this is a random test: if we repeat it multiple times we will get a slightly different answer each time.

Example 2: Chi-squared tests

Exam grades are to be compared between 16 boys and 19 girls in a single class. The data are

	A	B	C	D
boys	3	4	5	4
girls	8	8	3	0

The null hypothesis is that there is no difference between boys and girls in exam performance.

In other words, a girl and boy chosen at random have the same probability of obtaining any particular grade.

To apply the standard chi-squared test in this case we would calculate the table of expected values under H_0 and then calculate the test statistics

Calculating for the data we find $T_{obs} = 7.907$, which has a p-value of 0.048.

However, as a rule of thumb, to use the χ^2 test, the expected number of counts in each cell should be at least 5. In this case, 4 of the 8 values are less than 5, which means the assumptions used to show that T has a χ^2 distribution with 3 degrees of freedom do not hold.

Consider using a Monte Carlo test to perform the test.

1. Under H_0, probabilities of obtaining each grade are given by the estimates

	A	B	C	D
probability	$\frac{11}{35}$	$\frac{12}{35}$	$\frac{8}{35}$	$\frac{4}{35}$

2. We then generate a new set of results for boys and girls; the boys' results are sampled from a Multinomial$(16, \frac{11}{35}, \frac{12}{35}, \frac{8}{35}, \frac{4}{35})$ distribution, and the girls' from a Multinomial$(19, \frac{11}{35}, \frac{12}{35}, \frac{8}{35}, \frac{4}{35})$. An example is shown below:

	A	B	C	D
boys	3	5	6	2
girls	4	5	7	3

3. Calculate T for these data, which for this simulated dataset is 5.323.

We then repeat $m-1$ times to get $T_1, \ldots, T_{m-1}$

We rank $T_1, \ldots, T_{m-1}$ together with the observed value of the test statistic T_{obs}. In R we generate 99 test statistics and find 75 to be less than T_{obs}, and 24 to be greater. In this case the null hypothesis is not rejected at the 5% level.

Notice that this is a different conclusion to that reached using the χ^2 test.

In this case, the Monte Carlo test should preferred as we are working with the true distribution of the test statistic and not an approximation.

In general, when conducting hypothesis tests we do not have to be so reliant on distributional approximations, and **we should always consider the option of working with exact distributions**.

Example 3: Testing for randomness in spatial patterns

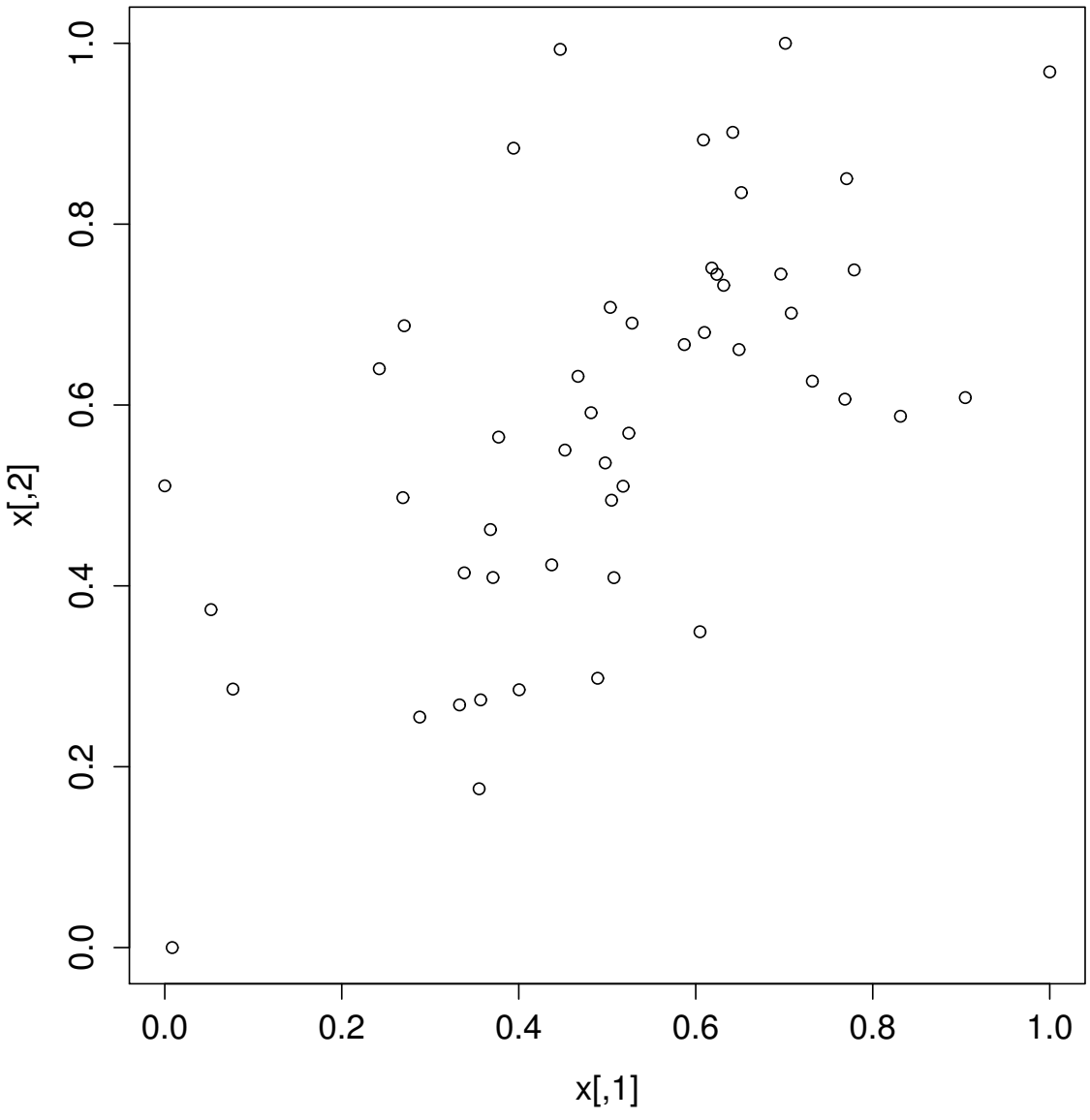
H_0: The spatial locations of each subject are randomly distributed over the unit square: both coordinates have $U[0, 1]$ distributions.

Various possibilities for test statistic. We will consider *nearest-neighbour* of each subject.

Let d_i denote distance from subject i to next closest subject, and define the test statistic T to be

$$T = \left(\sum_{i=1}^{50} d_i \right)^{-1}. \tag{2}$$

If locations are clustered, nearest neighbours will be small $\Rightarrow T$ will be large.


x[,2]
0.0
0.2
0.4
0.6
0.8
1.0
x[,1]
0.0
0.2
0.4
0.6
0.8
1.0

Under H_0, don't know theoretical sampling distribution of T. Straightforward to simulate values of T under H_0, so can estimate the critical values (such as the 95th percentile) we need for the hypothesis test.

1. Generate locations (x, y) of each subject by sampling x and y independently from $U[0, 1]$.
2. For each subject, find the closest observation and measure the distance to it to obtain the nearest-neighbour distance for that observation
3. Take the reciprocal of the sum of the 50 nearest-neighbour distances to get T_i.

Given a sample $T_1, \ldots, T_{m-1}$, we then rank $T_1, \ldots, T_{m-1}$ and the observed T_{obs} in order to give $T_{(1)}, \ldots, T_{(m)}$. For a test of size 5%, if T_{obs} is one of the $0.05 \times m$ largest values, then H_0 is rejected.

p-values

We can estimate the p-value $\mathbb{P}(T \geq T_{obs}|H_0)$ of a Monte Carlo test by looking at the number of observations greater than T_{obs}

$$\hat{p} = \frac{1}{m}\left(\sum_{i=1}^{m-1} \mathbb{I}_{T_i \geq T_{obs}} + 1\right)$$

Exercise: If $p = \mathbb{P}(T \geq T_{obs}|H_0)$ show that $\hat{p} \geq \frac{1}{m}$ and

$$\sum_{i=1}^{m-1} \mathbb{I}_{T_i \geq T_{obs}} \sim Bin(m-1, p)$$

so that the estimate $\hat{p}$ has expectation

$$\mathbb{E}(\hat{p}) = p + \frac{1-p}{m}$$

and is therefore a biased estimator of p. Note that for large m the bias is small.

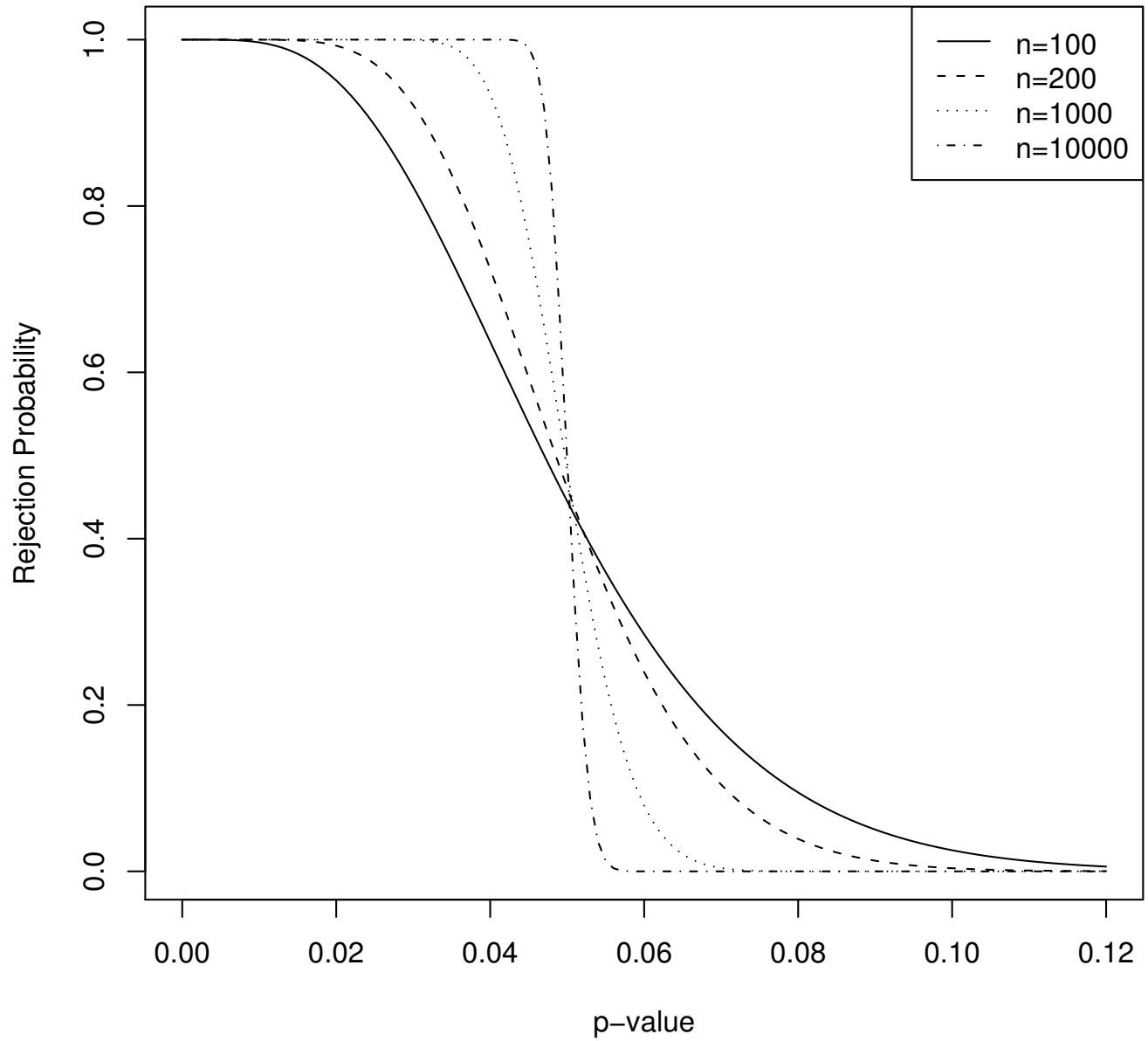
How large should m be?

We need to choose m sufficiently large so that the random sample $T_1, \ldots, T_{m-1}$ allows us to estimate the critical region to a sufficient degree of accuracy.

The Monte Carlo test has a random critical point and so 'blurs' the critical region.

- We reject H_0 if T_{obs} is one of the $k-1$ largest values in $\{T_1, \ldots, T_{m-1}\}$, where $m = k\alpha$.
- If p is the true p-value then we reject H_0 with probability

$$\begin{aligned} R(p) &= \sum_{r=0}^{k-1} \binom{m-1}{r} p^r (1-p)^{m-r-1} \\ &= \mathbb{P}(\text{Bin}(m-1, p) \leq k-1) \end{aligned}$$


n=100
n=200
n=1000
n=10000
Rejection Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
p−value

$R(p)$ can be interpreted as the proportion of times the Monte Carlo test will reject H_0 when we observe T_{obs}.

For p-values smaller than 0.05 we want $R(p)$ to be large and for p-values greater than 0.05 we want $R(p)$ to be small. We choose m to make this so.

We conclude from the figure that a sample size of $m = 100$ is usually acceptable as long as the results aren't interpreted too rigidly.

Of course, this is only an issue if generating test statistics requires substantial computational effort. If it is trivial to generate sample test statistics (which it is in all but the most complex of cases), then a much large value of m can be used.

2.2 Randomisation Tests

Monte Carlo tests allowed us to do hypothesis tests when the null hypothesis specified a complete distribution for the data, e.g., $H_0 : X_i \sim N(0, 1)$.

We now consider a second technique known as **randomisation tests** for deriving the sampling distribution of the test statistic, where no distributional assumptions about the data are required.

The general scenario under consideration is that of an investigation into whether or not a particular treatment/covariate/factor has an effect on some response.

Our aim is to test this without fully specifying a distribution for the data.

Example 1: Cholesterol data

A small study was conducted to investigate the effect of diet on cholesterol levels. Volunteers were randomly allocated to one of two diets, and cholesterol levels were recorded at the end of the trial period

Diet A	233	291	312	250	246	197	268	224
Diet B	185	263	246	224	212	188	250	148

The interest is in whether or not there is a significant difference between the mean cholesterol levels for the two groups. The null hypothesis is

$$H_0 : \text{mean cholesterol levels with the diets are equal}$$

A standard classical analysis of this data might be to assume

$$X_i^{(j)} \sim N(\mu_j, \sigma^2)$$

for $i = 1, \ldots, 8$ and $j = 1, 2$ with σ^2 an unknown common variance.

The standard test is then a two sample t-test, based on the statistic

$$T = \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{\sqrt{s^2/8 + s^2/8}}, \tag{3}$$

where s^2 is the pooled estimate of variance.

Then under H_0 (and assuming normality of the data!), the test statistic T has a t-distribution with 14 degrees of freedom.

For this data, the observed test statistic T_{obs} is 2.0034 with a p-value of 0.0649 for a two-sided test.

But what if we want to analyse the data **without assuming normality**? E.g., because the sample sizes are small

Randomization tests can be used to find a distribution for T without making any distributional assumptions about the data.

If H_0 is true, then any difference in the two sample means would be solely due to how the 16 individuals were assigned to the two groups. So if H_0 is true, what is the probability of observing a sizeable difference between the two group means?

It must be equal to the probability of assigning the individuals to the two groups in such a way that the imbalance occurs, as long as the individuals were assigned to the two groups at random in the actual study. This is the principle idea behind randomisation tests.

Randomisation Test

1. Suppose the 16 individuals in the study have been labelled

Diet A	1	2	3	4	5	6	7	8
Diet B	9	10	11	12	13	14	15	16

2. Randomly re-assign the 16 individuals to the two groups.
3. Re-calculate the test-statistic for this permuted data
4. Repeat 2 and 3 to obtain B sampled test-statistics, denoted $T_1, \ldots, T_B$.
5. For a two-sided test, the estimated p-value of the observed test statistic T_{obs} is

$$\frac{1}{B}\sum_{i=1}^{B} \mathbb{I}_{|T_i| \geq |T_{obs}|}$$

Using 10000 random permutations gave a p-value of 0.063.

Equivalent test statistics

The significance level of T_{obs} is determined using

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} I\{|T_i| \geq |T_{obs}|\}.$$

Notice that multiplying T_{obs} and all T_is by some constant would have no effect on significance level; ordering would be preserved. An **equivalent test statistic** is one that preserves ordering and hence does not change the p-value. In the example, an equivalent test statistic would be

$$T = \bar{X}_1 - \bar{X}_2. \tag{4}$$

ie no need to compute the denominator in Equation (3).

Exact randomisation tests

Could consider systematically *every* possible permutation, rather than a random sample of permutations to determine the significance level.

- This is known as an **exact** randomisation test or **permutation** test
- Can be computationally demanding/impracticable if number of possible permutations is large.
- A large sample of random permutations should be sufficient.

Outliers

In parametric tests, outlying observations in the data can cause problems.

- In the comparison of means problem, an outlier can increase the difference $\bar{X}^{(1)} - \bar{X}^{(2)}$ and will inflate the within group variance.
- Consequently the true significance of the test statistic may be underestimated.

In a randomisation test, you are comparing the relative size of the observed test statistic to its value under alternative random permutations.

Hence, the outlier will not have the same effect.

Example 2

This is illustrated in some data from a study reported in Ezinga (1976) for two treatments A and B:

A	0.33	0.27	0.44	0.28	0.45	0.55	0.44	0.76	0.59	0.01
B	0.28	0.80	3.72	1.16	1.00	0.63	1.14	0.33	0.26	0.63

The sample group means are $\bar{X}_A = 0.412$ and $\bar{X}_B = 0.995$, and the observed test statistic for a two sample t-test is $T = 1.78$.

For a two-tailed test this gives a p-value of 0.11, so not significant at the 5% level. Using a randomisation test, T is now significant at the 5% level with a p-value of about 0.03.

Exercise: Check this conclusion in R

Example 3: Analysis of Variance

Randomisation tests are applicable in many different contexts. Analysis of variance is another example. Below are responses measured on four treatment groups:

Group A	-0.10	-1.10	0.74	-3.80	
Group B	0.94	-0.30	0.67	0.86	1.19
Group C	-0.25	0.84	0.04	0.25	
Group D	0.99	0.08	0.98	0.75	0.53

Test the null hypothesis

$$H_0 : \text{all four groups have equal means}$$

Qn: What classical hypothesis test would you use?

Alternatively, a randomisation test could be applied:

1. Randomly re-assign the observations to the four treatments, keeping the numbers in each treatment the same.
2. Evaluate the test statistic

$$F = \frac{(\sum_{i=1}^{4} n_i(\bar{x}_i - \bar{x})^2)/3}{(\sum_{i=1}^{4}\sum_{j=1}^{n_i}(x_{i,j} - \bar{x}_i)^2)/14}$$

   for the permuated data.
3. Repeat steps 1 and 2 B times to obtain sampled test statistics $F_1, \ldots, F_B$.
4. Estimate the significance level of F_{obs} by

$$\frac{1}{B}\sum_{i=1}^{B} \mathbb{I}_{F_i \geq F_{obs}}.$$

Based on a sample size $B = 10000$, the estimated p-value for F_{obs} was 0.03, suggesting slightly stronger evidence against the null hypothesis (compared with the parametric test).

Example 4: One-sample randomisation tests

Randomisation tests can be used for one-sample problems, but under stricter assumptions. This is demonstrated with the following example:

Given observations

$$\{10.61, 9.46, 7.02, 11.68, 9.58, 11.96, 11.28, 7.63, 6.42, 8.85\}$$

drawn from some population with mean μ, test the null hypothesis

$$H_0 : \mu = 10.$$

It is not immediately obvious what can be permuted here. However, supposing the two following assumptions hold:

- Each observation has been sampled randomly from its population
- The population distribution is symmetric about its mean.

Now suppose H_0 is true, and consider randomly sampling a value X from the population, and then evalutating $Y = X - 10$. If the population distribution is symmetric about 10, then Y must have an equal probability of being positive or negative.

In this example, subtracting 10 from each observation and taking the resulting mean gives a sample mean of -0.551. We will use the absolute value of this sample mean as the test statistic, so $T_{obs} = 0.551$ (for a two-sided test).

$$T = \left| \frac{1}{B} \sum_{i=1}^{B} Y_i \right|$$

If H_0 is true, and both assumptions hold, then the observed sample mean could simply be due to an imbalance of positive and negative Y values. This can be tested as follows:

Fisher's Randomisation test

1. Subtract hypothesised population mean from each observations:

$$\{0.61, -0.54, -2.98, 1.68, -0.42, 1.96, 1.28, -2.37, -3.58, -1.15\}$$

2. Calculate the observed test statistics: $T_{obs} = 0.551$
3. With probability 0.5 for each observation, change the sign of $X - \mu$. E.g.

$$\{-0.61, -0.54, -2.98, -1.68, 0.42, 1.96, -1.28, -2.37, -3.58, -.15\}$$

4. Re-calculate the test-statistic for the new simulated observations: $T = 0.951$.
5. Repeate 3 and 4 to obtain B sampled test-statistics $T_1, \ldots, T_B$.
6. Estimate the significance of T_{obs} by $\frac{1}{B} \sum_{i=1}^{B} \mathbb{I}_{|T_i| \geq |T_{obs}|}$

With $B = 10000$, the estimated significance of T_{obs} is 0.4021.

Using a conventional t-test, the significance of T_{obs} is 0.3982, so there is close agreement between the two methods in this example.

Example 5

Two treatments A and B, unknown population means μ_A and μ_B.

treatment A	130	119	119	168	130
treatment B	154	115	169	137	186

Consider $H_0 : \mu_B - \mu_A = 20$.

How would we test this using a randomisation test?

Cannot just permute data and evaluate difference between the sample means, as population means not equal under H_0.

Suppose we were to add 20 to each observation in group A. Under H_0, what is the expectation of {20+ a response in group A}?

If H_0 is true, then this expectation is $20 + \mu_A = \mu_B$.
Adding 20 to each response in group A:

treatment A+20	150	139	139	188	150
treatment B	154	115	169	137	186

Under H_0 groups have equal population means. We can now use randomisation test in the usual way.

Summary of randomisation tests

Some argue that randomisation tests should always be used, as samples of data are never truly randomly drawn from the population of interest; some members of the population are always going to be more accessible than others.

On the other side, there is no theory to show that the results of a randomisation test can be generalised to the whole population; evidence against the null hypothesis is obtained for the observed sample only.

Consequently, in either case, a 'non-statistical' judgement has to made; that the sample can be treated as effectively random for a conventional test, or that the results can be generalised to the population for a randomisation test.

Two advantages of randomisation tests are that they can be used for any test statistic (i.e. in cases when it is not possible to analytically derive the distribution of the test statistic), and that we don't have to assume a particular distribution for the data.

Note that in most of the examples, almost identical results were obtained using the two methods. In this case, the randomisation test could be seen as a means of supporting the results from the parametric test.

The requirement for the randomisation test to be valid is that the subjects are assigned randomly to each treatment. If random allocation is not explicitly part of the experimental procedure then there needs to be the belief that the actual allocation was as likely to occur as any other.

2.3 Bootstrapping

The bootstrap is a method for assessing properties of a statistical estimator in a *non-parametric* framework. That is, we do not assume that the data are obtained from any parametric distribution (eg. normal, exponential etc).

The bootstrap is usually used to assess the variance of a statistical estimator but it is not exclusively used for this purpose.

The name comes from the story 'The Surprising Adventures of Baron Munchausen', where the main character pulls himself out of a swamp, by pulling on his own bootstraps.

The idea behind bootstrapping is that we can use the data muliple times to generate 'new' data sets to assess the properties of parameters.

Recap: CDFs

2.3.1 The Empirical Distribution Function

Define

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \leq x}$$

to be the *empirical distribution function (edf)* for data $\{X_1, \ldots, X_n\}$.

- $\hat{F}$ takes values in $\{0, \frac{1}{n}, \ldots, \frac{n}{n}\}$
- to sample from $\hat{F}$ we sample **WITH REPLACEMENT** from $\{X_1, \ldots, X_n\}$.

Note that $\hat{F}$ is a random quantity. We consider the edf to be an estimator for F. If X_i are all from distribution F then the following results hold.

Properties of the EDF - I

1. $\hat{F}(x)$ is an unbiased estimator of $F(x)$.

$$\mathbb{E}\hat{F}(x) = F(x)$$

Proof

Properties of the EDF - II

2. $\hat{F}(x) \to F(x)$ as $n \to \infty$ with probability 1.

   Proof

Properties of the EDF - III

3.

$$\frac{\sqrt{n}(\hat{F}(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \to N(0, 1) \text{ in distribution}$$

as $n \to \infty$.

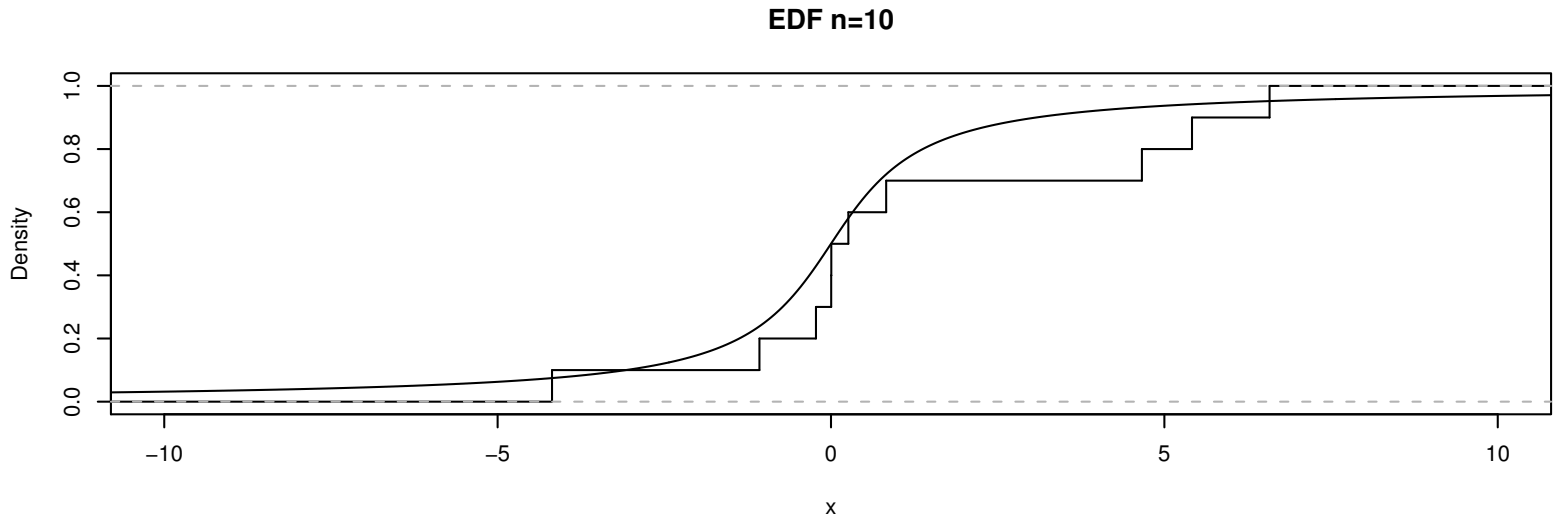
Proof:

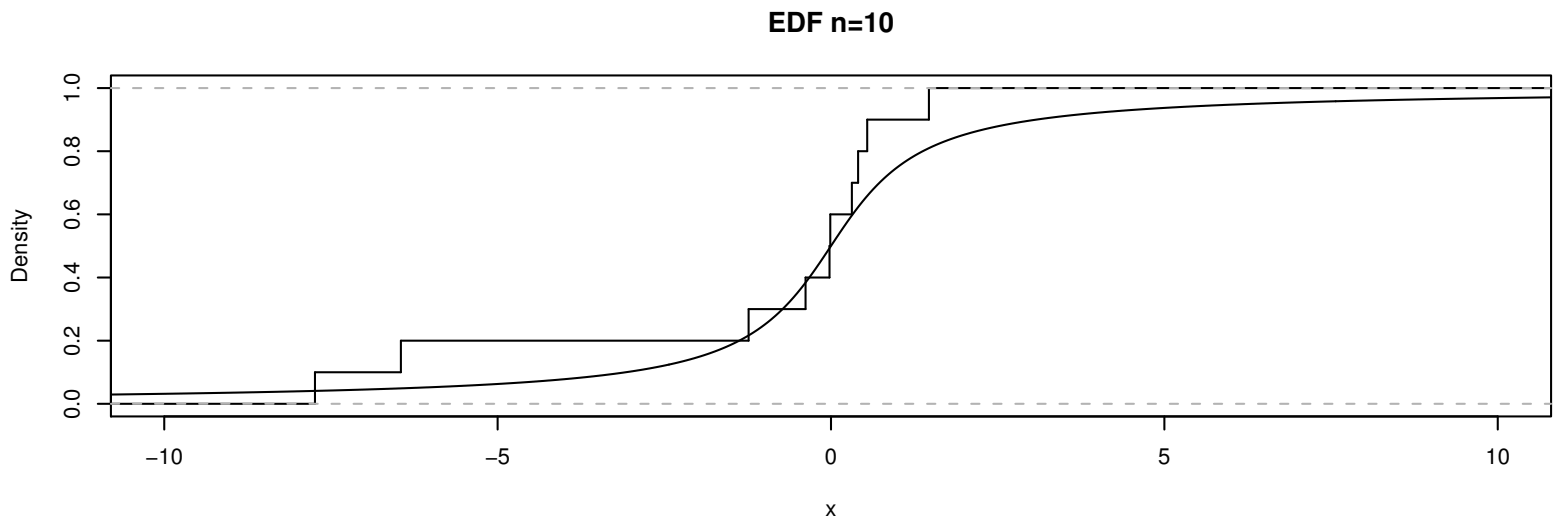
Properties of the EDF - IV

4 If X_i are an independent identically distributed sequence (so it doesn't matter if we change the order), then knowledge of $\hat{F}$ is equivalent to knowledge of $\{X_1, \ldots, X_n\}$.

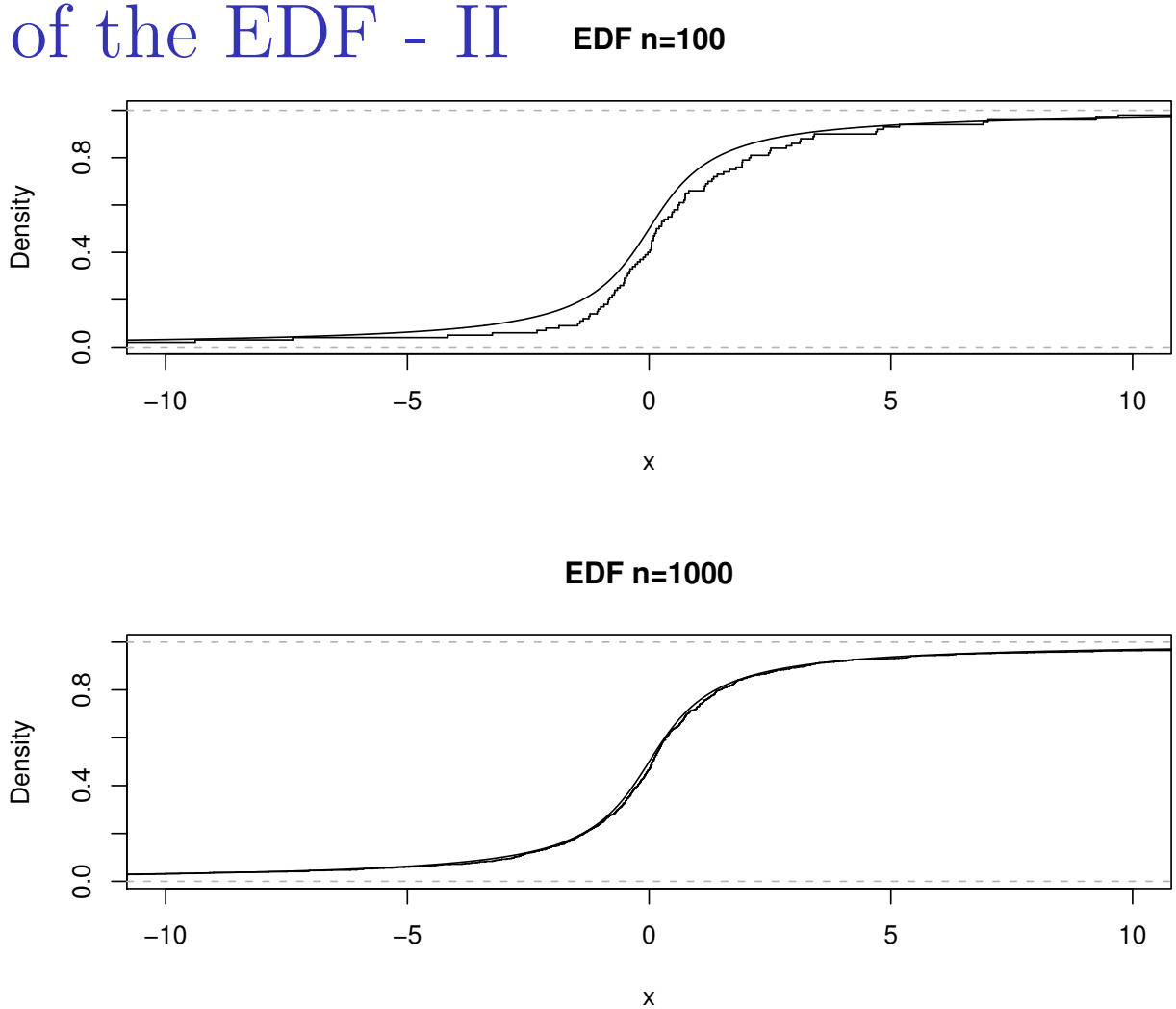
Example of the EDF

Suppose $X_i \sim$ Cauchy. Then we can examine the edf for increasing values of n.





Example of the EDF - II



Notice that we've repeated for $n = 10$ and that the edf is different each time. Notice also that the edf becomes more accurate as n gets larger.

Parameters, statistics and properties

Bootstrapping texts can sometimes be confusing because of the language usage. We say

- θ is a parameter if it is a property of the underlying population, i.e., $\theta = \theta(F)$.
- $\hat{\theta}$ is a statistic which estimates θ if $\hat{\theta}$ is a function of the sample $X_1, \ldots, X_n$ - this is equivalent to being a function of the empirical distribution function

$$\hat{\theta} = \theta(\hat{F}) \equiv \theta(X)$$

- We then talk about properties of $\hat{\theta}$ such as its bias, its expectation, its standard error etc. For bootstrapping applications these properties are usually sampling properties, that is, if we repeatedly collected similar samples, what properties would $\hat{\theta}$ have?

Difficulties arise when we note that properties are also parameters of the statistic $\hat{\theta}$ and that we estimate them with statistics of the statistics.

Plug-in Principle

For example, suppose we have a sample of size n, $\{X_1, \ldots, X_n\}$ say, from unknown density F.

Suppose that interest lies in some parameter θ of the distribution F which we write $\theta = \theta(F)$ where we consider θ to be a functional of the distribution F.

We estimate θ by $\hat{\theta}$ where $\hat{\theta}$ is a function of the observations $\{X_1, \ldots, X_n\}$. Usually we have that $\hat{\theta} = \theta(\hat{F})$, that is, if we apply the functional $\theta(\cdot)$ to the edf $\hat{F}$ we get the statistic $\hat{\theta}$.

The parameter θ and the statistic $\hat{\theta}$ are both found by using the functional $\theta(\cdot)$. For the parameter we have $\theta = \theta(F)$, and for the statistic we have $\hat{\theta} = \theta(\hat{F})$.

This is what we call the *plug-in principle*. To estimate parameter $\theta = \theta(F)$ when we don't know F, we plug-in the empirical distribution function $\hat{F}$ to find the estimator $\hat{\theta} = \theta(\hat{F})$.

Examples of parameters and the plug-in principle

1 **Population mean**

$$\theta = \theta(F) = \mathbb{E}_F X = \int x \mathrm{dF(x)} = \int \mathrm{xf(x)dx}$$

Use the plug-in principle

Here $\delta(x)$ is the Dirac delta function which is defined by its behaviour under integration

$$\int_A \delta(x-a)\mathrm{dx} = \begin{cases} 1 \text{ if } a \in A \\ 0 \text{ if } a \notin A \end{cases}$$

The delta function $\delta(x-a)$ is the derivative of the indicator function $\mathbb{I}_{x \leq a}$.

Examples of parameters and the plug-in principle

2 Population variance

$$\theta = \theta(F) = \mathbb{V}\mathrm{ar}_F(X) = \int (x - \mathbb{E}_F(X))^2 \mathrm{dF(x)}$$

Examples of parameters and the plug-in principle

3 Probability

$$\theta = \mathbb{P}_F(X > c) = \int_c^\infty \mathrm{dF(x)}$$

2.3.2 Estimating sampling properties with the bootstrap

For our estimates to be of any value, it is necessary to know their properties, such as the bias or the standard error:

- The bias is defined as

$$\text{bias}(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta$$

- The standard error is

$$\text{se}(\theta) = \sqrt{[\mathbb{E}(\hat{\theta} - \theta)^2]}$$

If we believed that X_i were from a specific parametric model

$$\text{e.g. } F = \Phi \text{ so that } X_i \sim N(\mu, \sigma^2)$$

then we could calculate the bias and standard error analytically. If these calculations were difficult or impossible (for example, if $\theta =$ trimmed mean) then we can use simulations from F to estimate the standard error and bias of the statistics.

What if we don't have a parametric model for F?

The bootstrap can be used to estimate the sampling distribution in this case.

The idea is that instead of sampling from the population of interest, i.e. from $F(\cdot)$, we instead sample with replacement from the sample $\{x_1, \ldots, x_n\}$, i.e. from $\hat{F}(\cdot)$.

Example 1: Heart-attack study

A controlled, randomized, double-blind study was carried out to investigate whether or not aspirin reduces the risk of heart attacks in healthy middle-aged men. Data from the study is

	heart attacks (fatal plus non-fatal)	subjects
aspirin	104	11037
placebo	189	11034

Heart-attack study -II

Define θ to be the true ratio of proportions of heart attacks in those with aspirin to those with a placebo, the relative risk.

From the data, the estimate of θ suggests that aspirin lowers the risk of a heart attack:

$$\hat{\theta} = \frac{104/11037}{189/11034} = 0.55$$

But how confident can we be?

Can we calculate a confidence interval for θ?

It is possible to derive a parametric confidence interval for θ from theory by assuming that the log relative risk is normally distributed. But what if we've forgotten how, or don't wish to assume normality?

Heart-attack study -III

Bootstrapping enables us to derive these intervals without assuming that the log relative risks are normally distributed:

1. Estimate the probability $\hat{p}_1$ of a patient with aspirin having a heart-attack:
$$\hat{p}_1 = \frac{104}{11037} = 0.00942$$
2. Estimate the probability $\hat{p}_2$ of a patient with a placebo having a heart-attack:
$$\hat{p}_2 = \frac{189}{11034} = 0.0171$$
3. Simulate data for a new experiment: sample r_1 from Binomial$(11037, 0.00942)$ and r_2 from Binomial$(11034, 0.0171)$. The new data is known as a bootstrap sample.
4. Obtain a new estimate of the ratio:
$$\hat{\theta}^*_s = \frac{r_1/11037}{r_2/11034}$$

Heart-attack study -IV

Steps 3 and 4 are then repeated a large number of times, to obtain a sample

$$\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B\}$$

We can then use the 2.5th and 97.5th percentiles of this sample as a 95% confidence interval for θ.

With $B = 10000$, performing this procedure in R gave a 95% interval of $(0.43, 0.69)$ for θ.

We will now formally introduce the bootstrap and look at some examples in detail.

The bootstrap

The basic idea behind the bootstrap is to find properties of statistics $\hat{\theta}$ by resampling from $\hat{F}$ (rather than F).

- If we could generate from F, we could simulate sample data sets $\{X_1^{(i)}, \ldots, X_n^{(i)}\}$ for $i = 1, \ldots, B$ and find $\hat{\theta}^{(i)}$. We can then learn properties of $\hat{\theta}$ from $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$.

But usually we don't know F and so can't produce these samples. Instead we can bootstrap. This involves two ideas:

(i) Replace F by $\hat{F}$.

(ii) We sample from $\hat{F}$ and find the properties of $\hat{\theta}$ under $\hat{F}$.

The bootstrap

The bootstrap algorithm

1. Generate B bootstrap replicates from $\hat{F}$.

$$\mathbf{X}^{*(i)} = \{X_1^{*(i)}, \ldots, X_n^{*(i)}\} \text{ for } i = 1, \ldots, B$$

2. Calculate B bootstrap parameter estimates

$$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$$

3. Calculate the property of interest for $\hat{\theta}$ from $\{\hat{\theta}_i^*\}_{i=1}^B$ e.g.

$$\text{se}_{boot}(\hat{\theta}) = \sqrt{\mathbb{E}_F(\hat{\theta} - \theta)^2} \approx \sqrt{\frac{1}{B-1}\sum(\hat{\theta}_i^* - \bar{\hat{\theta}})^2}$$

where $\bar{\hat{\theta}} = \frac{1}{B}\sum \hat{\theta}_i^*$

The bootstrap

We call iid samples of size n from $\hat{F}$ *bootstrap replicates*.

They can be generated by sampling with replacement from $\{x_1, \ldots, x_n\}$.

In R this can be achieved by using the command

```
sample(n=1, size=n, data=x, replace=T)
```

The bootstrap estimate of standard error

Suppose $\hat{\theta}(X)$ is some statistic based on $X = \{x_1, \ldots, x_n\}$ used for estimating parameter θ. The standard error of $\hat{\theta}(X)$ is

$$\text{se}(\hat{\theta}) = \sqrt{\mathbb{V}\text{ar}_F(\hat{\theta}(X))}$$

Here the variance is with respect to distribution F. The bootstrap estimate is found by

ii replacing F with $\hat{F}$.

$$\text{se}_F(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{n}})}{\approx} se_{\hat{F}}(\hat{\theta})$$

iiii Approximating $se_{\hat{F}}$ using simulation:

$$se_{\hat{F}}(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{B}})}{\approx} \left(\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}(X^{(b)}) - \bar{\hat{\theta}})^2 \right)^{\frac{1}{2}} =: \text{se}_{boot}$$

where $X^{*(b)} = \{X_1^{*(b)}, \ldots, X_n^{*(b)}\}$ is a bootstrap sample from $\{x_1, \ldots, x_n\}$ ie, se^2_{boot} is the variance of $\hat{\theta}(X^*)$ when X^* are drawn from $\hat{F}$.

The bootstrap estimate of standard error - II

To make this more explicit, note that the first step is using the plug-in principle again.

If we consider the variance of $\hat{\theta}$ to be a functional of F -

$$\mathbb{V}\text{ar}(\hat{\theta})[F] = \mathbb{E}_F(\hat{\theta} - \mathbb{E}_F(\hat{\theta}))^2$$

then when we plug-in $\hat{F}$ we find

$$\mathbb{V}\text{ar}_{\hat{F}}(\hat{\theta}) = \mathbb{E}_{\hat{F}}(\hat{\theta} - \mathbb{E}_{\hat{F}}\hat{\theta})^2.$$

The bootstrap estimate of standard error - III

The second step is then to estimate se_{boot} by simulation by replacing $\mathbb{V}\text{ar}_{\hat{F}}(\hat{\theta}(X^*))$ with an estimate

$$\mathbb{V}\text{ar}_{\hat{F}}(\hat{\theta}(X^*)) \approx \mathbb{V}\text{ar}_{boot}(\hat{\theta}(X^*)) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}(X^{*(b)}) - \bar{\hat{\theta}})^2$$

where

$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(X^{*(b)})$$

and where

$$X^{*(b)} = \{X_1^{*(b)}, \ldots, X_n^{*(b)}\}$$

are B bootstrap replicates from $\hat{F}$.

Bootstrap estimate of bias

The bias of an estimator $\hat{\theta}$ of parameter θ is defined as

$$\text{bias} = \mathbb{E}_F(\hat{\theta}) - \theta$$

i.e., how the mean of the estimator of θ differs from the true value of θ.

An estimate is found by replacing F by $\hat{F}$.

$$\text{bias}_{\hat{F}} = \mathbb{E}_{\hat{F}}(\hat{\theta}) - \hat{\theta}$$

That is, the difference between the expected value and the estimated value.

This, again, is the plug-in principle.

Bootstrap estimate of bias - II

We can estimate $\mathbb{E}_{\hat{F}}(\hat{\theta})$ from bootstrap samples as

$$\mathbb{E}_{\hat{F}}(\theta) \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{(b)}$$

giving us the bootstrap estimator of the bias to be

$$\text{bias}_{\hat{F}}(\hat{\theta}) \approx \text{bias}_{boot}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{(b)} - \hat{\theta}$$

So in general there are two approximation steps in the bootstrap procedure:

1. Replace F by $\hat{F}$.
2. Simulate from $\hat{F}$ to form the estimate of the property of interest.

The error in the first approximation scales with the number of data points (and so is fixed for any given problem). The error in the second approximation scales with B, the number of bootstrap replicates, and so can be controlled.

$$\mathrm{se}_F(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{n}})}{\approx} se_{\hat{F}}(\hat{\theta}) \overset{O_p(\frac{1}{\sqrt{B}})}{\approx} \mathrm{se}_{boot}(\hat{\theta})$$

Here $Y_n = O_p(x_n)$ means that Y_x/x_n is stochastically bounded, i.e., for any $\epsilon > 0$, there exists $M > 0$, such that for all n

$$\mathbb{P}(|Y_n/x_n| > M) < \epsilon$$

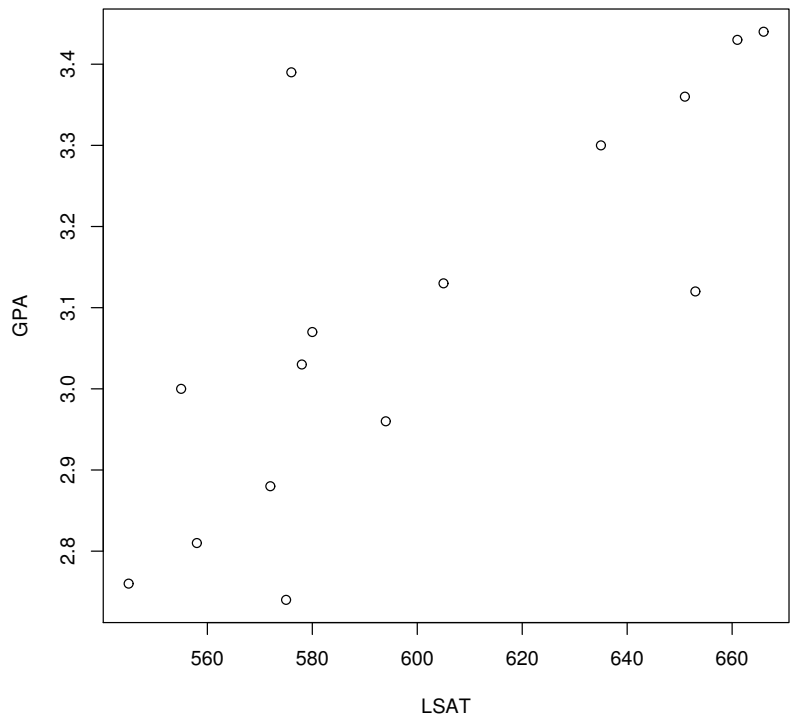
Lawschool example

A sample of 15 law schools was taken, and two measurements were made for each school:

x_i : LSAT, average score for the class on a national law test

y_i : GPA, average undergraduate grade-point average for the class

We are interested in the correlation coefficient between these two quantities, which we estimate to be $\hat{\theta} = 0.776$.



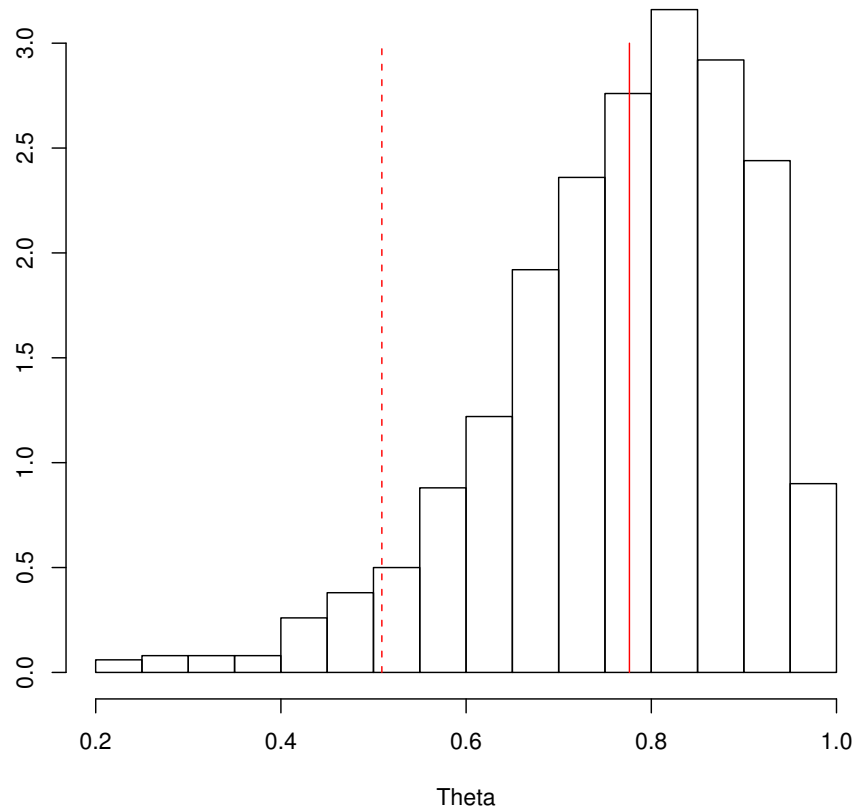
Lawschool example - II

But how accurate is this estimate of the correlation coefficient? We use the bootstrap to estimate the standard error of $\hat{\theta} = \text{cor}(LSAT, GPA)$.

1. Sample 15 data points with replacement from the observed data $z = \{(x_1, y_1), \ldots, (x_{15}, y_{15})\}$ to obtain new data z^*.
2. Evaluate the sample correlation coefficient $\hat{\theta}^*$ for the newly sampled data z^*.
3. Repeate steps 1 and 2 to obtain $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$.
4. Estimate the standard error of the sample correlation coefficient by the sample standard deviation of $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$.

Lawschool example - III

With $B = 1000$, I found the estimated standard error of $\hat{\theta}$ is 0.137. It can help to plot a histogram of the bootstrap replicates as this gives more information about the distribution of $\hat{\theta}$.



Bootstrap confidence interval

We have two methods of calculating CIs.

1. **Normal interval** Given an estimate of the standard error of $\hat{\theta}$, if we assume that the distribution of $\hat{\theta}$ is approximately normal, then an approximate 95% confidence interval is given by

$$\hat{\theta} \pm 1.96 \text{se}(\hat{\theta}^*)$$

   - For the law dataset we find a 95% CI for $\text{cor}(LSAT, GPA)$ of $[0.51, 1.04] \equiv [0.51, 1.00]$.

   This interval is not accurate unless the distribution of bootstrap samples is approximately normal.

Bootstrap confidence interval - II

2 **Percentile confidence interval**

For a 95% confidence interval, we need to find the two values l and u with

$$\mathbb{P}(\hat{\theta}^* > u) = 0.975$$
$$\mathbb{P}(\hat{\theta}^* < l) = 0.025$$

ie, we need to identify 2.5th and 97.5th percentiles from the distribution of $\hat{\theta}^*$. We can find this from the 2.5th and 97.5th percentiles of the sample $\{\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}\}$. We generally need a larger value of B to get accurate percentile estimates than that required to find an accurate estimate of the standard error.

- For the law dataset we find a 95% CI of $[0.45, 0.96]$.

Hypothesis testing with the bootstrap

Example: mice survival times

Treatment	94	197	16	38	99	141	23		
Control	52	104	146	10	50	31	40	27	46

Is there a difference between two group means?

- Denote 7 treatment observations by $\mathbf{x} = \{x_1, \ldots, x_7\}$, and 9 control observations by $\mathbf{y} = \{y_1, \ldots, y_9\}$.
- Could perform two-sample t test, assuming normally distributed responses and equal variances in the two groups.
- Define μ_X: population treatment mean, μ_Y: population control mean. For one-sided test of $H_0 : \mu_X = \mu_Y$, observed $p-$value is 0.1405.

Bootstrap hypothesis test

- Alternative to assuming normality
- Denote F_X: distribution of treatment survival time, F_Y: distribution of control survival time.
- Write null hypothesis as $H_0 : F_X = F_Y = F$, with F the single common distribution of all the responses.
- estimate F by $\hat{F}$, empirical cdf of all 16 observations.

Bootstrap two-sample significance test

1. Sample 16 values with replacement from $\{x_1, \ldots, x_7, y_1, \ldots, y_9\}$.
2. Set $\{x_1^*, \ldots, x_7^*\}$ to be the first 7 sampled values, and $\{y_1^*, \ldots, y_9^*\}$ to be the remaining 9 sampled values.
3. Calculate the bootstrap test statistic

$$T^* = \frac{\bar{x}^* - \bar{y}^*}{\hat{\sigma}^* \sqrt{1/7 + 1/9}}$$

   for the sampled data.
4. Repeat steps 1 to 3 B times to obtain $T_{(1)}^*, \ldots, T_{(B)}^*$.
5. Estimate the significance of the observed T_{obs} by

$$\frac{1}{N} \sum_{i=1}^{N} I\{T_{(i)}^* \geq T_{obs}\} \tag{5}$$

The Parametric bootstrap

Thus far we have been using the **non-parametric bootstrap**, ie,

- sample from $\hat{F}$ making no assumptions about the distribution of the data.

The **parametric bootstrap** can be used when we believe $F = F_\theta$, i.e. we have a parametric model for the data.

Then instead of sampling from $\hat{F}$, we sample from $F_{\hat{\theta}}$.

In the mice example, we would replace step 1. on slide 77 by

1a. Estimate population mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, e.g.,

$$\hat{\mu} = \frac{1}{16}(\sum x_i + \sum y_i)$$

1b. Sample 16 values from a $N(\hat{\mu}, \hat{\sigma}^2)$ distribution.

Steps 2-5 remain unchanged.

The Bootstrap and Regression

A formal regression type model has the structure

$$y_i = f(x_i, \beta) + \epsilon_i$$

where f is a specified function acting on the covariates x_i with parameters β, and ϵ_i is a realisation from a specified error structure. With this model framework, there are two alternative ways to bootstrap the model.

1. Fit the regression model, form the empirical distribution of the residuals, generate bootstrap replications of the data by substituting these back into the model, and re-fit the model to obtain bootstrap distributions of β. This is called *model-based resampling*.
2. Bootstrap from the pairs (x_i, y_i), re-fit the model to each realization, form the bootstrap distribution of β.

Model-based resampling

We will fit a model of the form

$$GPA_i = \beta_0 + \beta_1 LSAT_i + \epsilon_i$$

to the law data. A least squares fit to these data gives $\hat{\beta}_0 = 0.3794$ and $\hat{\beta}_1 = 0.0045$, but how accurate are these values? We can perform the following set of steps to find the standard error of these estimates

Model-based resampling - II

1. Find the fitted residuals

$$\hat{\epsilon}_i = GPA_i - \hat{\beta}_0 - \hat{\beta}_1 LSAT_i$$

2. Sample $\epsilon_1^*, \ldots, \epsilon_{15}^*$ with replacement from $\{\hat{\epsilon}_1, \ldots, \hat{\epsilon}_{15}\}$
3.

$$\text{Set } GPA_i^* = \hat{\beta}_0 + \hat{\beta}_1 LSAT_i + \epsilon_i^*$$

4. Fit the least squares regression to $\{(LSAT_1, GPA_1^*), \ldots, (LSAT_{15}, GPA_{15}^*)\}$ to find estimates $\beta = (\beta_0^*, \beta_1^*)$.
5. Repeat steps 2 to 4 B times to find bootstrap replicates

$$\{(\beta_0^{*(1)}, \beta_1^{*(1)}), \ldots, (\beta_0^{*(B)}, \beta_1^{*(B)})\}$$

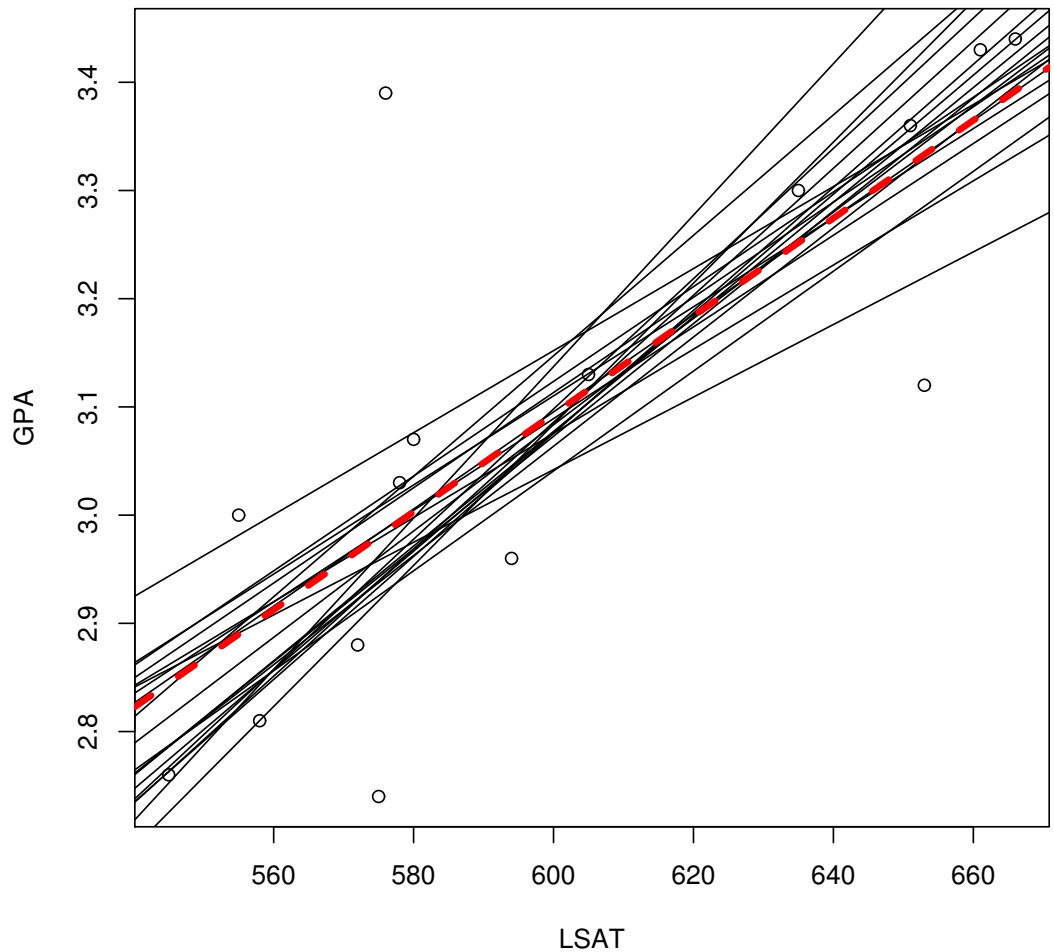
and use these replicates to estimate $\text{se}(\hat{\beta}_0)$ and $\text{se}(\hat{\beta}_1)$.

Model-based resampling - III

Using 1000 bootstrap replicates, I find the standard errors are

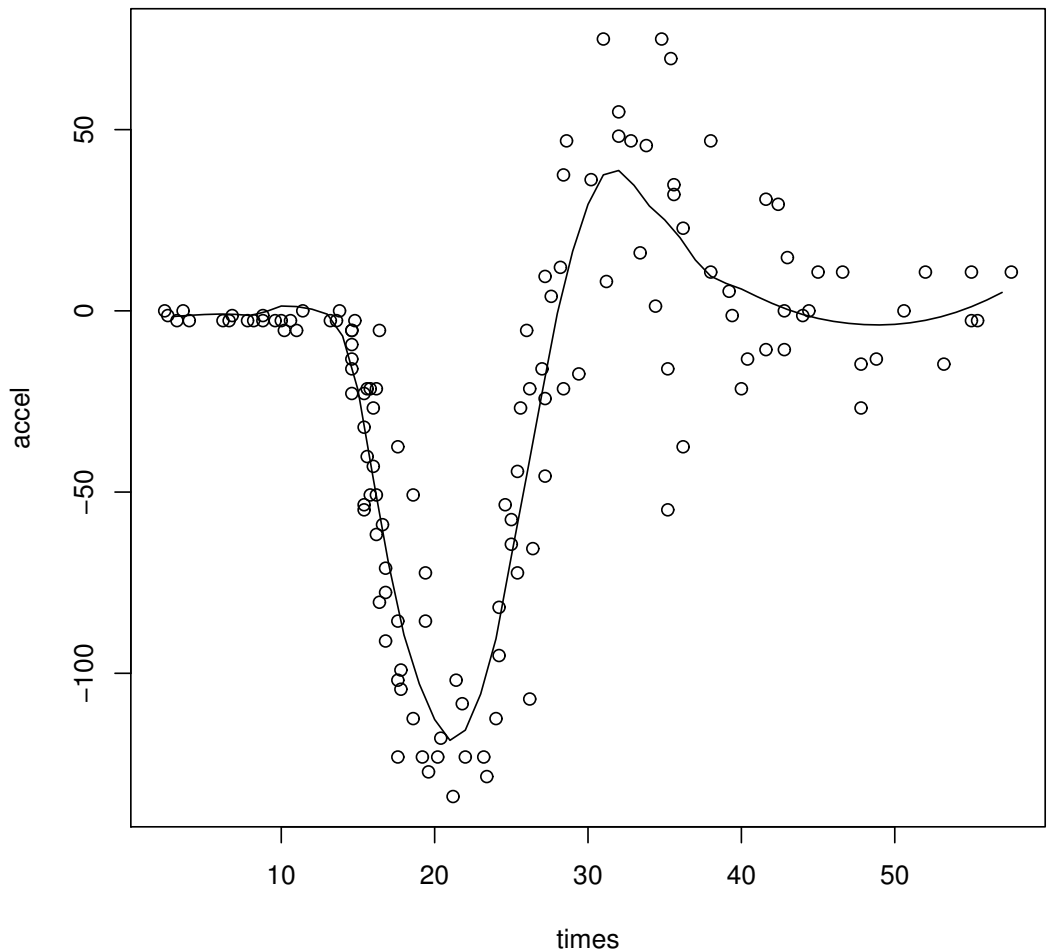
$$\text{se}(\hat{\beta}_0) = 0.586, \qquad \text{se}(\hat{\beta}_1) = 0.000973$$

and the plot shows a sample of 20 bootstrap regression lines.



Motorcycle example

We consider data providing measurements of acceleration against time for a simulated motorcycle accident. The data are shown in the figure.

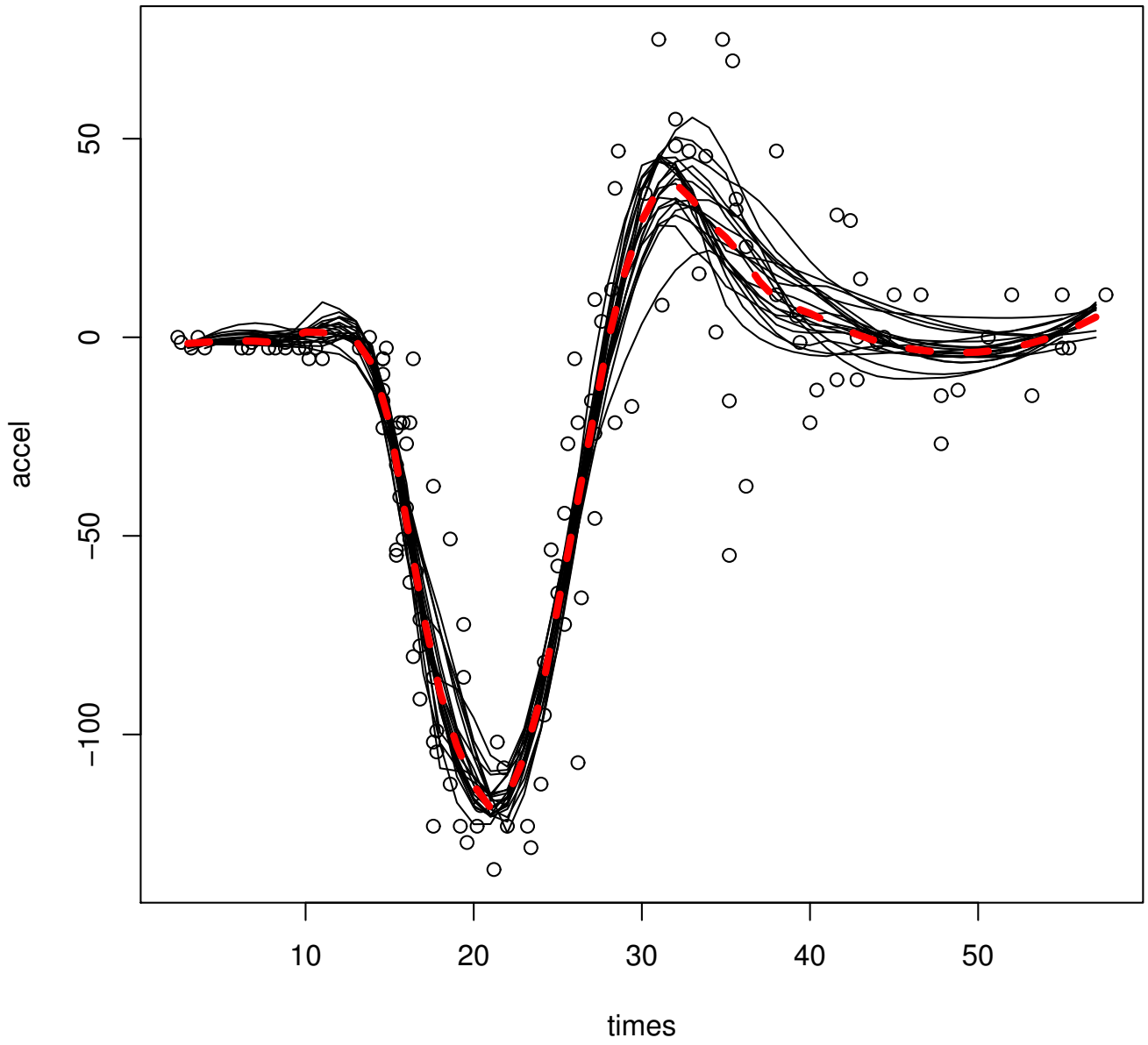


Motorcycle example

Clearly the relationship is nonlinear, and has structure that will not easily be modelled parametrically. We use the `loess()` command in R to fit a locally weighted least-squares regression line to the data (the details aren't important for this course, but for completeness sake we set `span`=1/3 which determines the proportion of the data to be included in the moving window which specifies which points are to be regressed upon.). The figure shows the best fit.

Because of the non-parametric structure of the model, classical approaches to the assessment of parameter precision are not available. We can get a sense of how accurate the parameters are by using a bootstrapping scheme (of the first type). This is achieved by simply bootstrapping the pairs (x,y) in the original plot and fitting `loess` curves to each simulated bootstrap series. A figure showing 20 bootstrap samples is shown below.

Motorcycle example



R code is available in `motorcycle.txt`.

Summary

1. Monte Carlo tests
   - Will work with any test statistic and hypothesis, but requires specification of distribution of data under null hypothesis
   - Only procedure out of three that produces 'completely new' data.
2. Randomization tests
   - Can generally only handle tests of no treatment effect between different treatment groups. One sample tests can be performed, but under stricter assumptions.
   - No distribution is required/assumed for the data, only that allocation of subjects to treatment groups is random.

3 Bootstrapping

- Arguably the most widely applicable method of the three.
- Main use is to construct confidence intervals
- Dependent on the empirical cdf being a good approximation to the true distribution.
- Accuracy ultimately depends on size of **original** sample.

2.4 Prediction errors and cross-validation

We fit models by minimizing some measure of error, eg, we fit a linear model by minimizing sum of squares

$$S(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

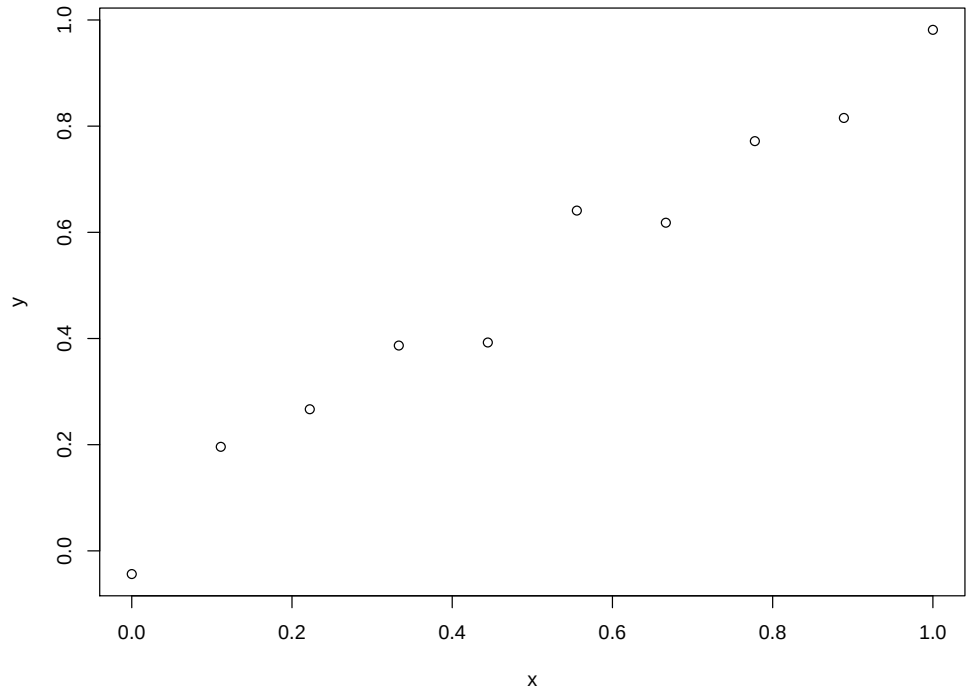
When choosing between competing models we might be tempted to take the model that achieves the lowest error rate on the training data.

However, the error we achieve on the training data is not the same as the error we expect when predicting new data.

We need to be careful when choosing between models not to over-fit and choose a model that is too complex.

Example: Over-fitting

Suppose we are given data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$



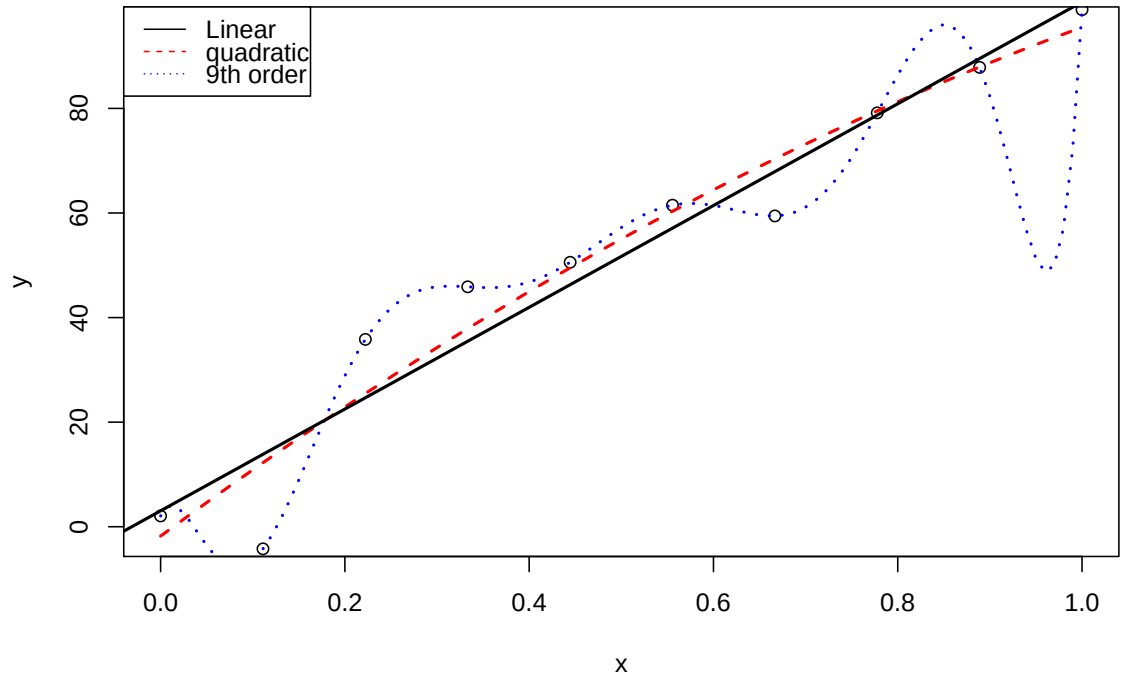
and we want to choose between

$$\mathcal{M}_1 : y = \beta_0 + \beta_1 x + \epsilon \qquad \mathcal{M}_2 : y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

$$\vdots$$

$$\mathcal{M}_d : y = \beta_0 + \beta_1 x + \ldots + \beta_d x^d + \epsilon$$

Example: Over-fitting



The plot shows the fitted curves for $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_9$. The residual sum of squares is 676 ($\mathcal{M}_1$), 590 ($\mathcal{M}_2$), and 0 ($\mathcal{M}_9$).

> *With four parameters I can fit an elephant, and with five I can make him wiggle his trunk. John von Neumann*

Example: Over-fitting

$\mathcal{M}_9$ is a perfect fit to the training data - the residual sum of squares is 0.

- With n data points, we can always find a polynomial of degree $n-1$ that fits perfectly.

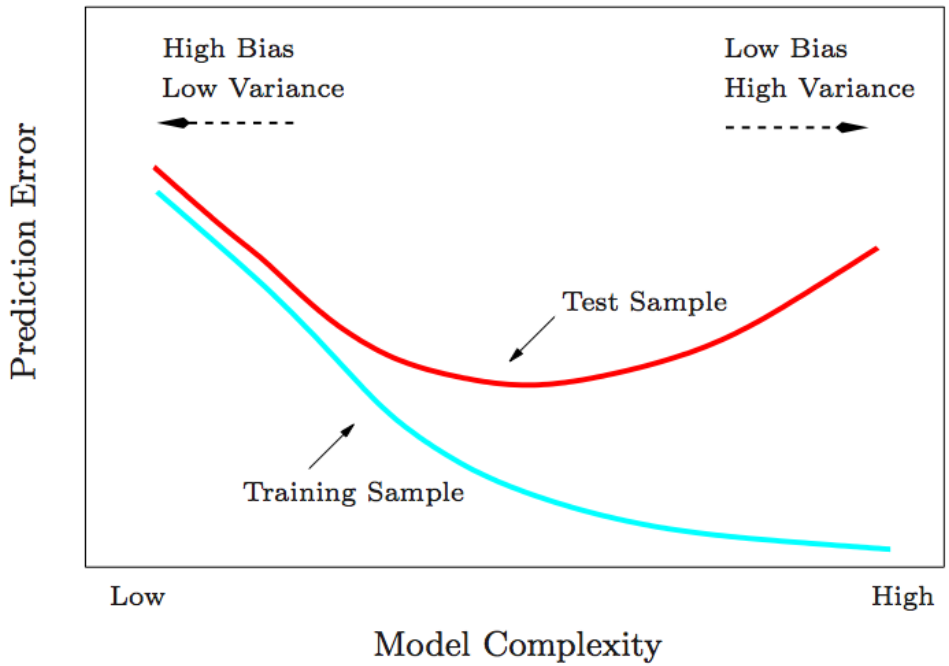
But $\mathcal{M}_9$ is over-fit - it is modelling the noise not the signal and would fail to accurately predict new data.

We know in general that fitting high order polynomials to regression data is not a sensible thing to do, but how can we demonstrate this?

- Some methods adjust the training error to account for the model complexity, e.g., AIC, BIC, C_p statistic.

Alternatively, in data-rich environments, we can simply split the data into a training set, and a test set. We fit the model on the training-set, and then test its predictive accuracy on the test set.

Training vs test set performance



Making a model more complex will **always** result in a better fit to the training data. But there is a **bias-variance** trade-off

- bias occurs from errors in the model structure, ie, from models that are too simplistic
- variance occurs from needing to estimate parameters - for complex models with many parameters, fitting can be sensitive to small fluctuations in the training set leading us to fit the noise rather than the signal.

Cross-validation

Cross-validation is means of efficiently assessing **predictive accuracy**, and extends the idea of having a test and training datasets.

Leave-one-out cross-validation (LOO-CV) For $i = 1, \ldots, n$

1. Fit the model to the reduced data set (or training set),

$$\{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$$

2. Obtain from the fitted model the predicted value $\hat{y}_i$ at x_i.
3. Compute the squared error $\epsilon_i^2 = (\hat{y}_i - y_i)^2$

An average squared prediction error can then be reported as $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$, or the root-mean-square (rms) prediction error as $\sqrt{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2}$.

All predictions are on held-out data (test data) and so this gives us a measure of a model's predictive skill.

k-fold cross-validation

Note that this is not the expected prediction error of the actual model (as we have only fit to $n-1$ data points), though it should be close if n is sufficiently large (so that the fit to $n-1$ points is very similar to the fit to n points).

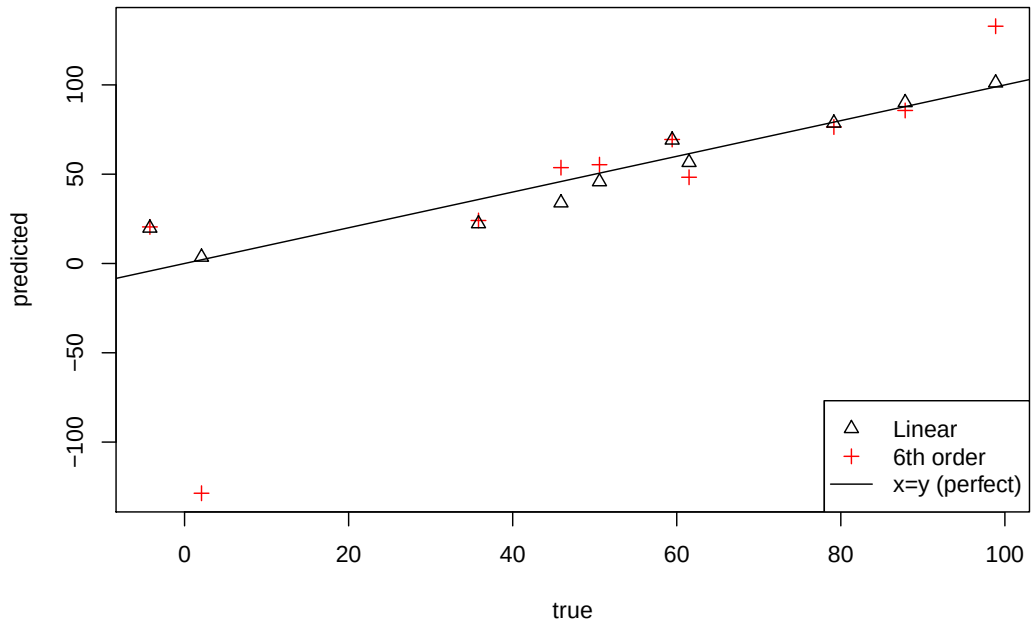
This approach left out one point at a time, and is called **leave-one-out cross-validation**.

K-fold cross-validation splits the data into K chunks of approximately equal size. Then for $k = 1, \ldots, K$,

- Delete chunk k from the data
- Fit the model to the rest of the data
- Use the fitted model to predict the data in chunk k and compute the prediction error.

Setting $K = n$ gives leave-one-out cross-validation!

If we do LOO-CV for our example, then we can plot the predicted value against the true observed value for different models. A perfect model would have predictions on the line $y = x$.



We can see that linear regression is much better in terms of predictive performance than the 6th degree polynomial.

The mean square prediction error for the straight line is 1065.8 whereas for the 8th order polynomial it is 1936.

What K should we use in K-fold cross-validation?

There is a variance-bias trade-off here too!

- The variance of our estimate of the predictive error grows as K gets larger
  - For large K, e.g. LOO-CV with $K = n$, the data doesn't typically get *shaken up* enough. In LOO-CV each fold only differs by two data points, and so estimates from each fold are highly correlated. Hence our estimate of the average prediction error has a high variance (ie is unreliable).
  - For small K, the folds are very different, and so the error estimates are less correlated, and we get a stable estimate.
- The bias of our estimate of the predictive error shrinks as K gets larger
  - Since each training set contains only $\frac{K-1}{K}n$ data points, rather than n, the estimate of the prediction error is usually biased upwards (ie is too large)
  - The bias is minimized for $K = n$, but this has high variance.

$K = 5$ or $K = 10$ are both usually considered good choices but it can vary between applications.

The `cvTools` package in R can be used to do cross-validation.