

PROBABILITY AND STATISTICS

II

wi2605

Part 1

E.A.Cator

DELFT UNIVERSITY OF TECHNOLOGY
SPRING 2004

©e.a.cator

Contents

1	Testing hypotheses	1
1.1	Testing hypotheses	1
1.2	Definitions	5
1.3	Power of a test	8
1.4	Neyman-Pearson test	10
1.5	Likelihood ratio test	12
1.6	P-value	15
1.7	Confidence regions for θ	18
1.8	Confidence region for $\rho(\theta)$ and one-sided confidence intervals . .	23
1.9	Solutions to the quick exercises	27
1.10	Exercises	29
2	Linear Models	33
2.1	Introduction	33
2.2	Linear models	35
2.3	Least squares estimator	36
2.4	Random vectors	39
2.5	Properties of the least squares estimator	43
2.6	Normality assumption	46
2.7	Testing Hypotheses	47
2.8	Testing without normality assumptions	55
2.9	Solutions to the quick exercises	58
2.10	Exercises	60
3	Nonparametric statistics	61
3.1	Introduction	61
3.2	Kernel density estimators	61
	Choice of bandwidth	64
	Mean Integrated Square Error	67
	Least squares cross-validation	68
	Testing for normality using kernel estimators	71
	Some simulation results	75
3.3	Monotone densities	79
	The maximum likelihood estimator of a monotone density	80
	Testing for exponentiality	87
	Some simulation results	88

3.4	Solutions to the quick exercises	89
3.5	Exercises	91
Tables		99

Chapter 1

Testing hypotheses

In this chapter we will repeat and generalize the notion of testing hypotheses. We will introduce the power of a test, with which we will be able to compare two different tests for the same hypothesis. The classical Neyman-Pearson test will be mentioned, an example of a uniformly most powerful test, and the likelihoodratio test will be defined. Some optimality properties of this test will also be mentioned. Finally, we will use hypothesis testing to define confidence intervals for certain parameters and show how one could use a bootstrap-like method to determine these intervals.

1.1 Testing hypotheses

Let us review a simple example to refresh our memory: suppose we have a batch of microchips. We know (or assume) that each microchip has the same probability p to be defect, independently from the other chips, we just don't know what is the value of p . We decide to check 20 chips. The outcome of this experiment is described by the number of defect chips, so the *data space* $\mathcal{X}$ of this experiment, or the possible outcomes of this experiment, is

$$\mathcal{X} = \{0, 1, \dots, 20\}.$$

Let us call the outcome of the experiment x , also called the *data*. We model the data as a realization of a random variable X . Now the unknown parameter p describes the probability distribution of the random variable X . After all, it is clear that $X \sim \text{Bin}(20, p)$, since we check 20 chips and each of them have probability p to be defect. Therefore, the collection $\{p : 0 \leq p \leq 1\}$ describes a *statistical model* of possible distributions for our random variable X describing the data. We will denote our statistical model by Θ .

Based on the data x we would like to say something about p , for example can we conclude that $p \leq 0.06$? It might be that the manufacturer of the chips claims that no more than 6% of their chips is defect, and we want to use our experiment to see if we can find a significant indication that this claim is false. This is how we should think of the *null hypothesis*. It is a statement about the true parameter p , and we use our experiment to see if we can find evidence against this statement. However, the statement of the null hypothesis gets the

benefit of the doubt: we really need quite strong evidence against it in order to reject it. Formally we will view the statement of the null hypothesis as being of the form: the true parameter lies in the subset H_0 of the statistical model Θ . So in this example we would have

$$H_0 = \{p : 0 \leq p \leq 0.06\}.$$

Our null hypothesis H_0 can therefore be viewed as a subset of Θ :

$$H_0 \subset \Theta.$$

In some cases you might not just want to find evidence against H_0 , but you want evidence that points in the direction of an alternative statement about the true parameter, called the *alternative hypothesis* H_1 . For instance, it might be that a consumer organization tested the microchips in some earlier experiment and they claimed that actually $p = 0.09$. Like with H_0 , we will view H_1 as a subset of Θ :

$$H_1 \subset \Theta.$$

It must always hold that $H_0 \cap H_1 = \emptyset$ (that is why it is called the *alternative hypothesis*), and often we will have that $H_0 \cup H_1 = \Theta$, so this corresponds to finding any evidence against H_0 . Let's also assume this in our example, so

$$H_1 = \{p : 0.06 < p \leq 1\}.$$

Our goal in hypothesis testing is determining whether or not we can reject the null hypothesis H_0 in favor of H_1 , based on the outcome of our experiment, that is, the data x . So we want to find a function ϕ of x , called the *test function*, that can only take the values 0, which corresponds to not rejecting H_0 , and 1, which corresponds to rejecting H_0 . In short:

$$\phi : \mathcal{X} \rightarrow \{0, 1\}.$$

As an example, we might decide to reject H_0 if we would find 3 or more defect chips in the 20 that we check. This would correspond to choosing the test function

$$\phi_1(x) = \begin{cases} 0 & \text{if } x = 0, 1, 2 \\ 1 & \text{if } x = 3, 4, \dots, 20. \end{cases}$$

The question is whether or not this is a reasonable test function. We would like to obey the following rule: if the null hypothesis is true, the probability of rejecting H_0 should be small. Of course we have to specify what is small; this will be done by choosing a *significance level* α between 0 and 1. For a given test function ϕ we immediately check that

$$P(\text{rejecting } H_0) = E[\phi(X)]. \quad (1.1)$$

Quick exercise 1.1 Check Equation (1.1) directly from the definition of a test function.

We want to have that the probability of rejecting H_0 when H_0 is true, is smaller than the significance level α (common choices for α are 0.01, 0.05 or 0.1, but it very much depends on the situation). However, we have to realize that the distribution of the random variable $\phi(X)$ depends on which parameter $p \in \Theta$ we consider. To denote this dependence we will use the subscript p , so $E_p[\phi(X)]$ denotes the expectation of $\phi(X)$ when $X \sim \text{Bin}(20, p)$. Likewise we will sometimes use notations like $P_p(X \geq 3)$. Using this notation we see that in order for ϕ to be a reasonable test function, so in order to have $P(\text{rejecting } H_0) \leq \alpha$ whenever H_0 is true, we must have

$$E_p[\phi(X)] \leq \alpha \quad \forall p \in H_0. \quad (1.2)$$

Let us check this for the test function ϕ_1 we chose in our example, if we choose our significance level $\alpha = 0.05$. It is clear that

$$E_p[\phi_1(X)] \leq E_{0.06}[\phi_1(X)] \quad \forall p \in H_0,$$

so we only need to check (1.2) for $p = 0.06$. Consider Table 1.1.

Table 1.1: The $\text{Bin}(20, 0.06)$ distribution.

k	$P(X \leq k)$	k	$P(X \leq k)$	k	$P(X \leq k)$
0	0.2901	7	1.0000	14	1.0000
1	0.6605	8	1.0000	15	1.0000
2	0.8850	9	1.0000	16	1.0000
3	0.9710	10	1.0000	17	1.0000
4	0.9944	11	1.0000	18	1.0000
5	0.9991	12	1.0000	19	1.0000
6	0.9999	13	1.0000	20	1.0000

Since $\phi_1(x) = 1$ precisely when $x \geq 3$, we get that

$$E_{0.06}[\phi_1(X)] = P_{0.06}(X \geq 3) = 1 - P_{0.06}(X \leq 2) = 0.1150.$$

This shows that this particular function is not a reasonable choice for a test function: the probability that we will reject H_0 when in fact it is true, is higher than 0.05, our significance level. Property (1.2) will from now on be incorporated into the definition of a test function (all formal definitions will follow after this introductory example).

Apparently, even when we find 3 defect chips in our batch of 20, it is not enough to conclude that $p > 0.06$, at least not if we choose our significance level to be 5%. From Tabel 1.1 we can easily see that if we define

$$\phi_2(x) = \begin{cases} 0 & \text{if } x = 0, 1, 2, 3 \\ 1 & \text{if } x = 4, 5, \dots, 20, \end{cases}$$

then this is a reasonable test function. It rejects H_0 whenever we find 4 or more defect chips. We also have for all $p \in H_0$:

$$E_p[\phi_2(X)] \leq E_{0.06}[\phi_2(X)] = 1 - P_{0.06}(X \leq 3) = 0.0290.$$

This shows an interesting phenomenon: it seems that our test function ϕ_2 is too conservative. We are willing to accept a probability of a type I error, i.e. rejecting H_0 if in fact H_0 is true, of 5%, but in this case the probability of a type I error is always smaller than or equal to 0.0290, so it seems that we would be willing to reject H_0 more often. However, making ϕ_2 less conservative by also rejecting H_0 if we find 3 defect chips goes too far: the probability of a type I error might then be bigger than 5%.

In order to be able to find a test function that is not too conservative but still reasonable, we extend the notion of a test function. It will still be a function on the data space $\mathcal{X}$, but it can now take any value in $[0, 1]$. We then interpret this value as the *probability of rejecting H_0* . Let's clarify this idea by considering the test function

$$\phi_3(x) = \begin{cases} 0 & \text{if } x = 0, 1, 2 \\ 0.2446 & \text{if } x = 3 \\ 1 & \text{if } x = 4, 5, \dots, 20. \end{cases}$$

This means that if we find 2 or less defect chips, we do not reject H_0 , if we find 3 defect chips we reject H_0 with 0.2446 probability and if we find 4 or more defect chips we always reject H_0 .

What is the upshot of allowing such a test function? Consider $P_p(\text{rejecting } H_0)$ for $p \in H_0$, using test function ϕ_3 . It is not hard to see that

$$P_p(\text{rejecting } H_0) \leq P_{0.06}(\text{rejecting } H_0) \quad \forall p \in H_0.$$

Quick exercise 1.2 Prove this last statement.

Furthermore,

$$\begin{aligned} P_{0.06}(\text{rejecting } H_0) &= P_{0.06}(X \geq 4) + P(\text{rejecting } H_0 \mid X = 3) \cdot P_{0.06}(X = 3) \\ &= 0.0290 + 0.2446 \cdot 0.0860 \\ &= 0.05. \end{aligned}$$

This shows that the test function ϕ_3 is to be preferred over the test function ϕ_2 , because it is able to reject H_0 more often, without violating the rule that the probability of a type I error must be smaller than or equal to 5%. We will come back to this in the next section.

Tests that use test functions which do not only take the values 0 or 1 are sometimes called *randomized tests*, for obvious reasons: even though the outcome of our experiment is known, whether we reject H_0 or not might still be random. This is the reason why these tests are not very widely used in practice, since they seem to bring some arbitrariness into the test. After all, when we find 3 defect chips, we don't know whether we will reject or not until we do some extra random experiment, which is independent of our data and therefore does not contain any information about the data. However, we might look at it from another way. We choose our test function before actually performing the experiment. At that time we can introduce a $U(0, 1)$ random variable U , independent of the data X , and say that we will reject H_0 if $\phi(X) \geq U$. This U

we will call a *decision variable*. Then we take a realization of U , and now the outcome of the test will be clear as soon as we have performed our experiment to get a realization of X . Saying that a different realization of U changes the outcome of the test is now comparable to saying that a different realization of X changes the outcome of the test, a property of the test that no one would object to. It is clear, however, that in our example, finding 3 defect chips will always be a difficult case.

1.2 Definitions

We will consider an *experiment* on a *data space* $\mathcal{X}$ to be a random variable X (the *data* or the outcome of the experiment will be a realization of X) with values in $\mathcal{X}$, together with a *statistical model* Θ , which describes the possible distributions of X we will consider. In this book we will only consider two types of possible dataspace $\mathcal{X}$: the discrete dataspace, where $\mathcal{X}$ is some finite or countable set, so that X will be a discrete random variable, and a continuous dataspace, where $\mathcal{X}$ is a subset of $\mathbb{R}^n$ and X is a continuous random variable. In the discrete case, every $\theta \in \Theta$ will describe a probability mass function $f_\theta(x)$ of X , so

$$f_\theta(x) = P_\theta(X = x) \quad (x \in \mathcal{X}).$$

We have seen an example of this in the previous section, where $\mathcal{X} = \{0, 1, \dots, 20\}$, $\Theta = [0, 1]$, and

$$f_\theta(k) = \binom{20}{k} \theta^k (1 - \theta)^{20-k} \quad (k = 0, 1, \dots, 20).$$

In the continuous case, each $\theta \in \Theta$ will describe a probability density $f_\theta(x)$ ($x \in \mathbb{R}^n$) for X , with $f_\theta(x) = 0$ if $x \notin \mathcal{X}$. An example might be that $\mathcal{X} = [0, \infty)$, $\Theta = (0, \infty)$ and each θ corresponds to an *Exp*(θ)-distribution for X , so

$$f_\theta(x) = \theta e^{-\theta x} \quad \text{for } x \geq 0.$$

Many experiments will work with what is called a *random sample*: in this case, the data consists of a series of n repeated measurements, so

$$X = (X_1, X_2, \dots, X_n).$$

Each X_i takes values in the *sample space*, which we will also denote by $\mathcal{X}$. The data space in this case will therefore be $\mathcal{X}^n$. We want to model the repeated measurements as being independent and identically distributed, often abbreviated to i.i.d., so each $\theta \in \Theta$ describes the distribution f_θ of each X_i in $\mathcal{X}$ and the distribution of X , sometimes denoted by $f_\theta^{(n)}$, is given by

$$f_\theta^{(n)}(x_1, \dots, x_n) = f_\theta(x_1) f_\theta(x_2) \dots f_\theta(x_n) \quad ((x_1, \dots, x_n) \in \mathcal{X}^n).$$

Here we use the well known fact that $X_1, \dots, X_n$ are independent, precisely when the density function (or the probability mass function) of $(X_1, \dots, X_n)$ is the product of the marginal density functions (resp. the marginal probability

mass functions). So if each X_i is $Exp(\theta)$ distributed, then the density of X is given by

$$f_{\theta}^{(n)}(x_1, \dots, x_n) = \theta^n e^{-\theta(x_1 + \dots + x_n)}.$$

A *null hypothesis* H_0 is a subset of Θ , that is

$$H_0 \subset \Theta.$$

An *alternative hypothesis* to H_0 is a subset $H_1 \subset \Theta$ such that $H_0 \cap H_1 = \emptyset$. After we choose a *significance level* α , a function

$$\phi : \mathcal{X} \rightarrow [0, 1]$$

is called a *test function* if

$$E_{\theta}[\phi(X)] \leq \alpha \quad \forall \theta \in H_0. \quad (1.3)$$

The interpretation of the test function is that after we find the data x , that is a realization of X , we reject H_0 with probability $\phi(x)$. Condition (1.3) ensures that the probability of rejecting H_0 while H_0 is in fact true (i.e. the probability of a type I error), is always smaller than or equal to α .

If we allow randomized tests, so we allow test functions to take values in $(0, 1)$, then we can show that for every test function ϕ we can find a test function $\phi' \geq \phi$ such that

$$\sup_{\theta \in H_0} E_{\theta}[\phi'(X)] = \alpha. \quad (1.4)$$

The stronger condition (1.4) is sometimes used instead of condition (1.3).

Most non-randomized test functions we will encounter can be written in the following form: there exists a *test statistic* T , which is nothing but a real-valued function of the data, so

$$T : \mathcal{X} \rightarrow \mathbb{R},$$

and a *critical region* $C \subset \mathbb{R}$ for our test statistic such that our test function ϕ can be written as

$$\phi(x) = 1_{\{T(x) \in C\}} = \begin{cases} 0 & \text{if } T(x) \notin C \\ 1 & \text{if } T(x) \in C. \end{cases}$$

This means that we reject H_0 if and only if our test statistic falls in the critical region. With a slight abuse of notation we will also write T for the random variable $T(X)$ and we will be interested in the distribution of T . Note that since

$$E_{\theta}[\phi(X)] = P_{\theta}(T \in C),$$

condition (1.3) can be rewritten for a test based on a test statistic:

$$P_{\theta}(T \in C) \leq \alpha \quad \forall \theta \in H_0. \quad (1.5)$$

Two test statistics (and their respective critical regions) are called *equivalent* if they reject for the same values of X . An example of this, which we will see

several times in this book, is when the critical region C of a test statistic T is of the form

$$C = [c, \infty).$$

Now consider a strictly increasing function h and define $T' = h(T)$. If we define $c' = h(c)$, then $T(X) \geq c \Leftrightarrow T'(X) \geq c'$, which shows that T and T' are equivalent, if we define $C' = [c', \infty)$ as the critical region for T' . Note that T and T' are also equivalent if the critical region is of the form $C = (-\infty, c]$.

Let's review one more famous example and see how our notations work. Suppose we model the length of male students by a $N(\mu, \sigma^2)$ distribution. We want to test the null hypothesis

$$H_0 : \mu \geq 1.80\text{m}$$

against the alternative

$$H_1 : \mu < 1.80\text{m},$$

with a significance level of $\alpha = 5\%$. In order to do this we measure n students. In this case we are working with a random sample. The sample space is

$$\mathcal{X} = \mathbb{R}.$$

We model our random variables $X_1, \dots, X_n$, the length of the n students, by a normal distribution, so

$$\Theta = \mathbb{R} \times (0, \infty)$$

and for $\theta = (\mu, \sigma)$ we get

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

This means that the density of our data $X = (X_1, \dots, X_n)$ in $\mathbb{R}^n$ is given by

$$f_\theta^{(n)}(x_1, \dots, x_n) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2}.$$

The well known test statistic for this test is the studentized mean, so

$$T = \sqrt{n} \frac{\bar{X}_n - 1.80}{S_n}$$

where

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

If $\mu = 1.80$, then T has a $t(n-1)$ distribution. Remember that the critical value $t_{n-1, \alpha}$ is defined such that if $Y \sim t(n-1)$, then

$$P(Y \geq t_{n-1, \alpha}) = \alpha.$$

Therefore we choose the critical region

$$C = (-\infty, -t_{n-1, \alpha}].$$

We reject our hypothesis when the value of our test statistic lies in C . We have to check condition (1.5): for $(\mu, \sigma) \in H_0$ we have

$$\begin{aligned} P_{(\mu, \sigma)}(T \leq -t_{n-1, \alpha}) &= P_{(\mu, \sigma)}\left(\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \leq -t_{n-1, \alpha} - \sqrt{n} \frac{\mu - 1.80}{S_n}\right) \\ &\leq P_{(\mu, \sigma)}\left(\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \leq -t_{n-1, \alpha}\right) \\ &= \alpha. \end{aligned}$$

1.3 Power of a test

Suppose John and Mary have data $x_1, \dots, x_{20}$, which they model as a random sample $X_1, \dots, X_{20}$ from an $Exp(\theta)$ distribution, where θ is unknown. They want to test the hypothesis

$$H_0 : \theta = 1$$

against

$$H_1 : \theta \neq 1.$$

Mary knows that $E_\theta[X_1] = 1/\theta$, so if $\bar{X}_{20}$ is far away from 1, this might be an indication to reject H_0 . Mary therefore suggests as a test statistic

$$T_1 = |\bar{X}_{20} - 1|.$$

However, John notices that

$$P_\theta(X_1 \geq \ln(2)/\theta) = \frac{1}{2},$$

so he figures that if the median of the data is far away from $\ln(2)$, this might be an indication to reject H_0 . He therefore suggests as a test statistic

$$T_2 = |\text{Med}(X_1, \dots, X_{20}) - \ln(2)|.$$

They use the computer to find two positive numbers c_1 and c_2 such that

$$P_{\theta=1}(T_1 > c_1) = 0.05 \quad \text{and} \quad P_{\theta=1}(T_2 > c_2) = 0.05.$$

Quick exercise 1.3 Describe precisely how you would find c_1 and c_2 .

In this way they end up with two valid test functions

$$\phi_1(X_1, \dots, X_n) = 1_{\{T_1 > c_1\}} \quad \text{and} \quad \phi_2(X_1, \dots, X_n) = 1_{\{T_2 > c_2\}}.$$

Which test function should they use? In fact, how can we compare these two test functions? To do this, we must realize one thing: we think a test is better if, when H_1 is true, it rejects H_0 more often. This means that the probability of rejecting H_0 should be bigger when $\theta \in H_1$. This leads us to the following definition:

POWER OF A TEST. Suppose that we test H_0 against H_1 , using a test function ϕ . The *power* β of this test is the following function on Θ :

$$\beta : \Theta \rightarrow [0, 1] : \beta(\theta) = E_\theta[\phi(X)].$$

So $\beta(\theta)$ is the probability of rejecting H_0 when the true parameter is θ . In particular,

$$\beta(\theta) \leq \alpha \quad \forall \theta \in H_0.$$

We say that a test function ϕ is *uniformly more powerful* than a test function ϕ' if

$$E_\theta[\phi(X)] \geq E_\theta[\phi'(X)] \quad \forall \theta \in H_1.$$

In other words, $\beta(\theta) \geq \beta'(\theta)$ for all $\theta \in H_1$. Note that ϕ only needs to be more powerful on the alternative hypothesis H_1 , since those are the values of θ for which we would like to reject.

We return to our example. In Figure 1.1 we plot β_1 (solid line) and β_2 (dotted line), the power of the test using T_1 and T_2 , respectively.

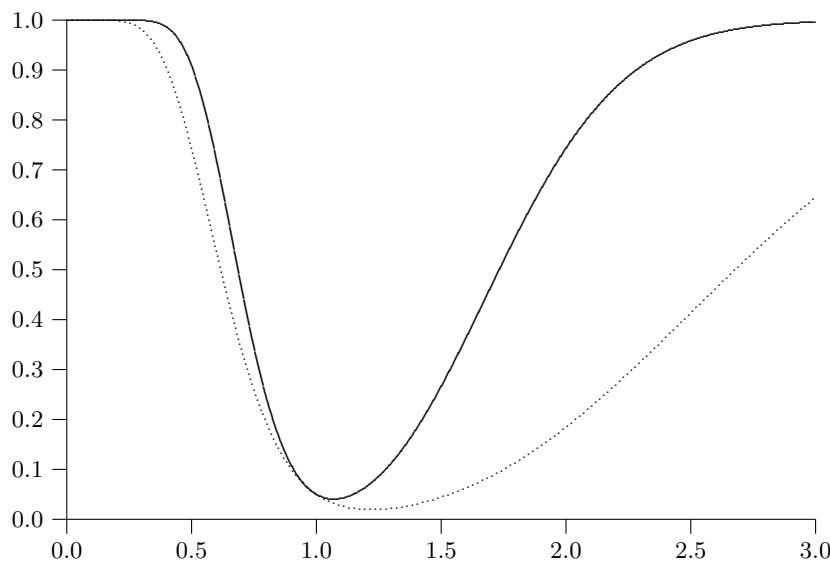


Figure 1.1: Power of T_1 and T_2 .

There are a couple of things we notice about this picture. First of all, it seems clear that the test using T_1 , so the one based on $\bar{X}_n$, performs better than the test using T_2 , based on $\text{Med}(X_1, \dots, X_n)$, but it is in fact not uniformly more powerful: there is a small interval just to the left of $\theta = 1$ where T_2 has (slightly) bigger power. The fact that the test using $\bar{X}_n$ performs better than the test using $\text{Med}(X_1, \dots, X_n)$ is directly related to the fact that the variance of $\bar{X}_n$ is smaller than the variance of $\text{Med}(X_1, \dots, X_n)$.

Both tests have power 5% at $\theta = 1$ and both tests actually have lower power than 5% for θ a little to the right of 1.

1.4 Neyman-Pearson test

Ideally, we would like to use a test which is *uniformly most powerful*, which means that for any other test, our test is more powerful for all $\theta \in H_1$. More formally, given a null hypothesis H_0 and an alternative hypothesis H_1 , a test function ϕ is called uniformly most powerful if for any other test function ϕ' we have

$$E_{\theta}[\phi(X)] \geq E_{\theta}[\phi'(X)] \quad \forall \theta \in H_1.$$

Remember that any test function ϕ has to satisfy

$$E_{\theta}[\phi(X)] \leq \alpha \quad \forall \theta \in H_0,$$

where α is the significance level.

In general, such a uniformly most powerful test does not exist; this is the case when for any test function we can find another test function that is more powerful for some $\theta \in H_1$. However, when H_1 consists of one point (H_1 is then called a *simple hypothesis*), we can show that there exists a uniformly most powerful test. The following famous lemma tells us what this test looks like in case H_0 is also a simple hypothesis.

Lemma 1.1 (Neyman-Pearson Lemma) *Let X be a continuous or discrete space. We are given two distributions of X , namely with densities (or probability mass functions) $f_0(x)$ and $f_1(x)$, and we want to test $H_0 = \{f_0\}$ against $H_1 = \{f_1\}$ at significance level $0 < \alpha < 1$. Then the test function*

$$\phi(x) = \begin{cases} 1 & \text{if } f_1(x)/f_0(x) > C \\ \gamma & \text{if } f_1(x)/f_0(x) = C \\ 0 & \text{if } f_1(x)/f_0(x) < C \end{cases}$$

where $C > 0$ and $\gamma \in [0, 1]$ are chosen such that

$$E_{f_0}[\phi(X)] = \alpha,$$

is the uniformly most powerful test.

Proof: We will show the Lemma only in the continuous case; the discrete case is completely analogous. Let ϕ be the Neyman-Pearson test function and let ϕ' be some other test function. Consider the difference of the power of the two

tests at f_1 :

$$\begin{aligned}
E_{f_1}[\phi(X)] - E_{f_1}[\phi'(X)] &= \int (\phi(x) - \phi'(x))f_1(x)dx \\
&= \int_{\{f_1(x) > Cf_0(x)\}} (1 - \phi'(x))f_1(x)dx - \int_{\{f_1(x) < Cf_0(x)\}} \phi'(x)f_1(x)dx + \\
&\quad + \int_{\{f_1(x) = Cf_0(x)\}} (\phi(x) - \phi'(x))f_1(x)dx \\
&\geq C \int_{\{f_1(x) > Cf_0(x)\}} (1 - \phi'(x))f_0(x)dx - C \int_{\{f_1(x) < Cf_0(x)\}} \phi'(x)f_0(x)dx + \\
&\quad + C \int_{\{f_1(x) = Cf_0(x)\}} (\phi(x) - \phi'(x))f_0(x)dx \\
&= C \int (\phi(x) - \phi'(x))f_0(x)dx.
\end{aligned}$$

For the inequality we use for the first term that $\phi' \leq 1$ and for the second term that $\phi' \geq 0$. Now we note that ϕ' is a test function, so

$$\int \phi'(x)f_0(x)dx \leq \alpha = \int \phi(x)f_0(x)dx.$$

This proves that

$$E_{f_1}[\phi(X)] \geq E_{f_1}[\phi'(X)],$$

so indeed that ϕ is the uniformly most powerful test. $\square$

Consider the example from the introduction. The manufacturer of a batch of microchips claims that 6% of the chips are defect. However, a consumer organization claims that actually 9% of the chips of this manufacturer are defect. We check 20 chips and we would like to know what is the Neyman-Pearson test function ϕ for testing $H_0 : p = 0.04$ against $H_1 : p = 0.09$ at significance level 5%. We know that

$$f_0(k) = \binom{20}{k} (0.06)^k (0.94)^{20-k} \quad (k = 0, 1, \dots, 20)$$

and

$$f_1(k) = \binom{20}{k} (0.09)^k (0.91)^{20-k} \quad (k = 0, 1, \dots, 20),$$

so we get

$$f_1(k)/f_0(k) = (1.5)^k (0.968)^{20-k} = (1.55)^k (0.968)^{-20}.$$

This is an increasing function of k , so ϕ will be of the form

$$\phi(k) = \begin{cases} 1 & \text{if } k > k_0 \\ \gamma & \text{if } k = k_0 \\ 0 & \text{if } k < k_0. \end{cases}$$

The constants γ and k_0 are determined by

$$E_{0.06}[\phi(X)] = 0.05.$$

Using Table 1.1 it is not hard to see that

$$\phi(x) = \begin{cases} 0 & \text{if } x = 0, 1, 2 \\ 0.2446 & \text{if } x = 3 \\ 1 & \text{if } x = 4, 5, \dots, 20. \end{cases}$$

Quick exercise 1.4 Calculate the power of the test at $p = 0.09$.

Note that whenever $p > 0.06$ and $H_1 = \{p\}$, we would get that

$$f_1(k)/f_0(k) = \left(\frac{47p}{3-3p}\right)^k \left(\frac{1-p}{0.94}\right)^{20}$$

is increasing in k (since $p > 0.06$ implies that $47p > 3 - 3p$), and therefore we would end up with the same ϕ as the uniformly most powerful test function. This shows that ϕ is in fact uniformly most powerful for the test $H_0 = \{0.06\}$ against $H_1 = (0.06, 1]$. Since Quick exercise (1.2) shows that the power $\beta(p)$ is increasing when $p \leq 0.06$, we get that ϕ is also the uniformly most powerful test function for testing $H_0 : p \leq 0.06$ against $H_1 : p > 0.06$. For if ϕ' is a test function for this test, it is also a test function for the test $H_0 : p = 0.06$ against $H_1 : p > 0.06$, and therefore the power of ϕ' in H_1 is smaller than or equal to the power of ϕ . This phenomenon is only true because of the special structure of our statistical model.

1.5 Likelihood ratio test

The Neyman-Pearson Lemma is not very useful in practical situations, because it only works for simple hypotheses. However, we can use it as an inspiration to find a test which can be used in almost any testing problem and which has many good properties: the likelihood ratio test.

THE LIKELIHOOD RATIO TEST. Suppose we are testing H_0 against H_1 . The likelihood ratio L is a test statistic defined by

$$L(x) = \frac{\sup_{\theta \in H_1} f_{\theta}(x)}{\sup_{\theta \in H_0} f_{\theta}(x)}$$

where f_{θ} is the density function (or probability mass function) of X described by θ . The likelihood ratio test function ϕ for a significance level α is defined by

$$\phi(x) = \begin{cases} 1 & \text{if } L(x) > C \\ \gamma & \text{if } L(x) = C \\ 0 & \text{if } L(x) < C, \end{cases}$$

where $C \geq 0$ and $\gamma \in [0, 1]$ are chosen such that

$$\sup_{\theta \in H_0} E_{\theta}[\phi(X)] = \alpha.$$

Note that large values of L point in the direction of the alternative hypothesis, because when L is large, it means that we can make the likelihood of the data much bigger by choosing a model in H_1 , indicating that H_1 “fits” the data better. It is immediately clear that if we suppose that $H_0 = \{f_0\}$ and $H_1 = \{f_1\}$, we get the Neyman-Pearson test.

Let us see how the likelihood ratio test works in our example of Section 1.3. John and Mary had $n = 20$ data points, all independent realizations of an $Exp(\theta)$ distribution, where $\theta > 0$. Furthermore, $H_0 = \{1\}$ and $H_1 = (0, 1) \cup (1, \infty)$. Given some $\theta > 0$, the likelihood is given by

$$f_\theta^{(n)}(x_1, \dots, x_n) = \theta^n e^{-\theta(x_1 + \dots + x_n)}.$$

Obviously,

$$\sup_{\theta \in H_0} f_\theta^{(n)}(x) = e^{-(x_1 + \dots + x_n)}.$$

In this case (and in many cases) there is a direct connection between the likelihood ratio statistic and the maximum likelihood estimator for θ . Indeed, the maximum likelihood estimator $\hat{\theta}$ maximizes the likelihood $f_\theta^{(n)}(x)$ over all possible $\theta \in \Theta$, but since in this case we have

$$\sup_{\theta \in H_1} f_\theta^{(n)}(x) = \sup_{\theta \in \Theta} f_\theta^{(n)}(x),$$

we conclude that

$$\sup_{\theta \in H_1} f_\theta^{(n)}(x) = f_{\hat{\theta}}^{(n)}(x).$$

It is well known (and can be easily verified) that here

$$\hat{\theta} = 1/\bar{x}_n,$$

which means that

$$f_{\hat{\theta}}^{(n)}(x) = \left(\frac{1}{\bar{x}_n}\right)^n e^{-\frac{1}{\bar{x}_n}(x_1 + \dots + x_n)} = e^{-n(1 + \log(\bar{x}_n))}.$$

So finally our likelihood ratio statistic becomes

$$L(x_1, \dots, x_n) = e^{n(\bar{x}_n - 1 - \log(\bar{x}_n))}.$$

Our test function ϕ will therefore be based on the statistic $T(x) = \bar{x}_n - \log(\bar{x}_n)$, since $L(x)$ is a strictly increasing function of $T(x)$. So ϕ will have the following form:

$$\phi(x) = \begin{cases} 1 & \text{if } \bar{x}_n - \log(\bar{x}_n) > C \\ 0 & \text{if } \bar{x}_n - \log(\bar{x}_n) \leq C, \end{cases}$$

for some $C > 1$ such that

$$E_1[\phi(X_1, \dots, X_n)] = 0.05.$$

Note that since $P_1(\bar{X}_n - \log(\bar{X}_n) = C) = 0$ for any $C \geq 0$, we do not need the extra γ which was in the definition of the likelihood ratio test; in fact, we do not need a randomized test. To find C we use the computer and find

$$C = 1.0968356.$$

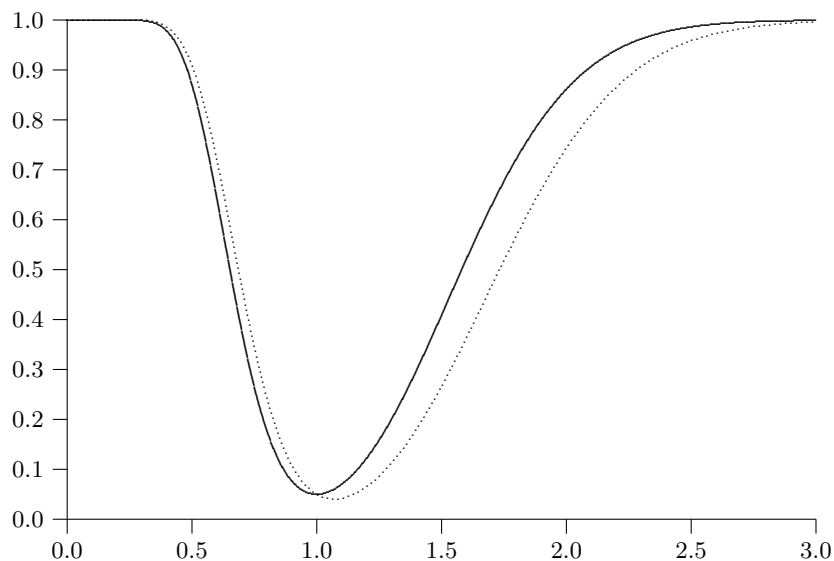


Figure 1.2: Power of likelihood ratio test and T_1 .

We plotted the power of the likelihood ratio test (drawn line) and the test based on $T_1(x) = |\bar{x}_n - 1|$ (dashed line) in Figure 1.2. We see that the likelihood ratio test has significantly more power for $\theta > 1$, whereas T_1 is slightly more powerful for $\theta < 1$. This becomes more apparent if we look at Figure 1.3, where we plotted the difference of the two powerfunctions. Another attractive feature

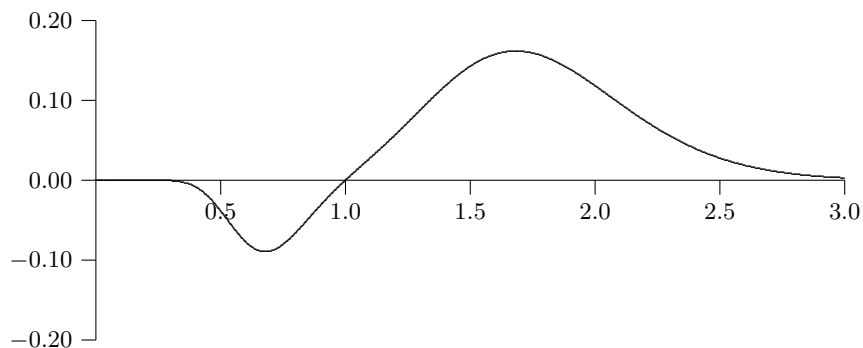


Figure 1.3: Difference of power.

of the likelihood ratio test in this example is that the power is greater than $\alpha = 0.05$ for all $\theta \in H_1$. A test with this property is sometimes called *unbiased*. In general we can say that if $\Theta \subset \mathbb{R}^p$, then, under some regularity properties, the likelihood ratio test is asymptotically optimal when working with samples, which means that when the sample size becomes large, the likelihood ratio test becomes almost a uniformly most powerful unbiased test. In fact, the likelihood ratio test is directly connected to the maximum likelihood estimator, and since this has good properties in many models, the likelihood ratio test can be expected to be a good candidate for a test in almost any model.

1.6 P-value

When choosing a test function, we often first choose a test statistic, for example the likelihood ratio, and then we choose our parameters of the test function in such a way that

$$\sup\{E_\theta[\phi(X)] : \theta \in H_0\} = \alpha.$$

This means that for each value of α we get a different test function; let's denote this test function by $\phi^{(\alpha)}$. In principle we are free to choose this family of test functions $\{\phi^{(\alpha)} : \alpha \in [0, 1]\}$, which are all test functions for the same testing problem, but at different significance levels. However, we will assume two conditions on such a family: $\phi^{(1)} \equiv 1$ and

$$\alpha_1 \leq \alpha_2 \Rightarrow \phi^{(\alpha_1)} \leq \phi^{(\alpha_2)}.$$

A family of test functions that satisfies these conditions is called a *monotone family* of test functions. These conditions are always met if we use the likelihood ratio test and define $\phi^{(1)} = 1$. To see this, recall that if $L(x)$ is the likelihood ratio, the test function is defined by

$$\phi^{(\alpha)}(x) = \begin{cases} 1 & \text{if } L(x) > C(\alpha) \\ \gamma(\alpha) & \text{if } L(x) = C(\alpha) \\ 0 & \text{if } L(x) < C(\alpha), \end{cases}$$

where $C(\alpha) \geq 0$ and $\gamma(\alpha) \in [0, 1]$ are chosen such that

$$\sup_{\theta \in H_0} E_\theta[\phi^{(\alpha)}(X)] = \alpha.$$

If $C(\alpha_1) < C(\alpha_2)$, then $\phi^{(\alpha_1)} \geq \phi^{(\alpha_2)}$. For if $x > C(\alpha_1)$, then $\phi^{(\alpha_1)}(x) = 1$, and if $x \leq C(\alpha_1) < C(\alpha_2)$, then $\phi^{(\alpha_2)}(x) = 0$. This proves that if $\alpha_1 < \alpha_2$, then $C(\alpha_1) < C(\alpha_2)$. If $C(\alpha_1) > C(\alpha_2)$, then we would have that $\phi^{(\alpha_1)} \leq \phi^{(\alpha_2)}$. Now suppose $C(\alpha_1) = C(\alpha_2)$. Then $\gamma(\alpha_1) \geq \gamma(\alpha_2)$ would imply that $\phi^{(\alpha_1)} \geq \phi^{(\alpha_2)}$, which would mean that $\alpha_1 \geq \alpha_2$. Therefore, we must have that $\gamma(\alpha_1) < \gamma(\alpha_2)$, and hence $\phi^{(\alpha_1)} \leq \phi^{(\alpha_2)}$.

It should be clear from the above that any family of test functions that is defined in a similar way based on some other statistic, will also be a monotone family. Now we are ready to define the P-value:

P-VALUE. Suppose that we test H_0 against H_1 , using a monotone family of test functions $\phi^{(\alpha)}$. Choose U uniform in $[0, 1]$. The *P-value* p of our data x is defined as:

$$p = \operatorname{argmin}\{\alpha : \phi^{(\alpha)}(x) > U\}.$$

So p is the lowest value of the significance level for which we still reject H_0 , when using the decision variable U .

The P-value gives you more information than just saying whether you reject a test or not (note that if $p < \alpha$, you reject H_0): a very low P-value gives you a lot of confidence that you correctly rejected H_0 . To calculate the P-value, we can often use the following proposition:

Proposition 1.2 Suppose we want to test H_0 based on a test statistic L in the following way: we choose our test function $\phi^{(\alpha)}$ at significance level α to be

$$\phi^{(\alpha)}(x) = \begin{cases} 1 & \text{if } L(x) > C(\alpha) \\ \gamma(\alpha) & \text{if } L(x) = C(\alpha) \\ 0 & \text{if } L(x) < C(\alpha), \end{cases}$$

where $C(\alpha) \geq 0$ and $\gamma(\alpha) \in [0, 1]$ are chosen such that

$$\sup_{\theta \in H_0} E_{\theta} [\phi^{(\alpha)}(X)] = \alpha.$$

Then the P -value p of the data x is given by:

$$p = \sup_{\theta \in H_0} [P_{\theta}(L(X) > L(x)) + UP_{\theta}(L(X) = L(x))].$$

Proof: Define p as in the last line of the proposition. First we note that

$$E_{\theta} [\phi^{(\alpha)}(X)] = P_{\theta}(L(X) > C(\alpha)) + \gamma(\alpha)P_{\theta}(L(X) = C(\alpha)).$$

This means that

$$\alpha = \sup_{\theta \in H_0} [P_{\theta}(L(X) > C(\alpha)) + \gamma(\alpha)P_{\theta}(L(X) = C(\alpha))].$$

Now suppose that we reject H_0 at significance level α . This means that either $L(x) > C(\alpha)$ or $L(x) = C(\alpha)$ and $\gamma(\alpha) > U$. Suppose $L(x) > C(\alpha)$. Then

$$\begin{aligned} \alpha &\geq \sup_{\theta \in H_0} [P_{\theta}(L(X) > C(\alpha))] \\ &\geq \sup_{\theta \in H_0} [P_{\theta}(L(X) \geq L(x))] \\ &\geq \sup_{\theta \in H_0} [P_{\theta}(L(X) > L(x)) + UP_{\theta}(L(X) = L(x))]. \end{aligned}$$

So $\alpha \geq p$. Now suppose that $L(x) = C(\alpha)$ and $\gamma(\alpha) > U$. Then

$$\begin{aligned} \alpha &= \sup_{\theta \in H_0} [P_{\theta}(L(X) > L(x)) + \gamma(\alpha)P_{\theta}(L(X) = L(x))] \\ &\geq \sup_{\theta \in H_0} [P_{\theta}(L(X) > L(x)) + UP_{\theta}(L(X) = L(x))]. \end{aligned}$$

This shows that $\alpha \geq p$ for all α for which we can reject H_0 . Therefore, p is smaller than or equal to the P -value.

Now suppose that we cannot reject H_0 for some significance level α . This means that either $L(x) < C(\alpha)$ or $L(x) = C(\alpha)$ and $\gamma(\alpha) \leq U$. Suppose $L(x) < C(\alpha)$. Then

$$\begin{aligned} \alpha &\leq \sup_{\theta \in H_0} [P_{\theta}(L(X) \geq C(\alpha))] \\ &\leq \sup_{\theta \in H_0} [P_{\theta}(L(X) > L(x))] \\ &\leq \sup_{\theta \in H_0} [P_{\theta}(L(X) > L(x)) + UP_{\theta}(L(X) = L(x))]. \end{aligned}$$

So $\alpha \leq p$. Now suppose that $L(x) = C(\alpha)$ and $\gamma(\alpha) \leq U$. Then

$$\begin{aligned}\alpha &= \sup_{\theta \in H_0} [\mathbb{P}_\theta(L(X) > L(x)) + \gamma(\alpha)\mathbb{P}_\theta(L(X) = L(x))] \\ &\leq \sup_{\theta \in H_0} [\mathbb{P}_\theta(L(X) > L(x)) + U\mathbb{P}_\theta(L(X) = L(x))].\end{aligned}$$

This shows that $\alpha \leq p$ for all α for which we cannot reject H_0 . Therefore, p must be equal to the P-value. $\square$

Note that if $L(X)$ is a continuous random variable, then our decision variable U is irrelevant, and we get that

$$\text{P-value} = \sup_{\theta \in H_0} [\mathbb{P}_\theta(L(X) > L(x))]. \quad (1.6)$$

Let us revisit the following example. Suppose we model the length of male students as $N(\mu, \sigma^2)$ distributed, with unknown μ and σ . We measure $n = 20$ male students and find $\bar{x}_n = 178\text{cm}$ and

$$s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2} = 3.2\text{cm}.$$

We want to test

$$H_0 : \mu \geq 180 \quad \text{against} \quad H_1 : \mu < 180.$$

First we check what our likelihood ratio statistic will be:

$$L(x) = \frac{\sup_{(\mu, \sigma) \in H_1} \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2}}{\sup_{(\mu, \sigma) \in H_0} \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2}}.$$

In this case it is easier to work with the log-likelihood ratio $l(x) = \log(L(x))$:

$$l(x) = \sup_{(\mu, \sigma) \in H_1} \left[-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2 - n \log(\sigma) \right] - \sup_{(\mu, \sigma) \in H_0} \left[-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2 - n \log(\sigma) \right].$$

If we define

$$\chi(\mu, \sigma) = -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2 - n \log(\sigma),$$

we see that

$$\frac{d}{d\sigma} \chi(\mu, \sigma) = 0 \Leftrightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

This value for σ corresponds to the maximum of $\chi(\mu, \sigma)$ for fixed μ . Also,

$$\frac{d}{d\mu} \chi(\mu, \sigma) = 0 \Leftrightarrow \mu = \bar{x}_n.$$

This also corresponds to a maximum (for fixed σ). In this case, $(\bar{x}_n, \sigma) \in H_1$, so if we want to maximize over H_0 , we have to choose μ as close as possible to $\bar{x}_n$, since $\chi(\mu, \sigma)$ is quadratic in μ . Define

$$s_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \quad \text{and} \quad s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - 180)^2.$$

Then

$$\begin{aligned} l(x) &= -\frac{1}{2}n - n \log(s_1) + \frac{1}{2}n + n \log(s_0) \\ &= \frac{1}{2}n \log \left(\frac{\sum_{i=1}^n (x_i - 180)^2}{ns_1^2} \right) \\ &= \frac{1}{2}n \log \left(\frac{\sum_{i=1}^n ((x_i - \bar{x}_n) + (\bar{x}_n - 180))^2}{ns_1^2} \right) \\ &= \frac{1}{2}n \log \left(\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2 + \sum_{i=1}^n (\bar{x}_n - 180)^2}{ns_1^2} \right) \\ &= \frac{1}{2}n \log \left(1 + \frac{(\bar{x}_n - 180)^2}{s_1^2} \right). \end{aligned}$$

It is easy to see, that if $\bar{x}_n > 180$, then

$$l(x) = -\frac{1}{2}n \log \left(1 + \frac{(\bar{x}_n - 180)^2}{s_1^2} \right).$$

We reject H_0 if $l(x)$ is bigger than a critical value. However, $l(x)$ is a strictly decreasing function of

$$T(x) = \sqrt{n} \frac{\bar{x}_n - 180}{s_n}.$$

Quick exercise 1.5 Check that $l(x)$ is a strictly decreasing function of $T(x)$.

This means that we will reject H_0 if $T(x)$ is smaller than a critical value. The P-value will therefore also be given by $\sup_{(\mu, \sigma)} P_{(\mu, \sigma)}(T(X) < T(x))$. We know that $T(X)$ is $t(19)$ distributed (for every choice of (μ, σ)), and $T(x) = -2.795$. This means that

$$P(T(X) < T(x)) = 0.006.$$

So our P-value is 0.006, which gives us enough indication that we can reject H_0 . Note that we have proved that the test using the test statistic T is in fact equivalent with the likelihood ratio test.

1.7 Confidence regions for θ

Hypothesis testing can be used to define confidence regions for the parameter $\theta \in \Theta$. Suppose we have a test function ϕ_θ , at significance level α , for each of the following tests:

$$H_0 = \{\theta\} \quad \text{against} \quad H_1 = \Theta \setminus \{\theta\}.$$

A natural choice would be to take the likelihood ratio test function as ϕ_θ . Then we define:

CONFIDENCE REGION FOR θ . A $(1 - \alpha)$ -confidence region for $\theta \in \Theta$ is constructed by choosing a U uniform in $[0, 1]$ and defining the random set

$$C(X) = \{\theta \in \Theta : \phi_\theta(X) \leq U\}.$$

This is the set of all parameters θ for which the hypothesis

$$H_0 = \{\theta\} \text{ against } H_1 = \Theta \setminus \{\theta\}$$

cannot be rejected at significance level α , if we use the test function ϕ_θ and the decision variable U .

A $(1 - \alpha)$ -confidence region $C(X)$ has the following desirable property:

Lemma 1.3 (Covering probability) *If $C(X)$ is a $(1 - \alpha)$ -confidence region for θ , then*

$$P_\theta(\theta \in C(X)) \geq 1 - \alpha \quad \forall \theta \in \Theta.$$

If ϕ_θ is such that $E_\theta[\phi_\theta(X)] = \alpha$, then

$$P_\theta(\theta \in C(X)) = 1 - \alpha.$$

This says that a $(1 - \alpha)$ -confidence region has a probability of at least $1 - \alpha$ to cover (or contain) the “true” underlying parameter θ .

Proof of Lemma 1.3: Since $\theta \in C(X)$ precisely when $\phi_\theta(X) \leq U$, we get

$$\begin{aligned} P_\theta(\theta \in C(X)) &= P_\theta(\phi_\theta(X) \leq U) \\ &= 1 - P_\theta(\phi_\theta(X) > U) \\ &= 1 - \int_{\mathcal{X}} \int_0^1 1_{\{\phi_\theta(x) > u\}} du f_\theta(x) dx \\ &= 1 - \int_{\mathcal{X}} \phi_\theta(x) f_\theta(x) dx \\ &= 1 - E_\theta[\phi_\theta(X)]. \end{aligned}$$

This, together with the remark that for our test function ϕ_θ we must have

$$E_\theta[\phi_\theta(X)] \leq \alpha,$$

proves the lemma. □

As an example, let us consider the traffic flow of heavy trucks in one direction on a highway at night, so that there will be free flowing traffic with relatively low intensity. At a given time, the location of trucks on the road can be modelled as a Poisson process with intensity λ . We assume that each truck i drives at a constant speed V_i , independent of the other trucks, where the speed V is distributed according to some unknown density $f(v)$ such that $E[V] < \infty$.

Now we start our experiment at $t = 0$, and we denote the times a truck passes our fixed location $x = 0$ by the side of the road. We claim that these time points also behave as a Poisson process, but with intensity $\theta = \lambda E[V]$. To see

this (heuristically), first note that seeing a truck pass at time t with speed v , tells us that at time 0 there was a truck at location $-vt$. However, due to the properties of a Poisson process, this tells us nothing about the location or the speed of any of the others trucks, and therefore we have no information about other time intervals. So the number of trucks that pass in disjoint time intervals are independent. Consider the probability that a truck passes in the time interval $[t, t + h]$, for small h . We condition on the speed of the truck that passes:

$$\begin{aligned}
 P(\text{truck at } [t, t + h]) &= \int_0^\infty P(\text{truck at } [t, t + h] \mid \text{speed} = v) f(v) dv \\
 &= \int_0^\infty P(\text{truck started at } [-vt - vh, -vt] \mid \text{speed} = v) f(v) dv \\
 &= \int_0^\infty (\lambda vh + O(h^2)) f(v) dv \\
 &= \lambda E[V] h + O(h^2)
 \end{aligned}$$

This shows that the intensity is constant and equal to $\theta = \lambda E[V]$.

In our model the number of trucks that pass each hour is $Pois(\theta)$ distributed, if λ is in cars/mile and V is in mph. We measure the number of trucks passing for 6 hours. Our findings are in Table 1.2.

Table 1.2: The number of trucks passing each hour.

hour	1	2	3	4	5	6
number of cars	1	0	3	1	1	0

Based on these $n = 6$ measurements $X_1, \dots, X_6$, we want to find a confidence interval for θ . In practice most people will try to use the normal approximation: according to the Central Limit Theorem we know that

$$\frac{\bar{X}_n - \theta}{\sqrt{\theta/n}}$$

is approximately standard normal for big n (note that $E[X_1] = \theta$ and $\text{Var}(X_1) = \theta$). If we use this in our case, we get

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \theta}{\sqrt{\theta/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.$$

We can rewrite this equation as

$$P\left(\left(\theta - \left(\bar{X}_n + \frac{z_{\alpha/2}^2}{2n}\right)\right)^2 \leq \left(\bar{X}_n + \frac{z_{\alpha/2}^2}{2n}\right)^2 - \bar{X}_n^2\right) \approx 1 - \alpha.$$

For our data, this leads to a 95% confidence interval for θ given by

$$C_1 = [0.458, 2.18].$$

Since we use an approximation to the real distribution, we do not expect Lemma 1.3 to hold for $C_1(X)$. Indeed, we can calculate the covering probability $P_\theta(\theta \in C_1(X))$ for every $\theta > 0$, and we get Figure 1.4. The apparently

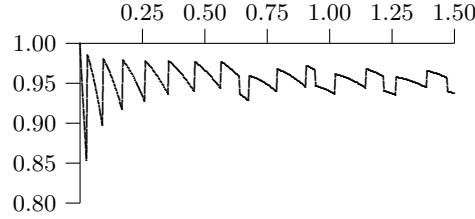


Figure 1.4: Coverage probability of normal confidence interval for different θ .

erratic nature of the covering probability is due to the discrete nature of the Poisson distribution. We see that the coverage probability can be as low as 85%, although for $\theta > 0.3$, the covering probability is always bigger than 93%.

Quick exercise 1.6 Give an exact formula for the coverage probability $P_\theta(\theta \in C_1(X))$, so that you can calculate it using a computer.

We now want to calculate an exact confidence interval for θ based on the likelihood ratio test. The likelihood is given by

$$f_\theta^{(n)}(x_1, \dots, x_n) = \frac{\theta^{x_1 + \dots + x_n} e^{-n\theta}}{x_1! \cdots x_n!}.$$

The maximum likelihood estimator for θ is given by

$$\hat{\theta} = \bar{x}_n.$$

This means that if $H_0 = \{\theta\}$ and $H_1 = (0, \infty) \setminus \{\theta\}$, our likelihood ratio statistic is given by

$$L(x) = \frac{\bar{x}_n^{n\bar{x}_n} e^{-n\bar{x}_n}}{\theta^{n\bar{x}_n} e^{-n\theta}} = e^{n\theta} e^{-n\bar{x}_n(1 + \log(\theta) - \log(\bar{x}_n))}.$$

If we define $S_n = X_1 + \dots + X_n$, then we know that $S_n \sim \text{Pois}(n\theta)$. Also,

$$L(X) = e^{n\theta} e^{n\theta e(\frac{S_n}{n\theta e} \log(\frac{S_n}{n\theta e}))}.$$

This shows that $L(X)$ is a strictly increasing function of $Y_n \log(Y_n)$, where

$$Y_n = \frac{S_n}{n\theta e}.$$

So our test function ϕ_θ has the following form:

$$\phi_\theta(x) = \begin{cases} 1 & \text{if } y_n \log(y_n) > K(\theta) \\ \gamma(\theta) & \text{if } y_n \log(y_n) = K(\theta) \\ 0 & \text{if } y_n \log(y_n) < K(\theta), \end{cases}$$

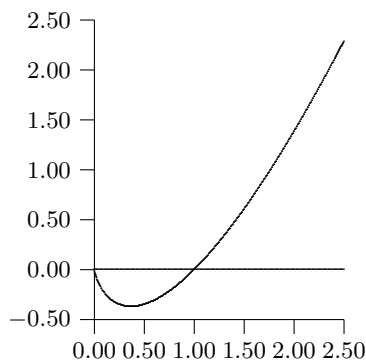


Figure 1.5: The function $x \mapsto x \log(x)$.

where $K(\theta) \in \mathbb{R}$ and $\gamma(\theta) \in [0, 1]$ are chosen such that

$$E_\theta[\phi_\theta(X)] = \alpha.$$

We should stop here and think a little bit if this test function makes sense. Figure 1.5 shows the function $x \mapsto x \log(x)$. If $K(\theta) < 0$, then we will reject $H_0 = \{\theta\}$ for very small values of y_n and for big values of y_n . In other words, we will reject if $\bar{x}_n$ is much smaller than θ or if $\bar{x}_n$ is much bigger than θ , which sounds reasonable. But if $K(\theta) > 0$, then we only reject for big y_n , so $\bar{x}_n$ is much bigger than θ . This means that even if we find $y_n = 0$ (so all $x_i = 0$), we still cannot reject H_0 . Of course this is possible if θ is very small. Indeed,

$$P_\theta(X_1 = 0, \dots, X_n = 0) = e^{-n\theta},$$

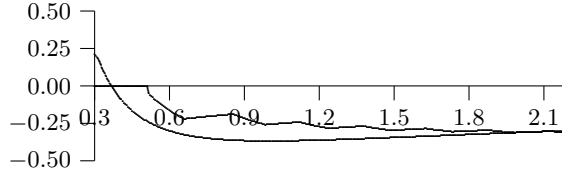
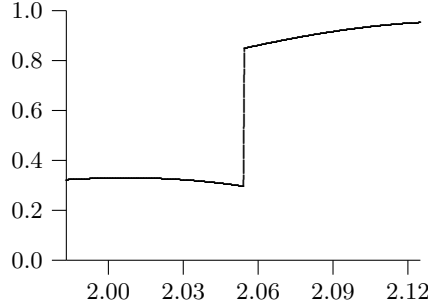
so if $e^{-n\theta} > \alpha$, or $\theta < -\log(\alpha)/n$, then we must have that $K(\theta) \geq 0$, or else the probability to reject H_0 when H_0 is true, will be bigger than α . Note that in general $K(\theta) \geq 0$ for even bigger θ (why?). Also note that $y_n \log(y_n)$ attains its minimum exactly when $\bar{x}_n = \theta$; can you see why this must be true?

Now we want to see how we can use these test functions to construct an exact confidence interval $C(x)$ for θ based on our data. In principle this means that we have to check for every $\theta > 0$ whether the null hypothesis $H_0 = \{\theta\}$ can be rejected using the test function ϕ_θ . So we also have to calculate $K(\theta)$ and $\gamma(\theta)$ for every θ and compare this to

$$y_n \log(y_n) = \frac{1}{e\theta} \log\left(\frac{1}{e\theta}\right).$$

Here we've use that for our data $s_n = 6$. It is a nice and challenging problem to give an exact solution for this, but we will leave this to the reader. Here we will just show some results: define $T(\theta) = -\log(e\theta)/(e\theta)$. In Figure 1.6 the two functions $K(\theta)$ and $T(\theta)$ are depicted.

We can show that $T(\theta) < K(\theta)$ when $\theta \in (1/e, 1.983148)$. Furthermore, $T(\theta) = K(\theta)$ when $\theta = 1/e$ or $\theta \in [1.983148, 2.125205)$. Now we need to know $\gamma(\theta)$ for θ in this last interval: see Figure 1.7.

**Figure 1.6:** The functions $K(\theta)$ and $T(\theta)$.**Figure 1.7:** The function $\gamma(\theta)$.

Let us take a realization of our decision variable U : $u = 0.32$. Using $\gamma(\theta)$ we can see that in this case

$$C(x) = (0.3637879, 1.983148) \cup [2.032871, 2.054353).$$

The confidence region we have constructed in this way has the property

$$P_{\theta}(\theta \in C(X)) = 1 - \alpha \quad \forall \theta > 0.$$

A remarkable feature of this region is that it has a positive probability of not being an interval, but a union of intervals. This is a somewhat pathological consequence of the fact that the Poisson distribution is discrete. The effect becomes negligible when θ is big or n is big.

1.8 Confidence region for $\rho(\theta)$ and one-sided confidence intervals

Sometimes you are not so much interested in the complete value of θ , but only in a real-valued *function* of the parameter θ . Let ρ be such a function, so

$$\rho : \Theta \rightarrow \mathbb{R}.$$

In case of the normal distribution, you could think of $\rho(\mu, \sigma) = \mu$, or in case $\Theta = \{f : \mathbb{R} \rightarrow [0, \infty) : \int_{\mathbb{R}} f(x)dx = 1\}$ (the set of all densities on $\mathbb{R}$), you could think of $\rho(f) = f(t)$, for some fixed $t \in \mathbb{R}$. We can use the ideas of the previous section to define a confidence region for $\rho(\theta)$. In this case, we want to test

$$H_0 = \{\theta : \rho(\theta) = r\} \quad \text{against} \quad H_1 = \{\theta : \rho(\theta) \neq r\},$$

and see for which values of r we cannot reject H_0 . Of course we would have to choose a test function ϕ_r for each test, for example the likelihood ratio test function. This leads to the following definition:

CONFIDENCE REGION FOR $\rho(\theta)$. A $(1 - \alpha)$ -confidence region for $\rho(\theta)$ is constructed by choosing a U uniform in $[0, 1]$ and defining the random set

$$C(X) = \{r \in \mathbb{R} : \phi_r(X) \leq U\}.$$

This is the set of all values $r \in \mathbb{R}$ for which the hypothesis

$$H_0 = \{\theta \in \Theta : \rho(\theta) = r\} \quad \text{against} \quad H_1 = \{\theta \in \Theta : \rho(\theta) \neq r\},$$

cannot be rejected at significance level α , if we use the test function ϕ_r and the decision variable U .

As a first example, we again consider the normal distribution and define $\rho(\mu, \sigma) = \mu$. Then we get

$$H_0 = \{(r, \sigma) : \sigma > 0\} \quad \text{and} \quad H_1 = \{(\mu, \sigma) : \mu \neq r, \sigma > 0\}.$$

The calculations at the end of Section 1.6 show that in this case, the likelihood ratio test is based on

$$T(x) = \sqrt{n} \frac{\bar{x}_n - r}{s_n}.$$

Since $T(X)$ is $t(n-1)$ distributed for all values of (μ, σ) , we will not reject H_0 if

$$-t_{n-1, \alpha/2} \leq \sqrt{n} \frac{\bar{x}_n - r}{s_n} \leq t_{n-1, \alpha/2}.$$

This leads to the following well-known confidence interval for μ :

$$[\bar{x}_n - t_{n-1, \alpha/2} s_n / \sqrt{n}, \bar{x}_n + t_{n-1, \alpha/2} s_n / \sqrt{n}].$$

As in the case for confidence regions for θ , we have the following lemma:

Lemma 1.4 (Covering probability) *If $C(X)$ is a $(1 - \alpha)$ -confidence region for $\rho(\theta)$, then*

$$P_\theta(\rho(\theta) \in C(X)) \geq 1 - \alpha \quad \forall \theta \in \Theta.$$

This says that a $(1 - \alpha)$ -confidence region has a probability of at least $1 - \alpha$ to cover (or contain) the “true” underlying value $\rho(\theta)$.

Proof of Lemma 1.4: Since $\rho(\theta) \in C(X)$ precisely when $\phi_{\rho(\theta)}(X) \leq U$, we get

$$\begin{aligned} P_\theta(\rho(\theta) \in C(X)) &= P_\theta(\phi_{\rho(\theta)}(X) \leq U) \\ &= 1 - P_\theta(\phi_{\rho(\theta)}(X) > U) \\ &= 1 - \int_{\mathcal{X}} \int_0^1 1_{\{\phi_{\rho(\theta)}(x) > u\}} du f_\theta(x) dx \\ &= 1 - \int_{\mathcal{X}} \phi_{\rho(\theta)}(x) f_\theta(x) dx \\ &= 1 - E_\theta[\phi_{\rho(\theta)}(X)]. \end{aligned}$$

This, together with the remark that for our test function $\phi_{\rho(\theta)}$ we must have

$$\sup_{\theta': \rho(\theta') = \rho(\theta)} E_{\theta'} [\phi_{\rho(\theta)}(X)] \leq \alpha,$$

proves the lemma. $\square$

Since ρ is real-valued, we are sometimes interested in finding an upper confidence limit for $\rho(\theta)$. This means that we want to find $r_u \in \mathbb{R}$ such that we are more than $1 - \alpha$ sure that $\rho(\theta) \leq r_u$. It turns out that the following definition works:

UPPER CONFIDENCE REGION FOR $\rho(\theta)$. Consider the hypotheses

$$H_0(r) = \{\theta \in \Theta : \rho(\theta) > r\} \quad \text{against} \quad H_1(r) = \{\theta \in \Theta : \rho(\theta) \leq r\}.$$

Choose a family of test functions ϕ_r at significance level α such that

$$r_1 \geq r_2 \Rightarrow \phi_{r_1}(X) \geq \phi_{r_2}(X) \text{ with probability } 1.$$

A $(1 - \alpha)$ -upper confidence region for $\rho(\theta)$ is constructed by choosing U uniform in $[0, 1]$ and defining the random set

$$C_u(X) = \{r \in \mathbb{R} : \phi_r(X) \leq U\}.$$

This is the set of all values $r \in \mathbb{R}$ for which the hypothesis $H_0(r)$ cannot be rejected at significance level α , if we use the test function ϕ_r and the decision variable U .

This definition requires an additional remark: if $r_1 \geq r_2$, then $H_0(r_1) \subset H_0(r_2)$. This shows that it is always possible to find a family ϕ_r that satisfies the condition in the definition, for example by using the likelihood ratio test functions. It is also clear that if $r \in C_u(x)$, then $(-\infty, r] \subset C_u(x)$, so $C_u(x)$ will always be an interval of the form $(-\infty, r_u)$ or $(-\infty, r_u]$, where $r_u = +\infty$ is allowed. The term upper confidence interval is therefore justified, if we have the following lemma:

Lemma 1.5 (Covering probability) *If $C_u(X)$ is a $(1 - \alpha)$ -upper confidence region for $\rho(\theta)$, then*

$$P_\theta(\rho(\theta) \in C_u(X)) \geq 1 - \alpha \quad \forall \theta \in \Theta.$$

This says that a $(1 - \alpha)$ -upper confidence region has a probability of at least $1 - \alpha$ to cover (or contain) the “true” underlying value $\rho(\theta)$. The proof of this lemma is identical (mutatis mutandis) to the proof of Lemma 1.4.

To be complete, we will also define a lower confidence region:

LOWER CONFIDENCE REGION FOR $\rho(\theta)$. Consider the hypotheses

$$H_0(r) = \{\theta \in \Theta : \rho(\theta) < r\} \quad \text{against} \quad H_1(r) = \{\theta \in \Theta : \rho(\theta) \geq r\}.$$

Choose a family of test functions ϕ_r at significance level α such that

$$r_1 \geq r_2 \Rightarrow \phi_{r_1}(X) \leq \phi_{r_2}(X) \text{ with probability } 1.$$

A $(1 - \alpha)$ -lower confidence region for $\rho(\theta)$ is constructed by choosing U uniform in $[0, 1]$ and defining the random set

$$C_l(X) = \{r \in \mathbb{R} : \phi_r(X) \leq U\}.$$

This is the set of all values $r \in \mathbb{R}$ for which the hypothesis $H_0(r)$ cannot be rejected at significance level α , if we use the test function ϕ_r and the decision variable U .

Here we find that $C_l(x)$ is of the form $[r_l, +\infty)$ or $(r_l, +\infty)$, and the analogue of Lemma 1.5 is also true.

Let us revisit the example where John and Mary had $n = 20$ observations $x_1, x_2, \dots, x_n$, which they modelled as an i.i.d. sample from an $Exp(\theta)$ distribution. Now we want to find a $(1 - \alpha)$ -upper confidence interval for $\rho(\theta) = \theta$ (here we already have that θ is a real-valued parameter). This means that we have to consider testing

$$H_0 : \theta > r \quad \text{against} \quad H_1 : \theta \leq r.$$

The likelihood of the data x is given by

$$f_{\theta}^{(n)}(x) = \theta^n e^{-\theta n \bar{x}_n}.$$

Suppose $r \geq 1/\bar{x}_n$ (remember that $1/\bar{x}_n$ is the maximum likelihood estimator of θ). Since $\theta \mapsto \theta^n e^{-\theta n \bar{x}_n}$ is a strictly decreasing function for $\theta > 1/\bar{x}_n$ (you might want to check this), we get for our likelihood ratio statistic $L_r(x)$:

$$L_r(x) = \frac{\bar{x}_n^{-n} e^{-n}}{r^n e^{-rn \bar{x}_n}} = e^{n(r \bar{x}_n - \log(r \bar{x}_n))} e^{-n}.$$

If $r < 1/\bar{x}_n$, we get

$$L_r(x) = e^{-n(r \bar{x}_n - \log(r \bar{x}_n))} e^n.$$

A picture of this function can be seen in Figure 1.8.

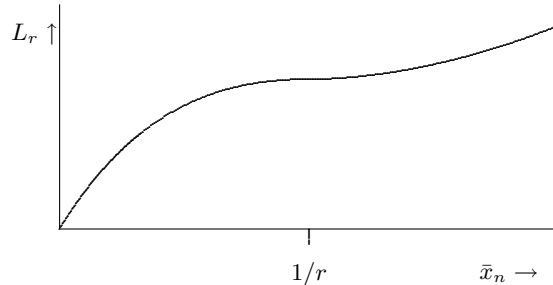


Figure 1.8: The function $L_r(x)$.

The P-value p_r of the likelihood ratio test is given by

$$p_r = \sup_{\theta > r} P_{\theta}(L_r(X) > L_r(x)) = \sup_{\theta > r} P_{\theta}(\bar{X}_n > \bar{x}_n)$$

The last equality follows from the fact that $L_r(X)$ is a strictly increasing function of $\bar{X}_n$. Now note that if $\theta > r$, then

$$P_{\theta}(\bar{X}_n > t) < P_r(\bar{X}_n > t) \quad (\forall t > 0).$$

Therefore

$$p_r = P_r(\bar{X}_n > \bar{x}_n).$$

We reject H_0 if $p_r < \alpha$, so we need to find r_u such that

$$P_{r_u}(\bar{X}_n > \bar{x}_n) = \alpha.$$

For this r_u we have that $C_u(x) = (0, r_u]$. Define $S_n = X_1 + \dots + X_n$. Then S_n is $Gam(n, \theta)$ distributed, which means that

$$P_r(S_n > s_n) = \frac{1}{(n-1)!} \int_{s_n}^{\infty} r^n t^{n-1} e^{-rt} dt = \frac{1}{(n-1)!} \int_{rs_n}^{\infty} t^{n-1} e^{-t} dt.$$

We use the computer to solve the equation

$$\frac{1}{(n-1)!} \int_a^{\infty} t^{n-1} e^{-t} dt = \alpha$$

with $n = 20$ and $\alpha = 0.05$. This gives us $a = 27.87924$. Therefore our 95%-upper confidence limit is given by

$$r_u = 27.87924/s_n = 1.393962/\bar{x}_n.$$

1.9 Solutions to the quick exercises

1.1 Since a test function ϕ can only take the values 0 or 1, we get the

$$E[\phi(X)] = 0 \cdot P(\phi(X) = 0) + 1 \cdot P(\phi(X) = 1) = P(\phi(X) = 1).$$

However, the event $\{\phi(X) = 1\}$ is by definition equal to the event $\{\text{reject } H_0\}$.

1.2 Look at the probability mass function of X :

$$P_p(X = k) = \frac{20!}{k!(20-k)!} p^k (1-p)^{20-k}.$$

As a function of p , we can see by taking the derivative that this function has only one maximum, let's call it p_k , that the function is increasing up to p_k and decreasing after p_k . Clearly, if $k \geq k'$, then $p_k \geq p_{k'}$. Furthermore, again by looking at the derivative, we easily see that $p_3 > 0.06$. Therefore, for any $p \leq 0.06$ and for any $k \geq 3$, we have

$$P_p(X = k) \leq P_{0.06}(X = k).$$

Since this implies that $P_p(X \geq 4) \leq P_{0.06}(X \geq 4)$ and that $P_p(X = 3) \leq P_{0.06}(X = 3)$, we conclude that

$$P_p(\text{rejecting } H_0) \leq P_{0.06}(\text{rejecting } H_0) \quad \forall p \in H_0.$$

1.3 A possibility using simulation would be to draw $U_1, \dots, U_{20}$ from a $U(0, 1)$ distribution and define

$$X_i = -\ln(U_i).$$

Check that $X_i \sim \text{Exp}(1)$. Then calculate T_1 and T_2 . Repeat this 10000 times, so you have 10000 realizations of T_1 and T_2 . Order these realizations and call the ordered realizations $T_1(1) \leq T_1(2) \leq \dots \leq T_1(10000)$, and the same for T_2 . Now define $c_1 = T_1(9500)$ and $c_2 = T_2(9500)$.

This approach however is not even accurate for the second decimal. Another option is to use a mathematical package to calculate the exact distribution of T_1 and T_2 . This gives $c_1 = 0.4325035866$ and $c_2 = 0.4334474220$. You might want to try this!

1.4 We have to calculate $E_{0.09}[\phi(X)]$. Since $\phi(k) = 1$ if $k \geq 4$, we get a contribution

$$\begin{aligned} P_{0.09}(X \geq 4) &= 1 - (0.91)^{20} - 20(0.09)(0.91)^{19} \\ &\quad - 190(0.09)^2(0.91)^{18} - \binom{20}{3}(0.09)^3(0.91)^{17} \\ &= 0.0993. \end{aligned}$$

However, there is also a probability of rejecting when $X = 3$. This gives a contribution

$$0.2446 \cdot P_{0.09}(X = 3) = 0.0409.$$

Together this gives a power of 0.1402.

1.5 It is clear that the function

$$f(t) = \begin{cases} -\log(1 + t^2) & t \geq 0 \\ \log(1 + t^2) & t \leq 0 \end{cases}$$

is strictly decreasing in t . Now we only have to remark that $s_1^2 = \frac{n}{n-1}s_n^2$.

1.6 We know that $X_i \sim \text{Pois}(\theta)$, so

$$X_1 + \dots + X_n \sim \text{Pois}(n\theta).$$

This follows from properties of the Poisson process. Define $S_n = X_1 + \dots + X_n$. Then, for $a \geq -\sqrt{n\theta}$,

$$\begin{aligned} P_\theta\left(\frac{\bar{X}_n - \theta}{\sqrt{\theta/n}} \leq a\right) &= P_\theta(S_n \leq n\theta + a\sqrt{n\theta}) \\ &= \sum_{i=0}^{\lfloor n\theta + a\sqrt{n\theta} \rfloor} \frac{(n\theta)^i}{i!} e^{-n\theta}. \end{aligned}$$

This means that

$$P_\theta(\theta \in C_1(X)) = \sum_{i=0}^{\lfloor n\theta + z_{\alpha/2}\sqrt{n\theta} \rfloor} \frac{(n\theta)^i}{i!} e^{-n\theta} - 1_{[-\sqrt{n\theta}, \infty)}(-z_{\alpha/2}) \sum_{i=0}^{\lfloor n\theta - z_{\alpha/2}\sqrt{n\theta} \rfloor} \frac{(n\theta)^i}{i!} e^{-n\theta}.$$

1.10 Exercises

1.1 According to an American statistical research company, we can model the income of a male republican (someone who votes for a republican candidate) as $N(30, 5^2)$ distributed, where we measure income in k\$ (i.e. 1000\$). They also claim that we can model the income of a male democrat as $N(27, 4^2)$.

Suppose we know a man that earns 25000\$. Test the hypothesis that this man is a republican against the hypothesis that he is a democrat at a significance level of 20%, using the Neyman-Pearson test, and calculate the power of the test if he is a democrat. Hint: use the log of the Neyman-Pearson test statistic and show that it is a strictly increasing function of $(x - 21\frac{2}{3})^2$. To find the critical value, neglect the appropriate tail probability.

1.2 John asks 7 of his fellow students if they smoke or not. It turns out that 3 smoke. Suppose the probability that a fellow student of John smokes is p . Given is the following table of the $\text{Bin}(7, 0.1)$ distribution.

Table 1.3: The $\text{Bin}(7, 0.1)$ distribution.

k	$P(X \leq k)$	k	$P(X \leq k)$
0	0.4783	4	0.9998
1	0.8503	5	1.0000
2	0.9743	6	1.0000
3	0.9973	7	1.0000

- Give the likelihood ratio test function to test the hypothesis $H_0 : p = 0.1$ against $H_1 : p > 0.1$, using a significance of 95%. Hint: show that the likelihood ratio statistic is a strictly increasing function of X .
- Show that this test is the uniformly most powerful test for this testing problem.
- Can John reject H_0 ? What is the P-value for this test? Take as a realization for the decision variable $u = 0.3$.

1.3 We model our data $x = (x_1, \dots, x_n)$ as realizations of a random sample $X_1, \dots, X_n \sim \text{Geo}(p)$, with $p \in (0, 1)$. This means that

$$P(X_i = k) = p(1 - p)^{k-1} \quad k = 1, 2, \dots$$

We want to test the hypothesis

$$H_0 : p \geq p_0 \quad \text{against} \quad H_1 : p < p_0.$$

- Define L as the likelihood ratio test statistic. Show that

$$\begin{aligned} \log L(x) = & 1_{\{\bar{x}_n \leq \frac{1}{p_0}\}} [\log(p_0) + \bar{x}_n \log(\bar{x}_n) + (\bar{x}_n - 1) \log(1 - p_0) - (\bar{x}_n - 1) \log(\bar{x}_n - 1)] + \\ & + 1_{\{\bar{x}_n > \frac{1}{p_0}\}} [-\log(p_0) - \bar{x}_n \log(\bar{x}_n) - (\bar{x}_n - 1) \log(1 - p_0) + (\bar{x}_n - 1) \log(\bar{x}_n - 1)]. \end{aligned}$$

- b. Show that L , as a function of $\bar{x}_n$, is strictly increasing.
- c. Suppose $Z_1, \dots, Z_n \sim \text{Exp}(1)$. Show that for all $p \in (0, 1)$

$$X_i^{(p)} := \lfloor \frac{-Z_i}{\log(1-p)} \rfloor + 1 \sim \text{Geo}(p).$$

Here $\lfloor a \rfloor$ is the largest integer k such that $k \leq a$. Hint: consider $P(X_i^{(p)} > k)$ for $k \geq 1$.

- d. Define $X^{(p)} = (X_1^{(p)}, \dots, X_n^{(p)})$ and show that if $p_1 \geq p_2$,

$$\bar{X}_n^{(p_1)} \leq \bar{X}_n^{(p_2)}.$$

- e. Show that for all $C > 0$,

$$\sup_{p \geq p_0} P_p(\bar{X}_n > C) = P_{p_0}(\bar{X}_n > C).$$

Hint: note that $P_p(\bar{X}_n > C) = P(\bar{X}_n^{(p)} > C)$.

- f. Explain how you would use simulations to determine the likelihood ratio test function.
- g. Show that the likelihood ratio test function is the uniformly most powerful test function for our test, by comparing it to the Neyman-Pearson test function for the test

$$H_0 : p = p_0 \quad \text{against} \quad H_1 : p = p_1,$$

where $p_1 < p_0$.

1.4 Suppose we have data on lifetimes of televisions. We have data for $n = 150$ televisions and their average lifetime $\bar{x}_n = 4.8$ years. We model these lifetimes as $\text{Exp}(\lambda)$ distributed.

- a. Test the hypothesis $H_0 : \lambda \leq 0.2$ against $H_1 : \lambda > 0.2$, using the likelihood ratio test, at a significance of 5%. You may use that the equation

$$\frac{1}{(n-1)!} \int_0^a s^{n-1} e^{-s} ds = 0.05$$

has as a solution $a = 130.439$ (for $n = 150$). Also note that if $X_1, \dots, X_n \sim \text{Exp}(\lambda)$, then, if $S_n = X_1 + \dots + X_n$,

$$S_n \sim \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}.$$

- b. Show that the above test is the uniformly most powerful test, by comparing it to the Neyman-Pearson test for $H_0 : \lambda = 0.2$ against $H_1 : \lambda = \lambda_1$, where $\lambda_1 > 0.2$.

- c. Construct a 95% lower confidence interval for λ using the likelihood ratio test. Compare this to the 95% lower confidence interval you get if we approximate the distribution of $\bar{X}_n$ with an appropriate normal distribution.

1.5 Let $A, B, H_0 \subset \Theta$ such that $A \subset B$ and $H_0 \cap B = \emptyset$, where Θ describes a statistical model for a random variable X . Consider two testing problems: test I

$$H_0 \quad \text{against} \quad H_1 = A$$

and test II

$$H_0 \quad \text{against} \quad H_1 = B.$$

Fix a significance level α .

- a. Show that for any testfunction ϕ for test II, there exists a testfunction ϕ' for test I such that

$$E_\theta[\phi'(X)] \geq E_\theta[\phi(X)] \quad \forall \theta \in A.$$

This says that if the alternative is smaller, you can get greater power.

- b. Let $\mathcal{X} = \{1, 2, 3\}$ and $\Theta = \{0, 1, 2, 3\}$, where the 4 probability mass functions are described below:

x	1	2	3
$f_0(x)$	1/3	1/3	1/3
$f_1(x)$	2/3	0	1/3
$f_2(x)$	5/12	0	7/12
$f_3(x)$	0	1/3	2/3

Furthermore we define $H_0 = \{0\}$, $A = \{1, 2\}$ and $B = \{1, 2, 3\}$. Choose $\alpha = 1/3$. Show that there exists $\theta \in A$ such that the likelihood ratio testfunction for test I has *lower* power at θ than the likelihood ratio testfunction for test II.

- c. Show that there does not exist a testfunction ϕ for test I in the situation of **b.** that has *strictly* better power than the likelihood ratio testfunction for test II at all $\theta \in A$.

Chapter 2

Linear Models

In this chapter we will introduce general linear models. We will use matrix notation, calculate the least squares estimator and derive likelihood ratio tests under the assumption of normality. We will also use the bootstrap to perform tests without normality assumptions.

2.1 Introduction

Consider the following data set, regarding the abrasion of rubber and how this relates to the hardness and the tensile strength of the rubber. The data are taken from [6] and are originally from [2] page 239.

Table 2.1: Abrasion loss of rubber.

Abrasion loss	Hardness	Tens. Str.	Abrasion loss	Hardness	Tens. Str.
372	45	162	196	68	173
206	55	233	128	75	188
175	61	232	97	83	161
154	66	231	64	88	119
136	71	231	249	59	161
112	71	237	219	71	151
55	81	224	186	80	165
45	86	219	155	82	151
221	53	203	114	89	128
166	60	189	341	51	161
164	64	210	340	59	146
113	68	210	283	65	148
82	79	196	267	74	144
32	81	180	215	81	134
228	56	200	148	86	127

Each of the $n = 30$ measurements contains three values, which makes it difficult to make an insightful plot of the data, that might suggest some meaningful

relationship between the abrasion loss on the one hand and the hardness and tensile strength on the other. With three measurements, this might still be possible, but with even higher dimensional data, we would have to think of some other way to analyze the data.

Quick exercise 2.1 Think of an informative plot (or plots) you could make of the Abrasion loss data.

Let us call the measured abrasion loss of the i^{th} measured sample of rubber y_i . Furthermore, we call the hardness of the i^{th} rubber sample x_{1i} and the tensile strength of the i^{th} rubber sample x_{2i} . We would like to explain or predict the abrasion loss y , when we know the hardness x_1 and the tensile strength x_2 of some rubber sample. The variable y is called the *dependent variable* and the variables x_1 and x_2 are called the *regressors* or *independent variables*.

We will use the following scheme to predict y , given the regressors x_1 and x_2 : we model y as a realization of a random variable Y , whose distribution depends on x_1 and x_2 in the following way:

$$Y = g(x_1, x_2; \theta) + U \quad (2.1)$$

where $g(\cdot; \theta)$ is called the *regression function* depending on some unknown parameter $\theta \in \Theta$ and U is a random variable with a distribution that does not depend on x_1 or x_2 , such that

$$\mathbb{E}[U] = 0.$$

Equation (2.1) is known as a general *regression model*. The vector $(y_1, \dots, y_n)$ is modelled as a realization of the random variables $(Y_1, \dots, Y_n)$, where we suppose that all Y_i are independent. In fact,

$$Y_i = g(x_{1i}, x_{2i}; \theta) + U_i$$

and we suppose that all U_i are independent and identically distributed (with expectation 0). The statistical problem will be to estimate the unknown θ and, if relevant, the unknown distribution of the U_i 's. We treat the regressors $x_{11}, \dots, x_{1n}$ and $x_{21}, \dots, x_{2n}$ as known constants.

There are different ways to interpret model 2.1. One way is to think of $g(x_1, x_2; \theta)$ as the “true” value of the independent variable, that is perturbed by a measurement error U (so the outcome y is the true value plus some measurement error). In the case of the abrasion loss data, it seems more appropriate to think of $g(x_1, x_2; \theta)$ as the average abrasion loss of a rubber sample, given a hardness x_1 and a tensile strength x_2 . The actual abrasion loss also depends on other factors beyond our control, and this is modelled by the random effect U .

We would like to point out that the most restrictive assumption in the general regression model is that the U_i 's are identically distributed, that is, their distribution do not depend on the regressors x_1 and x_2 . For example, if Y is the outcome of a measurement of some distance x (so $Y = x + U$ where U is the measurement error), then it would seem natural to assume that the variance of U is bigger, as x gets bigger (one can measure the distance between two ends

of a table more accurately than the distance to the moon). The fact that the variance of U would depend on the regressors is called the *heteroscedasticity* of the data. The assumption that the variance does not depend on the regressors is called the *homoscedasticity* of the data. We will point out ways to check the assumed homoscedasticity of the data.

2.2 Linear models

In the previous section we introduced a general regression model, with some regression function $g(\cdot; \theta)$. However, we will concentrate on a very specific kind of regression function, namely one that is linear in the unknown parameter θ : in this case, we usually denote the unknown parameter by β and assume that $\beta \in \mathbb{R}^p$. The model then becomes:

$$Y_i = g_1(x_{1i}, x_{2i})\beta_1 + \dots + g_p(x_{1i}, x_{2i})\beta_p + U_i.$$

This is a general *linear model* in case we have two regressors. As a linear model for the abrasion loss, one might think of

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + U_i. \quad (2.2)$$

The parameter β_0 is called the *intercept* and usually has index 0 (so in this case, we have that $p = 3$). Note that

$$Y_i = \beta_0 + \beta_1 x_{1i}^2 + \beta_2 e^{-x_{2i}} + \beta_3 x_{1i} x_{2i} + U_i \quad (2.3)$$

is also an example of a linear model, since the model is linear in the parameter $\beta = (\beta_0, \dots, \beta_3)$.

We can write model (2.2) as follows:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}. \quad (2.4)$$

The matrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$$

is called the *regression-matrix*, or *X-matrix* or *design-matrix*. It only depends on the regressors (and on the chosen model, of course) and in fact, it is the only part of the model (2.4) that depends on the regressors (note that we have assumed that the distribution of the U_i 's does not depend on the regressors). We also define the vectors

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}.$$

Using these matrix notations, the most general linear model with parameter $\beta \in \mathbb{R}^p$ is:

$$Y = X\beta + U. \quad (2.5)$$

Here, the $n \times p$ regression-matrix X depends on the regressors, and is assumed to be known and constant.

Quick exercise 2.2 Write down the regression-matrix X for the model (2.3).

2.3 Least squares estimator

We have written down a general form of our linear model:

$$Y = X\beta + U.$$

Here, Y and U are n dimensional (random) vectors, X is an $n \times p$ regression-matrix of known constants and β is a p -dimensional parameter vector. Based on our regression matrix X and $y = (y_1, \dots, y_n)^t$ (remember that we model the vector y as a realization of the random vector Y ; the superscript t indicates the transposition of a matrix or vector), how do we estimate β ? We want to choose that β , that somehow fits the data best. To make this more precise, we define for each $\beta \in \mathbb{R}^p$ the *residuals*

$$R_i(\beta) = Y_i - (X\beta)_i.$$

The realization of R_i will also be called the residual, so the residual for case i , $r_i(\beta)$, equals the difference between the actual measured value y_i and its estimation or prediction $(X\beta)_i$ given the parameter vector β . We would like the residuals to be as small as possible. For this reason we define the *residual sum of squares*:

$$\text{SS}_{\text{res}}(\beta) = \sum_{i=1}^n R_i(\beta)^2 = \sum_{i=1}^n (Y_i - (X\beta)_i)^2 = \|Y - X\beta\|^2. \quad (2.6)$$

The *least squares estimator* $\hat{\beta}$ is defined as that β that minimizes $\text{SS}_{\text{res}}(\beta)$, the residual sum of squares (2.6).

To see how we can calculate the least squares estimate $\hat{\beta}$, we use the last equality in Equation (2.6): apparently, $\text{SS}_{\text{res}}(\hat{\beta})$ equals the squared distance between y and the linear space $R = \{X\beta : \beta \in \mathbb{R}^p\}$; see also Figure 2.1.

This means that $X\hat{\beta}$ is the orthogonal projection of Y onto R . Therefore, the vector $Y - X\hat{\beta}$ is orthogonal to every vector in R , which means that

$$\beta^t X^t (Y - X\hat{\beta}) = 0 \quad \forall \beta \in \mathbb{R}^p.$$

This implies the *normal equations* for $\hat{\beta}$:

$$X^t X \hat{\beta} = X^t Y. \quad (2.7)$$

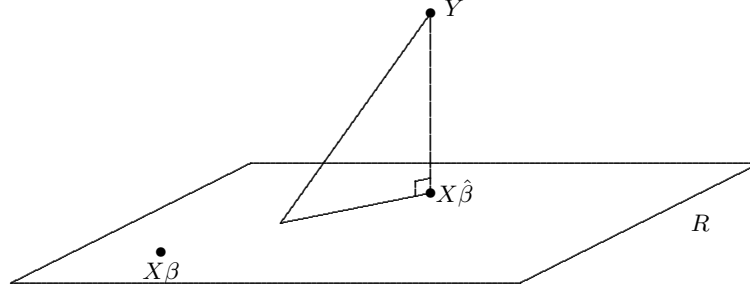


Figure 2.1: Picture of our linear model.

We will now make the assumption that our regression-matrix X is of full rank and that $p \leq n$. It is not hard to show that this assumption is equivalent to the fact that $X^t X$ is an invertible matrix, which leads us to the following equation:

$$\hat{\beta} = (X^t X)^{-1} X^t Y. \quad (2.8)$$

We can now use Equation (2.8) to calculate the least squares estimate when we use model (2.4) for the abrasion loss data:

$$X^t X = \begin{pmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{pmatrix} = \begin{pmatrix} 30 & 2108 & 5414 \\ 2108 & 152422 & 376562 \\ 5414 & 376562 & 1015780 \end{pmatrix},$$

so

$$(X^t X)^{-1} = \begin{pmatrix} 2.8639 & -2.2545 \cdot 10^{-2} & -6.9069 \cdot 10^{-3} \\ -2.2545 \cdot 10^{-2} & 2.5544 \cdot 10^{-4} & 2.5467 \cdot 10^{-5} \\ -6.9069 \cdot 10^{-3} & 2.5467 \cdot 10^{-5} & 2.8357 \cdot 10^{-5} \end{pmatrix}.$$

Furthermore,

$$X^t y = \begin{pmatrix} 5263 \\ 346867 \\ 921939 \end{pmatrix},$$

so

$$\hat{\beta} = (X^t X)^{-1} X^t y = \begin{pmatrix} 885.1611 \\ -6.5708 \\ -1.3743 \end{pmatrix}.$$

To see whether our model is reasonable, we could look at the residuals. Figure 2.2 is a scatter plot of the residuals against the fitted values, that is a scatter plot of $((X\hat{\beta})_i, r_i)$ ($r_i = r_i(\hat{\beta})$):

There does not seem to be a clear relationship with the distribution of the residuals and the regressors, which is in accordance with our assumption of homoscedasticity. Also the size of the residuals is relevant: small residuals suggest a better fit of the model. However, we have to be careful: if we would extend our model with a lot of parameters, we would be able to get a very small residual sum of squares, but we would have adapted to data so much, that the predictive value of our model will be very low. In other words, if we

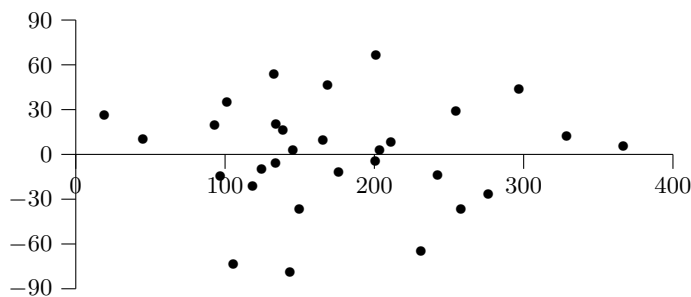


Figure 2.2: Scatter plot of residuals versus fitted values.

would measure another 30 samples of rubber and again fit this large number of parameters, we would find completely different values for the parameters. We should always try to minimize the number of parameters.

In this example we only have three parameters, which does not seem excessive. We look at the residual sum of squares:

$$SS_{\text{res}} = SS_{\text{res}}(\hat{\beta}) = \|y - X\hat{\beta}\|^2 = 35950.$$

To determine whether this number is big or small, we compare it to the residual sum of squares belonging to the model

$$Y_i = \beta_0 + U_i.$$

In this model, that does not depend on the regressors, we easily see that $\hat{\beta}_0 = \bar{Y}_n$. This means that the residual sum of squares of this model, which is called the *corrected total sum of squares*, equals

$$SS_{\text{corr.tot.}} = \sum_{i=1}^n (y_i - \bar{y}_n)^2 = 225010.$$

This $SS_{\text{corr.tot.}}$ will always be bigger than SS_{res} if the original model contains an intercept (can you see why?), but how much bigger gives us some information about how well our model explains the variation in the measured y_i 's. Therefore we define the *Multiple R-squared*, or just *R-squared*, by

$$R^2 = \frac{SS_{\text{corr.tot.}} - SS_{\text{res}}}{SS_{\text{corr.tot.}}} = 0.84.$$

This will always be a number between 0 and 1, and closer to 1 means a better fit of the model. In models where U_i is seen as a (small) measurement error, we would like to see $R^2 > 0.9$. If we think of the U_i as a random deviation from some expected value, $R^2 > 0.8$ is considered reasonable, but much lower values of R^2 are not uncommon. A low value of R^2 does not always mean that the model is wrong: it might just be that the chosen regressors are not able to predict the independent variable very well.

In models without an intercept, we introduce the *total sum of squares* as

$$SS_{\text{total}} = \sum_{i=1}^n y_i^2$$

and define

$$R^2 = \frac{SS_{\text{total}} - SS_{\text{res}}}{SS_{\text{total}}}.$$

This R-squared cannot be directly compared to the one we get if we do use an intercept.

Quick exercise 2.3 Show that for the standard linear model

$$Y_i = \beta_0 + \beta_1 x_i + U_i,$$

Equation (2.8) indeed gives the well known least squares estimator.

2.4 Random vectors

To study the statistical properties of the least squares estimator, we need to extend our notion of random variables to random vectors. A random vector

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

has as its components random variables Y_i , which may depend on each other. In fact, we can describe the distribution of the random vector Y by describing the common distribution of the random variables $(Y_1, \dots, Y_n)$. For a continuous random vector, the case which we will be interested in, this means that there exists a density f on $\mathbb{R}^n$, such that for any appropriate set $B \subset \mathbb{R}^n$

$$P(Y \in B) = \int_B f(y_1, \dots, y_n) dx_1 \dots dx_n.$$

It would go beyond the scope of this book to define precisely what we mean by appropriate set (we would need the notion of a measurable set, see for example [1]), but let us just say that any set we will consider in this book will be appropriate.

Just as for random variables, we can define the expectation of a random vector:

$$E[Y] = \begin{pmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{pmatrix}.$$

We easily check the following rule: let $Y \in \mathbb{R}^{n_1}$ and $Z \in \mathbb{R}^{n_2}$ be two random vectors and let A be a constant $m \times n_1$ matrix and B a constant $m \times n_2$ matrix. Then $AY + BZ$ is a random vector in $\mathbb{R}^m$ and

$$E[AY + BZ] = AE[Y] + BE[Z]. \quad (2.9)$$

Quick exercise 2.4 Prove Rule (2.9) by checking it component wise.

We would also like to extend the notion of variance to random vectors. Of course each component of the random vector Y has its own variance, but different components may be correlated to one another. That is why we make the following definition:

COVARIANCE MATRIX OF A RANDOM VECTOR. Suppose we have a random vector $Y \in \mathbb{R}^n$. We define $C(Y)$, the *covariance matrix* of Y , as the following $n \times n$ matrix:

$$C(Y) = E[(Y - E[Y])(Y - E[Y])^t]$$

$$= \begin{pmatrix} \text{Var}(Y_1) & \text{Cov}(Y_1, Y_2) & \dots & \text{Cov}(Y_1, Y_n) \\ \text{Cov}(Y_1, Y_2) & \text{Var}(Y_2) & \dots & \text{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_1, Y_n) & \text{Cov}(Y_2, Y_n) & \dots & \text{Var}(Y_n) \end{pmatrix}.$$

In this definition we have used the convention that the expectation of a random matrix is also calculated component wise.

From the definition it is clear the one can see the covariance structure of Y , that is, the correlations between the components of Y , from the covariance matrix $C(Y)$. Note that if Y is one dimensional (so Y is just a random variable), then $C(Y) = \text{Var}(Y)$. We have the following lemma:

Lemma 2.1 *Let $Y \in \mathbb{R}^n$ be a random vector, let A be a constant $m \times n$ matrix and let $c \in \mathbb{R}^m$ be a constant vector. Then*

$$C(AY + c) = AC(Y)A^t.$$

Proof: We use the linearity of the expectation (Rule (2.9)):

$$\begin{aligned} C(AY + c) &= E[(AY + c - E[AY + c])(AY + c - E[AY + c])^t] \\ &= E[(AY - AE[Y])(AY - AE[Y])^t] \\ &= E[A(Y - E[Y])(Y - E[Y])^t A^t] \\ &= AE[(Y - E[Y])(Y - E[Y])^t] A^t \\ &= AC(Y)A^t. \end{aligned} \quad \square$$

The covariance matrix is a symmetric, positive semi-definite matrix, that is

$$a^t C(Y) a \geq 0 \quad \forall a \in \mathbb{R}^n.$$

This follows immediately from Lemma 2.1: $a^t C(Y) a = \text{Var}(a^t Y) \geq 0$.

The multivariate normal distribution Let $Z_1, \dots, Z_n$ be independent, standard normal random variables. The random vector

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$$

has by definition a *standard normal n -multivariate distribution*, which we denote as

$$Z \sim N_n(0, I).$$

Clearly,

$$E[Z] = 0 \quad \text{and} \quad C(Z) = I,$$

where I denotes the identity matrix.

We now want to define general multivariate normal distributions, with a prescribed expectation $\mu \in \mathbb{R}^n$ and a prescribed covariance matrix Σ . For this, we note that if Σ is symmetric and positive semi-definite, then there exist a matrix B such that $\Sigma = BB^t$; this B is unique up to unitary transformations. We can think of B as a square root of Σ . We use this property in the following definition:

GENERAL MULTIVARIATE NORMAL DISTRIBUTION. Let $Z \sim N_n(0, I)$. Let $\mu \in \mathbb{R}^n$ and Σ be a symmetric, positive semi-definite $n \times n$ matrix. Choose B such that $\Sigma = BB^t$ and define

$$Y = \mu + BZ.$$

Then, by definition, Y has an n -multivariate normal distribution with parameters μ and Σ , which we denote as

$$Y \sim N_n(\mu, \Sigma).$$

The definition suggests that the distribution of Y does not depend on the choice of B , something we can indeed prove. Furthermore, it immediately follows that

$$E[Y] = \mu + BE[Z] = \mu$$

and

$$C(Y) = BC(Z)B^t = BB^t = \Sigma,$$

which shows that indeed μ is the expectation of Y and Σ is its covariance matrix.

The following lemma states that having a multivariate normal distribution is preserved under linear transformations:

Lemma 2.2 Let $Y \sim N_n(\mu, \Sigma)$. Let A be a constant $m \times n$ matrix and $c \in \mathbb{R}^m$. Then

$$AY + c \sim N_m(A\mu + c, A\Sigma A^t).$$

The proof of this lemma is a clever exercise in linear algebra, which we will omit. Of course we can check that $AY + c$ has the correct expectation and covariance matrix.

Quick exercise 2.5 Let

$$Y \sim N_2 \left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix} \right).$$

Show that $Y_1 + Y_2$ is normally distributed and calculate $P(Y_1 + Y_2 > 2)$.

The following corollary is a well known property of multivariate normal distributions:

Corollary 2.3 *Let $Y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^k$ be random vectors such that $(Y^t, Z^t)^t$ is multivariate normally distributed and such that for each $1 \leq i \leq n$ and $1 \leq j \leq k$ we have*

$$\text{Cov}(Y_i, Z_j) = 0.$$

Then Y and Z are independent.

Proof: The condition on Y and Z implies that

$$\Sigma = C \begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} C(Y) & 0 \\ 0 & C(Z) \end{pmatrix}.$$

Choose B_1 and B_2 such that

$$C(Y) = B_1 B_1^t \quad \text{and} \quad C(Z) = B_2 B_2^t.$$

If we define

$$B = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix},$$

then

$$BB^t = \Sigma.$$

Furthermore, let $W \sim N_n(0, I)$ and $V \sim N_k(0, I)$ such that W and V are independent. This means that

$$\begin{pmatrix} W \\ V \end{pmatrix} \sim N_{n+k}(0, I).$$

Finally, define $\mu_1 = E[Y]$ and $\mu_2 = E[Z]$. Then

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + B \begin{pmatrix} W \\ V \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} B_1 W \\ B_2 V \end{pmatrix}.$$

Here, $\stackrel{d}{=}$ denotes “has the same distribution as”. This proves that Y is independent from Z , since Y depends on W and Z on V . $\square$

2.5 Properties of the least squares estimator

Consider again the general linear model

$$Y = X\beta + U.$$

To determine some properties of the least squares estimator $\hat{\beta}$, found in (2.8), we formalize the assumption on the random vector U :

GAUSS-MARKOV CONDITIONS. When we consider a linear model

$$Y = X\beta + U,$$

the Gauss-Markov conditions on the random vector U are

1. $E[U] = 0$.
2. $C(U) = \sigma^2 I$.

Here σ^2 is an unknown constant.

With these conditions we can immediately deduce the following proposition:

Proposition 2.4 *Let $\hat{\beta}$ be the least squares estimator given by*

$$\hat{\beta} = (X^t X)^{-1} X^t Y.$$

Under the Gauss-Markov conditions we have

$$E[\hat{\beta}] = \beta \quad \text{and} \quad C(\hat{\beta}) = \sigma^2 (X^t X)^{-1}.$$

This shows that $\hat{\beta}$ is an unbiased estimator of β .

Proof of Proposition 2.4: Since $Y = X\beta + U$, we get

$$E[Y] = X\beta + E[U] = X\beta \quad \text{and} \quad C(Y) = C(U) = \sigma^2 I.$$

Therefore,

$$\begin{aligned} E[\hat{\beta}] &= E[(X^t X)^{-1} X^t Y] \\ &= (X^t X)^{-1} X^t E[Y] \\ &= (X^t X)^{-1} X^t X \beta \\ &= \beta. \end{aligned}$$

Furthermore,

$$\begin{aligned} C(\hat{\beta}) &= C((X^t X)^{-1} X^t Y) \\ &= (X^t X)^{-1} X^t C(Y) X (X^t X)^{-1} \\ &= \sigma^2 (X^t X)^{-1} X^t X (X^t X)^{-1} \\ &= \sigma^2 (X^t X)^{-1}. \quad \square \end{aligned}$$

It is clear that the covariance matrix $C(\beta)$ contains information about the accuracy of our least squares estimator. However, it still contains the unknown σ^2 . Note that

$$\text{Var}(U_i) = E[U_i^2] = \sigma^2,$$

so the weak law of large numbers (see [3]) tells us that

$$\frac{1}{n} \sum_{i=1}^n U_i^2 \xrightarrow{p} \sigma^2 \quad (n \rightarrow \infty).$$

If we want to estimate σ^2 , it seems like a good idea to look at the average of the U_i^2 's. The problem is of course, that we do not know the U_i . The closest thing would be the squared residuals R_i^2 . This suggests looking at SS_{res} . The following proposition tells us how to find an unbiased estimator of σ^2 , using SS_{res} .

Proposition 2.5 *Consider the linear model*

$$Y = X\beta + U$$

under the Gauss-Markov conditions. Let $\beta \in \mathbb{R}^p$ with $p < n$ and suppose that X is of full rank. Define

$$S^2 = \frac{\text{SS}_{\text{res}}}{n - p}.$$

Then $E[S^2] = \sigma^2$, so S^2 is an unbiased estimator of σ^2 .

Proof: Define as before

$$R = \{X\beta : \beta \in \mathbb{R}^p\} \subset \mathbb{R}^n.$$

Since X is of full rank, we know that $\dim(R) = p$. Choose an orthonormal basis $e_1, \dots, e_p$ of $\mathbb{R}^p$ such that

$$R = \langle e_1, \dots, e_p \rangle.$$

See also Figure 2.3.

Define the $n \times n$ matrix

$$E = (e_1 | e_2 | \dots | e_n),$$

so the i^{th} column of E is e_i . Then E is a unitary matrix, that is, $E^t E = I$. Also define the random vector

$$Z = E^t Y, \quad \text{or} \quad Y = EZ = Z_1 e_1 + \dots + Z_n e_n.$$

Then

$$C(Z) = \sigma^2 E^t I E = \sigma^2 I.$$

Now note that $X\hat{\beta}$ is the orthogonal projection of Y onto R , but since $Y = Z_1 e_1 + \dots + Z_n e_n$, we get

$$X\hat{\beta} = Z_1 e_1 + \dots + Z_p e_p.$$

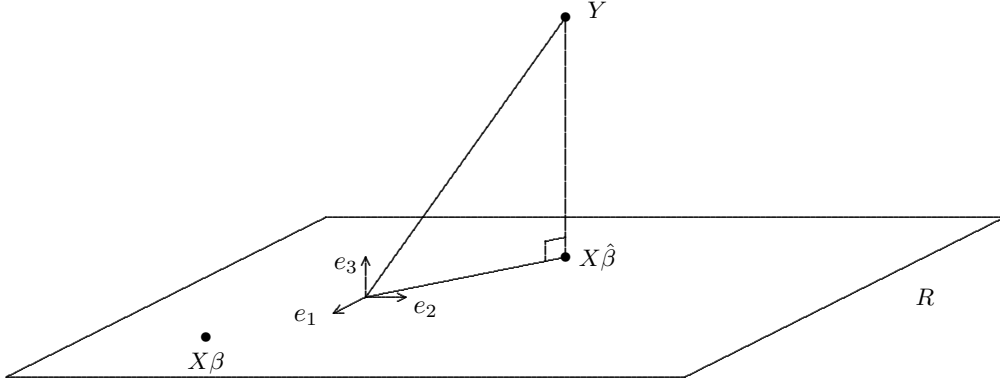


Figure 2.3: Picture of our linear model.

Therefore,

$$Y - X\hat{\beta} = Z_{p+1}e_{p+1} + \dots + Z_n e_n.$$

This means that (Pythagoras)

$$\text{SS}_{\text{res}} = \|Y - X\hat{\beta}\|^2 = Z_{p+1}^2 + \dots + Z_n^2.$$

Now note that if $i \geq p + 1$,

$$\text{E}[Z_i] = (E^t \text{E}[Y])_i = e_i^t X\beta = 0,$$

because $e_i \perp R$ and $X\beta \in R$. Therefore, for $i \geq p + 1$,

$$\text{E}[Z_i^2] = \text{Var}(Z_i) = \sigma^2.$$

This shows that

$$\text{E}[\text{SS}_{\text{res}}] = (n - p)\sigma^2,$$

which in turn proves that $\text{E}[S^2] = \sigma^2$. $\square$

We are now able to estimate the covariance matrix of $\hat{\beta}$ for our abrasion loss model. We had already seen that

$$\text{SS}_{\text{res}} = 35950.$$

Since in our model $p = 3$ and $n = 30$, we get

$$S^2 = \text{SS}_{\text{res}}/27 = 1331.5.$$

An estimate for σ , the standard deviation of the U_i 's, would be

$$\hat{\sigma} = S = 36.5.$$

Furthermore, an estimate for the covariance matrix is given by

$$\hat{C}(\hat{\beta}) = S^2(X^t X)^{-1} = \begin{pmatrix} 3813.3 & -30.018 & -9.1964 \\ -30.018 & 0.3401 & 0.0339 \\ -9.1964 & 0.0339 & 0.0378 \end{pmatrix}.$$

We define the *standard error* of $\hat{\beta}_i$, denoted by $\text{SE}(\hat{\beta}_i)$, as the estimated standard deviation of $\hat{\beta}_i$, so

$$\text{SE}(\hat{\beta}_i) = \sqrt{(\hat{C}(\hat{\beta}))_{ii}} = S\sqrt{(X^t X)^{-1}_{ii}}.$$

For the abrasion loss data, this leads to Table 2.2:

Table 2.2: Estimation of coefficients of abrasion loss data.

	Value	Standard Error
Intercept	885.16	61.752
Hardness	-6.5708	0.5832
Tensile Strength	-1.3743	0.1943

We would like to use this information to make a confidence interval for the coefficients, or to test some hypothesis, but for this we need extra assumptions.

2.6 Normality assumption

In this section we will make an assumption on the distribution of the random vector U in the linear model

$$Y = X\beta + U.$$

This is necessary for making confidence intervals or testing hypotheses.

NORMALITY ASSUMPTION. We consider a linear model

$$Y = X\beta + U,$$

and assume that the random vector U has the following multivariate normal distribution:

$$U \sim N_n(0, \sigma^2 I).$$

Here σ^2 is an unknown constant.

Note that the Gauss-Markov conditions follow from the normality assumption. Since U has a multivariate normal distribution, so has Y , and since $\hat{\beta}$ is a linear transformation of Y , it also has a multivariate normal distribution and we can use Lemma 2.4 to see that

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X^t X)^{-1}). \quad (2.10)$$

Quick exercise 2.6 Give the distribution of Y under the normality assumption.

To be able to use (2.10), we also need to estimate σ^2 , but this will influence the distributions. We can prove that in fact, the random variable SS_{res} (and therefore also the random variable S^2) is independent of the random vector $\hat{\beta}$. Furthermore, we have the following proposition:

Proposition 2.6 Consider the i^{th} component of the least squares estimator $\hat{\beta}$, $\hat{\beta}_i$. We have

$$\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{(X^t X)_{ii}^{-1}}} \sim N(0, 1) \quad (2.11)$$

and

$$\frac{\hat{\beta}_i - \beta_i}{S \sqrt{(X^t X)_{ii}^{-1}}} = \frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)} \sim t(n - p). \quad (2.12)$$

Here, $t(n - p)$ denotes the t -distribution with $n - p$ degrees of freedom.

For an introduction of the t -distribution, see [3].

Quick exercise 2.7 Prove Equation (2.11).

We can use Equation (2.12) to make a 95% confidence interval for β_1 in our abrasion loss data. From Table 2.2 we see that

$$\hat{\beta}_1 = -6.5708 \quad \text{and} \quad \text{SE}(\hat{\beta}_1) = 0.5832.$$

Now we use Table 3.6 to find the critical value of the t -distribution with 27(= $n - p$) degrees of freedom at 2.5%:

$$t_{27, 0.025} = 2.052.$$

So (2.12) implies that

$$\begin{aligned} 95\% &= P\left(-2.052 < \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} < 2.052\right) \\ &= P\left(\hat{\beta}_1 - 2.052 \text{ SE}(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + 2.052 \text{ SE}(\hat{\beta}_1)\right) \end{aligned}$$

This means that a 95% confidence interval for β_1 is given by

$$(-6.7508 - 2.052 \cdot 0.5832, -6.7508 + 2.052 \cdot 0.5832) = (-7.768, -5.374).$$

In general we have

CONFIDENCE INTERVAL FOR β_i . A $(1 - \alpha)$ confidence interval for β_i is given by

$$\left(\hat{\beta}_i - t_{n-p, \alpha/2} \text{ SE}(\hat{\beta}_i), \hat{\beta}_i + t_{n-p, \alpha/2} \text{ SE}(\hat{\beta}_i)\right).$$

2.7 Testing Hypotheses

Within the linear model

$$Y = X\beta + U,$$

with $\beta \in \mathbb{R}^p$ and $\sigma^2 = \text{Var}(U_i) > 0$, we consider the following testing problem:

$$H_0 : \beta \in B_0 \quad \text{against} \quad H_1 : \beta \notin B_0. \quad (2.13)$$

In this hypothesis, we do not restrict the parameter σ^2 . In fact, our complete parameter space is given by

$$\Theta = \mathbb{R}^p \times (0, \infty).$$

Therefore, if we view H_0 as a subset of Θ , we get

$$H_0 = \{(\beta, \sigma^2) : \beta \in B_0, \sigma^2 > 0\}.$$

We will also assume that $H_1 = \Theta \setminus H_0$ is dense in Θ , so B_0 is in a certain sense a very thin subset of $\mathbb{R}^p$. Usually, B_0 will be an affine subspace of $\mathbb{R}^p$ (that is, a translated linear subspace) of dimension strictly lower than p , for example

$$B_0 = \{\beta \in \mathbb{R}^p : \beta_1 = 1 \text{ and } \beta_2 = 0\}.$$

To calculate the likelihood ratio test statistic for test (2.13), we remark that X is viewed as a known constant matrix, and Y has the following distribution under the normality assumption:

$$Y = X\beta + U \sim N_n(X\beta, \sigma^2 I).$$

Note that Corollary 2.3 implies that all U_i are independent and identically distributed (namely, $U_i \sim N(0, \sigma^2)$), which means that all

$$Y_i = (X\beta)_i + U_i \sim N((X\beta)_i, \sigma^2),$$

are also independent. Given a realization $(y_1, \dots, y_n)$, we can write down the likelihood as a function of the parameter $\theta = (\beta, \sigma^2)$:

$$\begin{aligned} f_\theta(y_1, \dots, y_n) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2} \frac{(y_1 - (X\beta)_1)^2}{\sigma^2}} \cdots \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2} \frac{(y_n - (X\beta)_n)^2}{\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} e^{-\frac{1}{2} \frac{\sum_{i=1}^n (y_i - (X\beta)_i)^2}{\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} e^{-\frac{1}{2} \frac{\text{SS}_{\text{res}}(\beta)^2}{\sigma^2}}. \end{aligned}$$

The likelihood ratio test statistic is given by

$$L(y) = \frac{\sup_{\theta \in H_1} f_\theta(y)}{\sup_{\theta \in H_0} f_\theta(y)}. \quad (2.14)$$

Since H_1 is dense in Θ , we see that to calculate the numerator of (2.14), we need to find the maximum likelihood estimator of $\theta = (\beta, \sigma)$. It is clear that

$$f_\theta(y) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} e^{-\frac{1}{2} \frac{\text{SS}_{\text{res}}(\beta)^2}{\sigma^2}}$$

is maximal, whenever $\text{SS}_{\text{res}}(\beta)$ is minimal, which proves that the maximum likelihood estimator of β is equal to the least squares estimator $\hat{\beta}$ (under the normality assumption). We had already defined

$$\text{SS}_{\text{res}} = \text{SS}_{\text{res}}(\hat{\beta}).$$

Now we need to maximize the function

$$\sigma \mapsto \frac{1}{\sigma^n} e^{-\frac{1}{2} \frac{\text{SS}_{\text{res}}^2}{\sigma^2}}.$$

By taking the log and differentiating to σ , we get for the maximum likelihood estimator $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\text{SS}_{\text{res}}}{n}. \quad (2.15)$$

For the denominator of (2.14), we see that we have to minimize $\text{SS}_{\text{res}}(\beta)$ over $\beta \in B_0$. For this we define

$$\text{SS}_{\text{res}, H_0} = \inf_{\beta \in B_0} \text{SS}_{\text{res}}(\beta).$$

This is also known as the residual sum of squares under the null hypothesis. Similarly to (2.15), to maximize the likelihood $f_\theta(y)$ over $\theta \in H_0$, we need to choose for σ^2 :

$$\hat{\sigma}_{H_0}^2 = \frac{\text{SS}_{\text{res}, H_0}}{n}.$$

This means that (2.14) becomes

$$L(y) = \left(\frac{\text{SS}_{\text{res}, H_0}}{\text{SS}_{\text{res}}} \right)^n.$$

We have shown that the likelihood ratio test rejects H_0 whenever

$$\frac{\text{SS}_{\text{res}, H_0}}{\text{SS}_{\text{res}}} \geq C,$$

where C is chosen such that

$$\sup_{\theta \in H_0} \mathbb{P}_\theta \left(\frac{\text{SS}_{\text{res}, H_0}}{\text{SS}_{\text{res}}} \geq C \right) = \alpha,$$

if α is the significance level of the test.

We need to define one more distribution:

THE F -DISTRIBUTION. Let $Z \sim N_m(0, I)$ and $W \sim N_n(0, I)$ be independent multivariate standard normal random vectors. Define the random variable

$$F = \frac{Z_1^2 + \dots + Z_m^2}{W_1^2 + \dots + W_n^2} \cdot \frac{n}{m} = \frac{\|Z\|^2/m}{\|W\|^2/n}.$$

Then F has by definition an F -distribution with m and n degrees of freedom, denoted by

$$F \sim F(m, n).$$

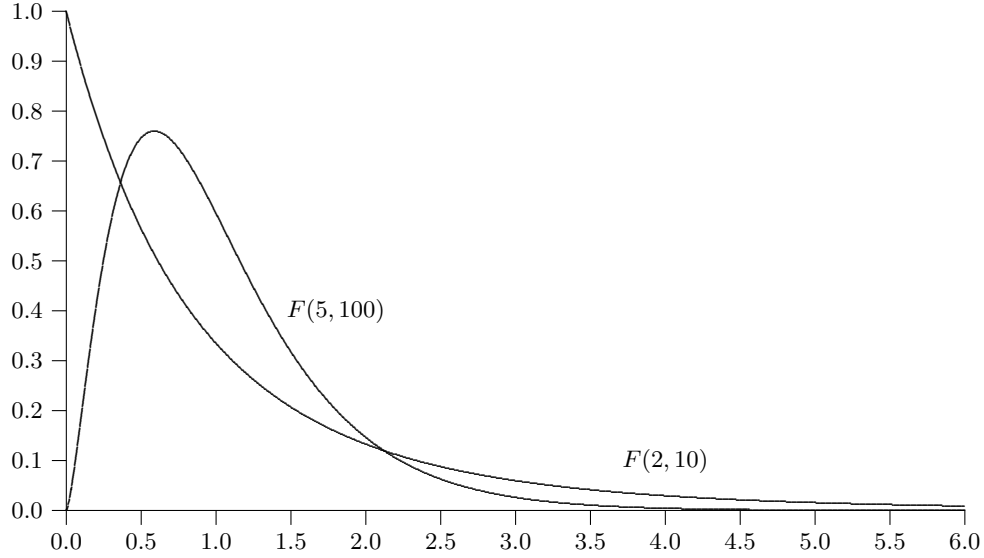


Figure 2.4: Densities of F -distributions.

Figure 2.4 shows a two densities of F -distributions with different degrees of freedom. Critical values for the F -distribution at different degrees of freedom can be found in Table 3.7. Note that the table only gives critical values for a significance level $\alpha = 5\%$.

We can prove that if B_0 is an affine subspace of $\mathbb{R}^p$ with dimension $p - q$, and X is of full rank, then under the null hypothesis we have

$$F := \left(\frac{\text{SS}_{\text{res}, H_0}}{\text{SS}_{\text{res}}} - 1 \right) \frac{n - p}{q} = \frac{\text{SS}_{\text{res}, H_0} - \text{SS}_{\text{res}}}{\text{SS}_{\text{res}}} \frac{n - p}{q} \sim F(q, n - p).$$

Clearly, the likelihood ratio test is equivalent to a test based on this F , which turns out to have a distribution under H_0 which is independent of the unknown parameters σ^2 and $\beta \in B_0$. The likelihood ratio test rejects $H_0 : \beta \in B_0$ whenever

$$\frac{\text{SS}_{\text{res}, H_0} - \text{SS}_{\text{res}}}{\text{SS}_{\text{res}}} \frac{n - p}{q} \geq F_\alpha(m, n).$$

Here, $F_\alpha(m, n)$ denotes the right critical value of the F -distribution with m and n degrees of freedom, that is, if $F \sim F(m, n)$, then

$$P(F \geq F_\alpha(m, n)) = \alpha.$$

See also Table 3.7. To summarize:

THE F -TEST. Consider the linear model

$$Y = X\beta + U$$

under the normality assumptions (so $U \sim N_n(0, \sigma^2 I)$). Consider the test

$$H_0 : \beta \in B_0 \quad \text{against} \quad H_1 : \beta \notin B_0,$$

where B_0 is an affine subspace of R^p of dimension $p - q$. Let X be of full rank. Then under H_0 we have that

$$\frac{SS_{\text{res}, H_0} - SS_{\text{res}}}{SS_{\text{res}}} \frac{n - p}{q} \sim F(q, n - p)$$

and we reject H_0 in favor of H_1 at a significance level α whenever

$$\frac{SS_{\text{res}, H_0} - SS_{\text{res}}}{SS_{\text{res}}} \frac{n - p}{q} \geq F_\alpha(m, n). \quad (2.16)$$

As an example, we will consider our abrasion loss data and test the hypothesis

$$H_0 : \beta_1 = -7 \text{ and } \beta_2 = 0.$$

So

$$B_0 = \{(\beta_0, \beta_1, \beta_2)^t \in \mathbb{R}^3 : \beta_1 = -7 \text{ and } \beta_2 = 0\}$$

and $\dim(B_0) = 1$, which means that $q = 2$. Note that q equals the number of constraints on the parameter vector β .

To perform the test (2.16), we need to calculate SS_{res, H_0} . This can be done by introducing a new linear model. Define

$$\tilde{Y}_i = Y_i - (-7 \cdot x_{1i}).$$

Since according to H_0 we have

$$Y_i = \beta_0 - 7 \cdot x_{1i} + 0 \cdot x_{2i} + U_i,$$

we get that under H_0

$$\tilde{Y}_i = \beta_0 + U_i.$$

This is again a linear model, for which we can calculate the residual sum of squares.

Quick exercise 2.8 Show that in this example,

$$SS_{\text{res}, H_0} = \sum_{i=1}^n \left(\tilde{Y}_i - \frac{1}{n} \sum_{j=1}^n \tilde{Y}_j \right)^2.$$

It turns out that in this example,

$$SS_{\text{res}, H_0} = 114454.3.$$

Since $SS_{\text{res}} = 35950$, we get for our test statistic

$$\begin{aligned} F &= \frac{SS_{\text{res}, H_0} - SS_{\text{res}}}{SS_{\text{res}}} \frac{n - p}{q} \\ &= \frac{114454.3 - 35950}{35950} \frac{30 - 3}{2} \\ &= 29.48. \end{aligned}$$

In Table 3.7 we see that $F_{0.05}(2, 27) = 3.35 < F$, which means that we can reject H_0 at a significance level of 95%: there is enough evidence to conclude that (β_1, β_2) differs significantly from $(-7, 0)$, assuming that our original linear model is adequate.

We would like to give one more example. Consider the data set given in Table 2.3, concerning the density and the Janka hardness of 36 Australian eucalypt hardwoods. The data are taken from [6] and are originally from [11] page 43, Table 3.7.

Table 2.3: Density and hardness of Australian hardwood.

Density	Hardness	Density	Hardness	Density	Hardness
24.7	484	39.4	1210	53.4	1880
24.8	427	39.9	989	56.0	1980
27.3	413	40.3	1160	56.5	1820
28.4	517	40.6	1010	57.3	2020
28.4	549	40.7	1100	57.6	1980
29.0	648	40.7	1130	59.2	2310
30.3	587	42.9	1270	59.8	1940
32.7	704	45.8	1180	66.0	3260
35.6	979	46.9	1400	67.4	2700
38.5	914	48.2	1760	68.8	2890
38.8	1070	51.5	1710	69.1	2740
39.3	1020	51.5	2010	69.1	3140

Janka hardness of wood is quite difficult to measure. whereas the density can be determined relatively easily. We would therefore like to describe the relation between the Janka hardness and the density. A scatter plot of this data can be seen in Figure 2.5. Let us call the hardness of the i^{th} sample of wood y_i and the density x_i .

Looking at the scatterplot of Figure 2.5, an obvious model would be the standard linear model

$$Y_i = \beta_0 + \beta_1 x_i + U_i. \quad (2.17)$$

Calculating the least squares estimator gives us

$$(\hat{\beta}_0, \hat{\beta}_1) = (-1160.5, 57.51)$$

and a residual sum of squares

$$\text{SS}_{\text{res}} = 1.14 \cdot 10^6. \quad (2.18)$$

To see if our model is reasonable, we calculate R-squared:

$$R^2 = \frac{\text{SS}_{\text{corr.tot.}} - \text{SS}_{\text{res}}}{\text{SS}_{\text{corr.tot.}}} = 0.949.$$

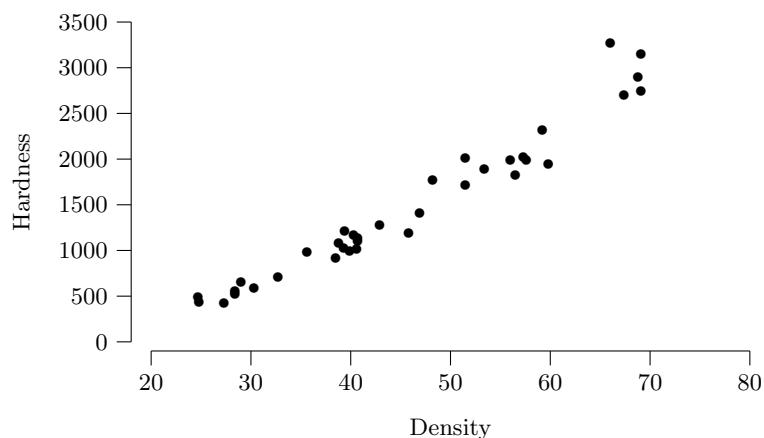


Figure 2.5: Scatterplot for Janka hardness data.

This looks promising, but we should also look at the residuals

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

This gives the scatter plot in Figure 2.6.

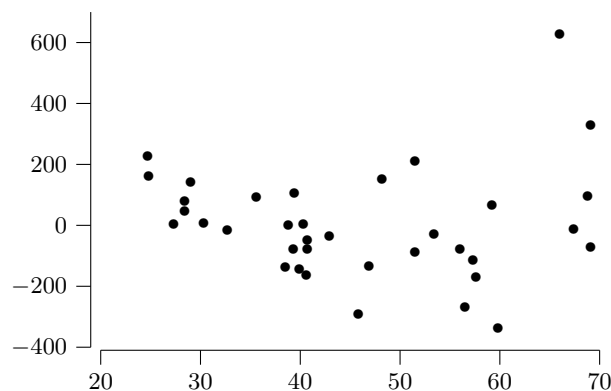


Figure 2.6: Residuals r_i versus x_i for Janka hardness data.

The residuals seem to have a slight tendency to be positive for small and big x_i , and to be negative for average x_i . This might be explained by a quadratic dependency of the hardness y_i on the density x_i . How can we determine whether there really is a significant quadratic dependency? One method could be the following: consider the linear model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + U_i$$

under the normality conditions, and determine whether or not we can reject the hypothesis

$$H_0 : \beta_2 = 0.$$

If we can reject H_0 , then the quadratic term is apparently significant; if we cannot reject H_0 , then we would decide against adding a quadratic term to our model.

To test H_0 , we need to calculate the residual sum of squares of the new model:

$$\text{SS}_{\text{res}} = 8.63 \cdot 10^5.$$

The model under H_0 equals the original model (2.17), so we get

$$\text{SS}_{\text{res}, H_0} = 1.14 \cdot 10^6.$$

This means that, since $q = 1$,

$$\begin{aligned} F &= \frac{\text{SS}_{\text{res}, H_0} - \text{SS}_{\text{res}}}{\text{SS}_{\text{res}}} \frac{n - p}{q} \\ &= \frac{1.14 \cdot 10^6 - 8.63 \cdot 10^5}{8.63 \cdot 10^5} \frac{36 - 3}{1} \\ &= 10.55. \end{aligned}$$

Also,

$$F_{0.05}(1, 33) \leq F_{0.05}(1, 30) = 4.17 \leq F.$$

This means that we can reject H_0 : it seems that the quadratic term is significant.

We would like to point out that when testing

$$H_0 : \beta_i = c,$$

we can show that

$$F = \frac{\text{SS}_{\text{res}, H_0} - \text{SS}_{\text{res}}}{\text{SS}_{\text{res}}} \frac{n - p}{q} = \left(\frac{\hat{\beta}_i - c}{\text{SE}(\hat{\beta}_i)} \right)^2.$$

We already know that under H_0 ,

$$\frac{\hat{\beta}_i - c}{\text{SE}(\hat{\beta}_i)} \sim t(n - p)$$

and we can indeed show that

$$t(n - p)^2 \stackrel{d}{=} F(1, n - p).$$

This means that we can reject H_0 whenever

$$\frac{\hat{\beta}_i - c}{\text{SE}(\hat{\beta}_i)} < -t_{n-p, \alpha/2} \quad \text{or} \quad \frac{\hat{\beta}_i - c}{\text{SE}(\hat{\beta}_i)} > t_{n-p, \alpha/2}.$$

In our Janka hardness example, we get

$$\frac{\hat{\beta}_2}{\text{SE}(\hat{\beta}_2)} = 3.248,$$

which is indeed bigger than $t_{33, 0.025} \leq t_{30, 0.025} = 2.042$.

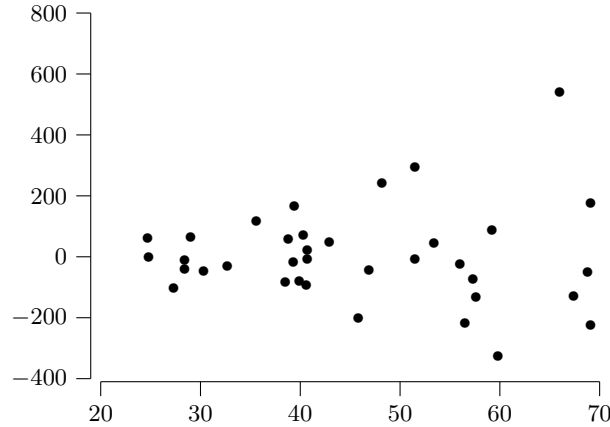


Figure 2.7: Residuals r_i versus x_i for Janka hardness data with model including quadratic term.

We could also determine a P-value for our test, but we need a computer for that. According to Equation (1.6), the P-value of this test is given by

$$\text{P-value} = P(F(1, 33) > 10.55) = 0.00267.$$

Finally, we could also look at the residuals for the model including the quadratic term; see Figure 2.7. Although there is no visible trend in the residuals, the plot does clearly suggest that the data are heteroscedastic: the variance seems to increase as the density gets higher. We might have to change our model to allow for this increase in variance. Note that the least squares estimator will still be a good estimator in such a model.

2.8 Testing without normality assumptions

After fitting a linear model, we can look at the residuals

$$R_i = Y_i - (X\hat{\beta})_i.$$

If $\hat{\beta}$ is close to the “true” β , we would get $R_i \approx U_i$. Under the normality conditions, this would imply that the residuals r_i can approximately be seen as independent realizations of $N(0, \sigma^2)$ distributed random variables. This gives us a way to check the normality assumptions. We have already seen that looking at the residuals might reveal the heteroscedasticity of the data, whereas the normality assumption implies homoscedasticity. In this section we will assume homoscedasticity of the data, so we will assume that the residuals can be modelled as realizations of independent, identically distributed random variables (namely, the U_i ’s), but we will not assume that these random variables are normally distributed.

Consider again the abrasion loss data. We will work with the model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + U_i. \quad (2.19)$$

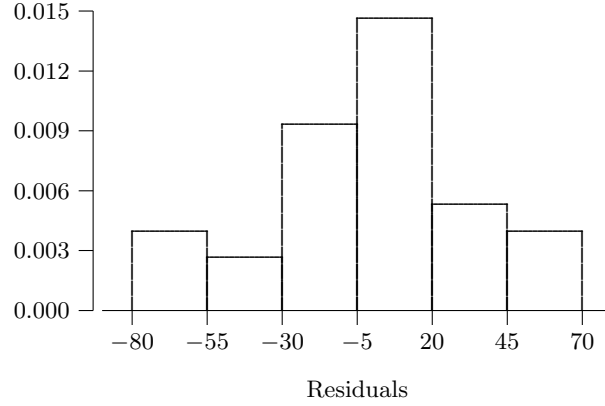


Figure 2.8: Histogram of the residuals of the abrasion loss data.

We have already calculated $\hat{\beta}$ (see for example Table 2.2), so we can make a histogram of the residuals r_i , see Figure 2.8.

In the following chapter we will see ways to test the null hypothesis that the r_i 's are realizations of i.i.d. normal random variables, but for now we take the asymmetry and bimodality of the histogram to be reasons enough not to assume a normal distribution for the U_i 's. The question now is: how do we test a null hypothesis, if we do not have the normality assumption? The answer: bootstrap!

Suppose we want to test the hypothesis

$$H_0 : \beta \in B_0 \quad \text{against} \quad H_1 : \beta \notin B_0,$$

where B_0 is a subset of $\mathbb{R}^p$ such that $\mathbb{R}^p \setminus B_0$ is dense in $\mathbb{R}^p$ (B_0 is in this sense a thin subset of $\mathbb{R}^p$). We will use the same test statistic as under the normality assumption:

$$F = \frac{\text{SS}_{\text{res}, H_0} - \text{SS}_{\text{res}}}{\text{SS}_{\text{res}}} \frac{n - p}{q}.$$

The problem is that we do not know the distribution of this F under H_0 , since we have not made any assumption about the distribution of the U_i 's. To approximate this distribution, we define the following bootstrap model as an approximation to our actual model under H_0 :

$$Y_i^* = (X\hat{\beta}_{H_0})_i + U_i^* = \hat{\beta}_{H_0,0} + \hat{\beta}_{H_0,1}x_{i1} + \hat{\beta}_{H_0,2}x_{i2} + U_i^*, \quad (2.20)$$

where the U_i^* 's are distributed according to the empirical distribution F_n of the residuals r_i and $\hat{\beta}_{H_0}$ is the least squares estimator belonging to the original model (2.19), but with β restricted to B_0 , that is

$$\text{SS}_{\text{res}}(\hat{\beta}_{H_0}) = \inf_{\beta \in B_0} \|Y - X\beta\|^2.$$

In fact, model (2.20) contains no unknown parameters; it is just used to find the distribution of F^* , which is the analogue of F , so

$$F^* = \frac{\text{SS}_{\text{res}, H_0}^* - \text{SS}_{\text{res}}^*}{\text{SS}_{\text{res}}^*} \frac{n - p}{q},$$

where

$$\text{SS}_{\text{res}}^* = \inf_{\beta \in \mathbb{R}^p} \|Y^* - X\beta\|^2$$

and

$$\text{SS}_{\text{res}, H_0}^* = \inf_{\beta \in B_0} \|Y^* - X\beta\|^2.$$

The bootstrap idea is that the distribution of F^* (which is known, in principle) is approximately equal to the distribution of F under H_0 , since $\hat{\beta}_{H_0}$ is, when the null hypothesis H_0 is actually true, approximately the “true” β belonging to model (2.19), and the distribution of U_i^* is close to the “true” distribution of U_i .

How are we going to determine a critical value for F^* ? Simple! We will just simulate a bootstrap data set $y_1^*, \dots, y_n^*$ by calculating the vector $X\hat{\beta}_{H_0}$ and adding to each component a realization of U_i^* , which is nothing more than a random pick out of all the residuals $r_1, \dots, r_n$, each with equal probability $1/n$. This is equivalent to the fact that U_i^* is distributed according to the empirical distribution function F_n . Note that since the original model (2.19) contains an intercept, we will have that

$$\sum_{i=1}^n r_i = 0,$$

which implies that $E[U_i^*] = \bar{r}_n = 0$ (see also Exercise 2.2). This means that for our bootstrap model (2.20), we actually have the Gauss-Markov conditions.

Once we have a bootstrap data set $y_1^*, \dots, y_n^*$, we calculate F_1^* , our first realization of F^* . Then we repeat these steps a lot of times, let us say a 1000 times, so we have $F_1^*, \dots, F_{1000}^*$. Then we need to find the $(1 - \alpha) \cdot 100$ empirical percentile of these $m = 1000$ values, so

$$c_\alpha^* = F_{(\lfloor \alpha(m+1) \rfloor)}^* + (\alpha(m+1) - \lfloor \alpha(m+1) \rfloor)(F_{(\lfloor \alpha(m+1) \rfloor + 1)}^* - F_{(\lfloor \alpha(m+1) \rfloor)}^*),$$

where $F_{(i)}^*$ denotes the i^{th} order statistic of $F_1^*, \dots, F_n^*$. If m is big enough, we will have in good approximation that

$$P(F^* > c_\alpha^*) = \alpha.$$

The bootstrap idea is now that we also have in reasonable approximation that under H_0

$$P(F > c_\alpha^*) \approx \alpha.$$

This means that we will reject H_0 at significance level α if $F > c_\alpha^*$.

As an example, let us again test the hypothesis

$$H_0 : \beta_1 = -7 \text{ and } \beta_2 = 0$$

in model (2.19) for the abrasion loss data, using the empirical bootstrap. We calculate $\hat{\beta}_{H_0}$ as described in and just before Quick Exercise 2.8:

$$\hat{\beta}_{H_0} = \begin{pmatrix} 667.3 \\ -7 \\ 0 \end{pmatrix}.$$

We have already seen that

$$F = 29.48.$$

Now we calculate F^* 1000 times; the results are in Figure 2.9. We have added a dotted histogram that gives the probability of each bin according to the $F(2, 27)$ distribution. We can note two things from this histograms: the value of our

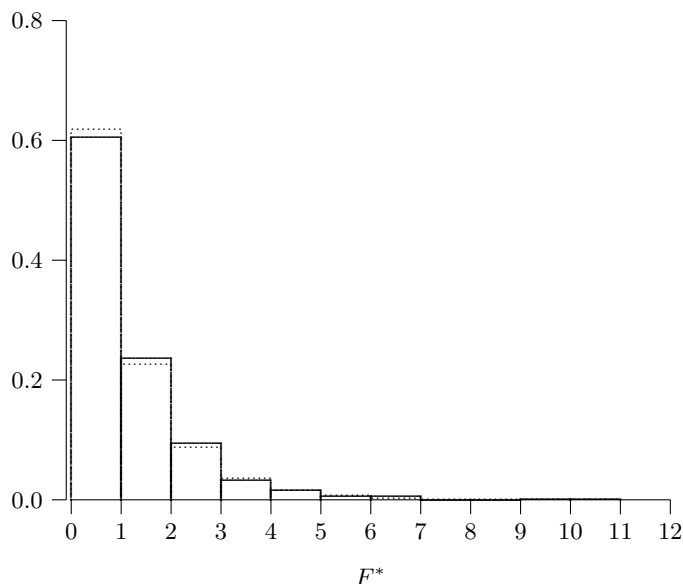


Figure 2.9: Histogram of 1000 bootstrapped F^* 's. Dashed bars indicate the $F(2, 27)$ distribution.

test statistic $F = 29.48$ is quite far away from all the bootstrapped values for F^* , which means that we can reject H_0 . Furthermore, the distribution of F^* is quite close to the $F(2, 27)$ distribution, which we would get if the U_i^* were normally distributed. This is one of the reasons the bootstrap method works quite well: the distribution of F does not depend a lot on the true distribution of the U_i 's.

2.9 Solutions to the quick exercises

2.1 We could make a scatter plot of abrasion loss against hardness and abrasion loss against tensile strength separately. We could also make a 3D-scatter plot of the data and rotate it to see the three dimensional structure. Of course many other ideas are possible!

2.2

$$X = \begin{pmatrix} 1 & x_{11}^2 & e^{-x_{21}} & x_{11}x_{21} \\ 1 & x_{12}^2 & e^{-x_{22}} & x_{12}x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n}^2 & e^{-x_{2n}} & x_{1n}x_{2n} \end{pmatrix}.$$

2.3 For the standard linear model, the regression-matrix is given by

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

This means that

$$X^t X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

so

$$(X^t X)^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Also,

$$X^t Y = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

which leads us to

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i Y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n Y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

and

$$\beta_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n.$$

2.4 We know that

$$(AY + BZ)_i = \sum_{k=1}^{n_1} A_{ik} Y_k + \sum_{l=1}^{n_2} B_{il} Z_l.$$

Now we use the linearity of expectation for random variables:

$$\begin{aligned} E[(AY + BZ)_i] &= E \left[\sum_{k=1}^{n_1} A_{ik} Y_k + \sum_{l=1}^{n_2} B_{il} Z_l \right] \\ &= \sum_{k=1}^{n_1} A_{ik} E[Y_k] + \sum_{l=1}^{n_2} B_{il} E[Z_l] \\ &= \sum_{k=1}^{n_1} A_{ik} E[Y]_k + \sum_{l=1}^{n_2} B_{il} E[Z]_l \\ &= (AE[Y] + BE[Z])_i. \end{aligned}$$

2.5 Define the matrix $A = \begin{pmatrix} 1 & 1 \end{pmatrix}$, then $Y_1 + Y_2 = AY$. Lemma 2.2 therefore shows that $Y_1 + Y_2$ is normally distributed. Also,

$$E[Y_1 + Y_2] = 2 + 1 = 3 \quad \text{and} \quad \text{Var}(Y_1 + Y_2) = AC(Y)A^t = 7.$$

So in fact, $Y_1 + Y_2 \sim N(3, (\sqrt{7})^2)$. This means that

$$\begin{aligned} P(Y_1 + Y_2 > 2) &= P \left(\frac{(Y_1 + Y_2) - 3}{\sqrt{7}} > \frac{2 - 3}{\sqrt{7}} \right) \\ &= P(N(0, 1) > -0.38) \\ &= 1 - P(N(0, 1) > 0.38) \\ &= 0.65. \end{aligned}$$

2.6

$$Y \sim N_n(X\beta, \sigma^2 I).$$

2.7 We know that

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X^t X)^{-1}).$$

This shows that $\hat{\beta}_i$ is a normally distributed random variable (Lemma 2.2). Furthermore,

$$E[\hat{\beta}_i] = E[\hat{\beta}]_i = \beta_i$$

and

$$\text{Var}(\hat{\beta}_i) = C(\hat{\beta})_{ii} = \sigma^2(X^t X)^{-1}_{ii}.$$

This proves that

$$\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{(X^t X)^{-1}_{ii}}} \sim N(0, 1).$$

2.8 We easily see that if

$$\tilde{Y}_i = \beta_0 + U_i,$$

then $\hat{\beta}_0 = \frac{1}{n} \sum_{j=1}^n \tilde{Y}_j$. This shows that

$$\begin{aligned} \text{SS}_{\text{res}, H_0} &= \sum_{i=1}^n (\tilde{Y}_i - \hat{\beta}_0)^2 \\ &= \sum_{i=1}^n \left(\tilde{Y}_i - \frac{1}{n} \sum_{j=1}^n \tilde{Y}_j \right)^2 \end{aligned}$$

2.10 Exercises

2.1 Suppose we have a random vector

$$Y \sim N_2 \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 9 & 1 \\ 1 & 4 \end{pmatrix} \right).$$

- a. Bereken $P(Y_1 + 2Y_2 > 3)$.
- b. Bepaal $a \in \mathbb{R}$ zodat Y_1 onafhankelijk is van $aY_1 + Y_2$.

2.2 Suppose we have a linear model

$$Y = X\beta + U$$

with an intercept, that is, one of the columns of X consists of only ones.

- a. Why would a column of only ones in the X -matrix imply that there exists an intercept in the model?
- b. Use the normal equations to show that if R_i is the i^{th} residual, then

$$\sum_{i=1}^n R_i = 0.$$

Chapter 3

Nonparametric statistics

In this chapter we will look at two examples of nonparametric models.

3.1 Introduction

In this chapter we will focus on two examples of *nonparametric models*. These are models where the possible distributions of the data cannot be described by a parametric family such as the multivariate normal distributions

$$\{N(\mu, \Sigma) : \mu \in \mathbb{R}^n, \Sigma \text{ a positive definite symmetric } n \times n \text{ matrix}\}$$

or the exponential distributions

$$\{Exp(\lambda) : \lambda > 0\}.$$

These families are called parametric since they can be described by a finite dimensional parameter. A typical example of a nonparametric model is

$$\{F : \mathbb{R} \rightarrow [0, 1] : F \text{ is a distribution function}\}.$$

This is the set of all possible distributions on $\mathbb{R}$. Another example might be

$$\{f : \mathbb{R} \rightarrow \mathbb{R} : f \geq 0, \int f(x)dx = 1 \text{ and } f \in C^2(\mathbb{R})\}.$$

This is the class of all densities on $\mathbb{R}$ which are twice continuously differentiable. It turns out that a lot of the general theory of Chapter 1 can be used, maybe in slightly altered form, for these nonparametric models, as we will see in the following sections.

3.2 Kernel density estimators

We consider a sample $X_1, \dots, X_n$ from a density f , where we only assume that the density is twice continuously differentiable. So in the notation of Chapter 1, we get that our statistical model equals

$$\Theta = \{f : \mathbb{R} \rightarrow \mathbb{R} : f \geq 0, \int f(x)dx = 1 \text{ and } f \in C^2(\mathbb{R})\}.$$

Given our sample, we would like to find an estimator for the underlying density $f \in \Theta$. To make things a bit more tangible, we consider the following data set, giving the length of the forearm (in inches) of 140 adult males; see Table 3.1. The data are taken from [6] and are originally from [8] page 357-462.

Table 3.1: Length of 140 forearms.

17.3	18.4	20.9	16.8	18.7	20.5	17.9	20.4	18.3	20.5
19.0	17.5	18.1	17.1	18.8	20.0	19.1	19.1	17.9	18.3
18.2	18.9	19.4	18.9	19.4	20.8	17.3	18.5	18.3	19.4
19.0	19.0	20.5	19.7	18.5	17.7	19.4	18.3	19.6	21.4
19.0	20.5	20.4	19.7	18.6	19.9	18.3	19.8	19.6	19.0
20.4	17.3	16.1	19.2	19.6	18.8	19.3	19.1	21.0	18.6
18.3	18.3	18.7	20.6	18.5	16.4	17.2	17.5	18.0	19.5
19.9	18.4	18.8	20.1	20.0	18.5	17.5	18.5	17.9	17.4
18.7	18.6	17.3	18.8	17.8	19.0	19.6	19.3	18.1	18.5
20.9	19.8	18.1	17.1	19.8	20.6	17.6	19.1	19.5	18.4
17.7	20.2	19.9	18.6	16.6	19.2	20.0	17.4	17.1	18.3
19.1	18.5	19.6	18.0	19.4	17.1	19.9	16.3	18.9	20.7
19.7	18.5	18.4	18.7	19.3	16.3	16.9	18.2	18.5	19.3
18.1	18.0	19.5	20.3	20.1	17.2	19.5	18.8	19.2	17.7

A histogram of this data can be seen in Figure 3.1.

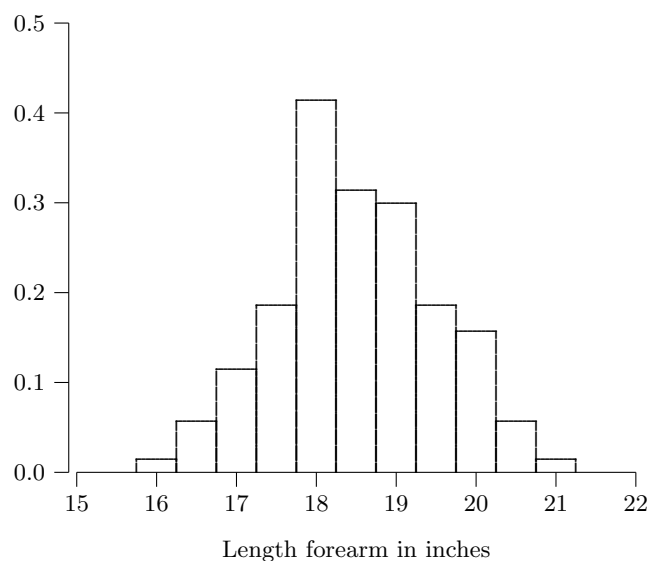


Figure 3.1: Histogram of the length of forearms.

If we assume that the distribution of the length of forearms can be described by a density $f \in \Theta$, how can we make an estimate of f ? Another obvious question in this case is whether it is plausible to model the length of forearms by a normal distribution.

The first thing we might think of is to try and find the maximum likelihood estimator of the underlying f , given our model Θ . However, it is not hard to see that we can make the likelihood as high as we like, by concentrating all the probability mass around the data points.

Quick exercise 3.1 Suppose our data is $\{1, 3, 4\}$. Find a sequence $f_n \in \Theta$ such that

$$\lim_{n \rightarrow \infty} f_n(1) \cdot f_n(3) \cdot f_n(4) = +\infty.$$

In [3], Chapter 15, the notion of a *kernel density estimate* is introduced. We will repeat its definition. Given a realization $x_1, \dots, x_n$ of a random sample $X_1, \dots, X_n$, a kernel density estimate $\hat{f}_{n,h}$ of the underlying density f is the function

$$\hat{f}_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - x_i}{h}\right), \quad (3.1)$$

where $h > 0$ is the *bandwidth* and K is a *kernel*.

DEFINITION OF A KERNEL. A bounded function $K : \mathbb{R} \rightarrow \mathbb{R}$ is called a *kernel* if

- (K1) K is a density, that is, $K \geq 0$ and $\int K(s)ds = 1$.
- (K2) K is symmetric around zero, that is, $K(s) = K(-s)$.
- (K3) K has a finite second moment, i.e., $\int s^2 K(s)ds < +\infty$.

Quick exercise 3.2 Show that for any choice of the bandwidth h , the function $\hat{f}_{n,h}$ defined by (3.1) is a density.

If we take K to be the standard normal density and $h = 0.35$, we get the kernel density estimate shown in Figure 3.2. In the same picture we show the density of a normal distribution with

$$\mu = \bar{x}_n = 18.80 \quad \text{and} \quad \sigma^2 = s_n^2 = 1.256.$$

From this picture there does not seem to be a reason to reject the hypothesis that the length of forearms is normally distributed: the two densities seem reasonably similar. However, we need some method to make this conclusion rigorous. Also, we have to keep in mind that the kernel density estimate in Figure 3.2 uses a particular choice of the bandwidth h . We could choose a bandwidth such that the two densities are very different, but that might not be a reasonable choice for h . This illustrates that it is important to have an objective way of choosing the bandwidth.

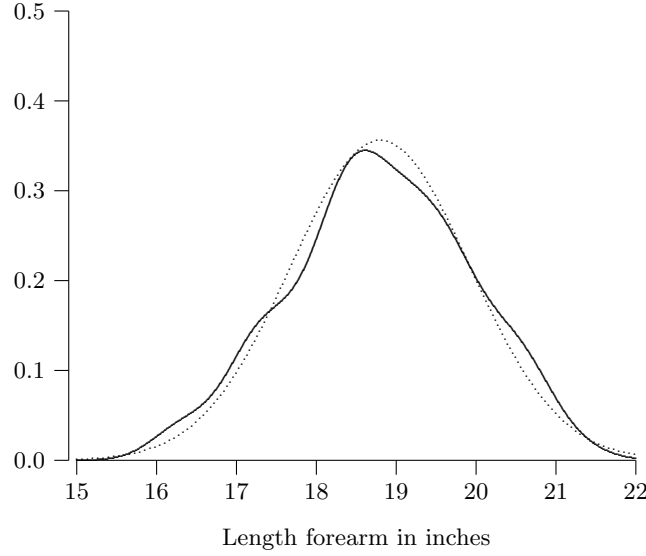


Figure 3.2: Kernel density estimate of the length of forearms, using a standard normal kernel and $h = 0.35$.

Choice of bandwidth

To have an idea of what would be a reasonable choice for the bandwidth h , we will first consider the kernel density estimator $\hat{f}_{n,h}$ as a random function, depending on the sample $X_1, \dots, X_n \sim f$:

$$\hat{f}_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - X_i}{h}\right).$$

With a slight abuse of notation we will write $\hat{f}_{n,h}(t)$ for the *estimator* of $f(t)$ (a random variable), as well as for the *estimate* of $f(t)$ (a realization of the random variable). It is straightforward to calculate the expectation of $\hat{f}_{n,h}(t)$:

$$\begin{aligned} \mathbb{E}[\hat{f}_{n,h}(t)] &= \frac{1}{nh} \sum_{i=1}^n \mathbb{E}\left[K\left(\frac{t - X_i}{h}\right)\right] \\ &= \frac{1}{nh} \sum_{i=1}^n \int K\left(\frac{t - x}{h}\right) f(x) dx \\ &= \int K(s) f(t - hs) ds. \end{aligned} \tag{3.2}$$

For the variance of $\hat{f}_{n,h}(t)$ we get:

$$\begin{aligned} \text{Var}(\hat{f}_{n,h}(t)) &= \frac{1}{n^2 h^2} \sum_{i=1}^n \text{Var}\left(K\left(\frac{t - X_i}{h}\right)\right) \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n \left(\int K\left(\frac{t - x}{h}\right)^2 f(x) dx - \left(\int K\left(\frac{t - x}{h}\right) f(x) dx \right)^2 \right) \\ &= \frac{1}{nh} \int K(s)^2 f(t - hs) ds - \frac{1}{n} \left(\int K(s) f(t - hs) ds \right)^2. \end{aligned} \tag{3.3}$$

It is intuitively clear that if n gets bigger (we get more data points), then our bandwidth h should get smaller: the data points x_i close to t are important for estimating $f(t)$, and when n is large, we can consider points close to t (that is, choose a small bandwidth) and still have enough points to make a good estimate. This means that we want to consider what happens if $n \rightarrow \infty$ and $h \downarrow 0$. In this case, we see that the second term in (3.3) goes to zero, so the variance is of order $1/(nh)$. We would like to get a small variance, so we choose $h \downarrow 0$ in such a way that $nh \rightarrow \infty$. In this situation we have the following proposition.

Proposition 3.1 *Consider a sample $X_1, \dots, X_n$ from a density f where f is twice continuously differentiable in the point $t \in \mathbb{R}$. Suppose that either f has a bounded second derivative on $\mathbb{R}$ or that K has a bounded support. Define the kernel density estimator*

$$\hat{f}_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - X_i}{h}\right).$$

Then

$$\mathbb{E}[\hat{f}_{n,h}(t)] = f(t) + \frac{1}{2}h^2 f''(t) \int s^2 K(s) ds + o(h^2), \quad h \downarrow 0$$

and

$$\text{Var}(\hat{f}_{n,h}(t)) = \frac{1}{nh} f(t) \int K^2(s) ds + o\left(\frac{1}{nh}\right), \quad h \downarrow 0 \text{ and } nh \rightarrow \infty.$$

Proof: Looking at (3.2) and (3.3), it should be clear that we want to use a Taylor expansion of f around t :

$$f(t - sh) = f(t) - hsf'(t) + \frac{1}{2}h^2 s^2 f''(t) + h^2 s^2 R(t, sh), \quad (3.4)$$

where the rest term $R(t, sh)$ remains bounded if sh remains bounded; if f has a bounded second derivative on $\mathbb{R}$, then $R(t, sh)$ remains bounded for all values of sh . Furthermore,

$$\lim_{h \rightarrow 0} R(t, sh) = 0.$$

Using (3.4) in (3.2) yields

$$\begin{aligned} \mathbb{E}[\hat{f}_{n,h}(t)] &= \int K(s) \left(f(t) - hsf'(t) + \frac{1}{2}h^2 s^2 f''(t) + h^2 s^2 R(t, sh) \right) ds \\ &= f(t) + \frac{1}{2}h^2 f''(t) \int s^2 K(s) ds + h^2 \int R(t, sh) s^2 K(s) ds. \end{aligned}$$

Here we have used the symmetry of K (and the fact that K has a finite first moment) to conclude that

$$\int sK(s) ds = 0.$$

We now use a famous theorem from integration theory:

LEBESGUE'S DOMINATED CONVERGENCE THEOREM. Suppose $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is such that

$$\lim_{h \rightarrow 0} g(x, h) = 0 \quad \forall x \in \mathbb{R}.$$

Also, there exists a function $\phi : \mathbb{R} \rightarrow \mathbb{R}$ and $\varepsilon > 0$ such that

$$|g(x, h)| \leq \phi(x) \quad (\forall x \in \mathbb{R} \quad \forall |h| < \varepsilon) \quad \text{and} \quad \int \phi(x) dx < +\infty.$$

Then

$$\lim_{h \rightarrow 0} \int g(x, h) dx = 0.$$

See for example [1] for a thorough treatment of this theorem. Since K has a finite second moment (condition (K3)), Lebesgue's Dominated Convergence Theorem tells us that

$$\lim_{h \rightarrow 0} \int R(t, sh) s^2 K(s) ds = 0.$$

This proves that

$$\mathbb{E}[\hat{f}_{n,h}(t)] = f(t) + \frac{1}{2} h^2 f''(t) \int s^2 K(s) ds + o(h^2), \quad h \downarrow 0.$$

In Equation (3.3) we only need the boundedness of f (which follows from the fact that f'' is bounded and that f is a density) or the fact that K has a bounded support in order to use Lebesgue's Dominated Convergence Theorem. This would give us

$$\lim_{h \rightarrow 0} \int K(s)^2 (f(t - hs) - f(t)) ds = 0$$

and

$$\lim_{h \rightarrow 0} h \left(\int K(s) f(t - hs) ds \right)^2 = 0.$$

These two equations show that

$$\text{Var}(\hat{f}_{n,h}(t)) = \frac{1}{nh} f(t) \int K(s)^2 ds + o\left(\frac{1}{nh}\right), \quad h \downarrow 0 \text{ and } nh \rightarrow \infty.$$

□

Quick exercise 3.3 Prove the last statement of the proof of Proposition 3.1.

We can use Proposition 3.1 to calculate the first order of the *Mean Squared Error*, which is defined by

$$\text{MSE}(\hat{f}_{n,h}(t)) = \mathbb{E}\left[\left(\hat{f}_{n,h}(t) - f(t)\right)^2\right].$$

We remark that

$$\begin{aligned}
 \text{MSE}(\hat{f}_{n,h}(t)) &= \mathbb{E} \left[\left(\hat{f}_{n,h}(t) - \mathbb{E}[\hat{f}_{n,h}(t)] + \mathbb{E}[\hat{f}_{n,h}(t)] - f(t) \right)^2 \right] \\
 &= \mathbb{E} \left[\left(\hat{f}_{n,h}(t) - \mathbb{E}[\hat{f}_{n,h}(t)] \right)^2 \right] + \left(\mathbb{E}[\hat{f}_{n,h}(t)] - f(t) \right)^2 + \\
 &\quad + 2\mathbb{E} \left[\hat{f}_{n,h}(t) - \mathbb{E}[\hat{f}_{n,h}(t)] \right] \left(\mathbb{E}[\hat{f}_{n,h}(t)] - f(t) \right) \\
 &= \text{Var}(\hat{f}_{n,h}(t)) + \text{Bias}[\hat{f}_{n,h}(t)]^2.
 \end{aligned}$$

Recall that the bias of the estimator $\hat{f}_{n,h}(t)$ of $f(t)$ is defined as

$$\text{Bias}[\hat{f}_{n,h}(t)] = \mathbb{E}[\hat{f}_{n,h}(t)] - f(t).$$

Proposition 3.1 now tells us that

$$\text{MSE}(\hat{f}_{n,h}(t)) = \frac{1}{nh} f(t) \int K(s)^2 ds + \frac{1}{4} h^4 \left(f''(t) \int s^2 K(s) ds \right)^2 + o(h^4) + o\left(\frac{1}{nh}\right). \quad (3.5)$$

Also, we see that the variance term in the MSE gets smaller as h gets bigger, whereas the bias term gets smaller as h gets smaller. This is called the *bias-variance trade-off*. It is not hard to see that the MSE has the smallest order if the variance and the bias-squared are of the same order, and that this happens when

$$h_n = cn^{-1/5}.$$

So asymptotically (that is, as $n \rightarrow \infty$) an optimal choice of the bandwidth would be to take the bandwidth h_n of the order $n^{1/5}$. In that case, the MSE will be of the order $n^{4/5}$, which means that the difference between the estimator $\hat{f}_{n,h}(t)$ and $f(t)$ is of the order $n^{2/5}$. To compare, in parametric models we usually get that the difference between the estimator and the parameter is of order $n^{1/2}$ (think of the mean of a normal distribution). The nonparametric model is so big, that we can estimate the parameter slower than in a parametric model. However, the difference in speed is not dramatic.

Mean Integrated Square Error

Up until now, we have only looked at the behavior of the kernel density estimator at a fixed point t . However, since we have introduced the kernel density estimator as an estimator for the entire underlying density f , we would like to have a global (that is, for all $t \in \mathbb{R}$ at the same time) performance measure of our estimator. A natural extension of the above defined MSE would be the *Mean Integrated Squared Error*, or MISE:

$$\text{MISE}(\hat{f}_{n,h}) = \mathbb{E} \left[\int \left(\hat{f}_{n,h}(t) - f(t) \right)^2 dt \right] = \int \mathbb{E} \left[\left(\hat{f}_{n,h}(t) - f(t) \right)^2 \right] dt. \quad (3.6)$$

Under some mild integrability assumptions, we can integrate Equation (3.5) for the MSE over t to conclude that

$$\begin{aligned} \text{MISE}(\hat{f}_{n,h}) &= \frac{1}{nh} \int K(s)^2 ds + \frac{1}{4} h^4 \left(\int f''(t)^2 dt \right) \left(\int s^2 K(s) ds \right)^2 + \\ &\quad + o(h^4) + o\left(\frac{1}{nh}\right). \end{aligned} \quad (3.7)$$

We can minimize Equation (3.7) over h to find an asymptotically optimal bandwidth:

$$h_n = \left(\frac{\int K(s)^2 ds}{\left(\int f''(t)^2 dt \right) \left(\int s^2 K(s) ds \right)^2} \right)^{1/5} \cdot n^{-1/5}.$$

Of course, this formula still contains the unknown constant $\int f''(t)^2 dt$, but there are good methods to estimate this constant. For details, see [9] or [10]

Least squares cross-validation

So far we have looked at an asymptotically optimal choice for the bandwidth h . However, we might want to try and minimize $\text{MISE}(\hat{f}_{n,h})$ directly, for fixed n . The problem is of course that $\text{MISE}(\hat{f}_{n,h})$ depends on the unknown underlying density f , so we need to do some kind of approximation. First we will rewrite the MISE:

$$\begin{aligned} \text{MISE}(\hat{f}_{n,h}) &= \text{E} \left[\int \hat{f}_{n,h}(t)^2 dt - 2 \int \hat{f}_{n,h}(t) f(t) dt \right] + \int f(t)^2 dt \\ &= \text{E} \left[\int \hat{f}_{n,h}(t)^2 dt \right] - 2 \text{E} \left[\int \hat{f}_{n,h}(t) f(t) dt \right] + \int f(t)^2 dt \end{aligned} \quad (3.8)$$

If we minimize the MISE over h , we can forget about the third term in (3.8), since it does not depend on h . The first term depends on f (through the expectation), but if we assume that our kernel density estimate will be reasonable for the optimal choice of h (whatever that turns out to be), we can make the following approximation:

$$\text{E} \left[\int \hat{f}_{n,h}(t)^2 dt \right] \approx \int \hat{f}_{n,h}(t)^2 dt \quad (3.9)$$

For the second term in (3.8), we introduce the following estimator:

$$\hat{f}_{n,h}^{(i)}(t) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{t - X_j}{h}\right).$$

This is just the standard kernel density estimator, but now based on the sample without X_i . This estimator is also known as the *leave-one-out* estimator. The

merit of this estimator lies in the following relation:

$$\begin{aligned}
\mathbb{E}[\hat{f}_{n,h}^{(i)}(X_i)] &= \frac{1}{(n-1)h} \sum_{j \neq i} \mathbb{E}\left[K\left(\frac{X_i - X_j}{h}\right)\right] \\
&= \frac{1}{(n-1)h} \sum_{j \neq i} \int \int K\left(\frac{t-s}{h}\right) f(s)f(t) ds dt \\
&= \int \left(\frac{1}{nh} \sum_{i=1}^n \int K\left(\frac{t-s}{h}\right) f(s) ds\right) f(t) dt \\
&= \int \left(\mathbb{E}\left[\frac{1}{nh} \sum_{i=1}^n \int K\left(\frac{t-X_i}{h}\right)\right]\right) f(t) dt \\
&= \mathbb{E}\left[\int \hat{f}_{n,h}(t) f(t) dt\right].
\end{aligned}$$

This shows that it might be reasonable to make the following approximation for the second term of (3.8):

$$\mathbb{E}\left[\int \hat{f}_{n,h}(t) f(t) dt\right] \approx \frac{1}{n} \sum_{i=1}^n \hat{f}_{n,h}^{(i)}(X_i). \quad (3.10)$$

The two approximations (3.9) and (3.10) lead to the following method of choosing the bandwidth, called the *least squares cross-validation method*:

LEAST SQUARES CROSS-VALIDATION METHOD. Suppose we have a random sample $X_1, \dots, X_n$ from an unknown density f . Consider the kernel density estimator

$$\hat{f}_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - X_i}{h}\right).$$

The least squares cross-validation method chooses as the bandwidth the minimizer of the so-called *estimated MISE*, that is

$$\hat{h}_n = \arg \min_{h>0} \int \hat{f}_{n,h}(t)^2 dt - \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,h}^{(i)}(X_i) \quad (3.11)$$

where $\hat{f}_{n,h}^{(i)}$ is the leave-one-out estimator, so

$$\hat{f}_{n,h}^{(i)}(X_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right).$$

We have to be a bit careful when using the least squares cross validation method directly for the forearm data (see Remark 3.1), but it gives us a bandwidth

$$h = 0.57.$$

This kernel density estimate is shown in Figure 3.3. In the same picture we show the previous kernel density estimate ($h = 0.35$, found by visual inspection) and the best fitting normal density.

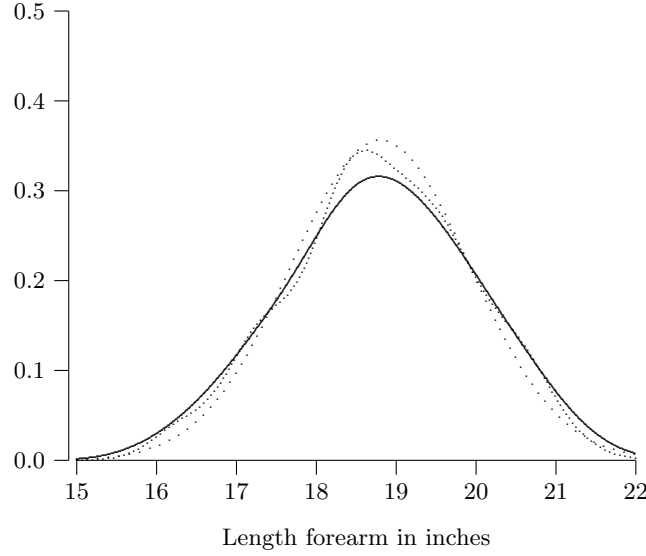


Figure 3.3: Kernel density estimate of the length of forearms, using a standard normal kernel and $h = 0.57$.

Remark 3.1 (Least squares cross-validation with ties in the data.)

In the forearm data there are a lot of ties. This is because the measured length was rounded off to 1/10 of an inch. This creates some difficulty for the least squares cross-validation method. To see why, let us see what happens to

$$\int \hat{f}_{n,h}(t)^2 dt - \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,h}^{(i)}(x_i) \quad (3.12)$$

when $h \downarrow 0$. We will assume that K has a support $[-1, 1]$, but if $K(s) \rightarrow 0$ fast enough (as $|s| \rightarrow \infty$), we get similar conclusions. Define

$$n(i) = \#\{j \neq i : x(j) = x(i)\}.$$

Note that $n(i) = 0$ for all i if there are no ties. Then if

$$h < \frac{1}{2} \min\{|x_i - x_j| : x_i \neq x_j\},$$

we get for $x_i \neq x_j$

$$K\left(\frac{t - x_i}{h}\right) K\left(\frac{t - x_j}{h}\right) = 0 \quad \forall t \in \mathbb{R}.$$

Therefore, the first term of Equation (3.12) becomes

$$\begin{aligned} \int \hat{f}_{n,h}(t)^2 dt &= \frac{1}{n^2 h^2} \sum_{i,j=1}^n \int K\left(\frac{t - x_i}{h}\right) K\left(\frac{t - x_j}{h}\right) dt \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n (1 + n(i)) \int K\left(\frac{t - x_i}{h}\right)^2 dt \\ &= \frac{1}{nh} \int K(s)^2 ds + \frac{1}{n^2 h} \sum_{i=1}^n n(i) \int K(s)^2 ds. \end{aligned}$$

The second term of Equation (3.12) becomes

$$\begin{aligned} \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,h}^{(i)}(x_i) &= \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i}^n K\left(\frac{x_i - x_j}{h}\right) \\ &= \frac{2}{n(n-1)h} \sum_{i=1}^n n(i)K(0). \end{aligned}$$

This shows that if $n(i) = 0$ for all i (so there are no ties), then (3.12), the estimate of the MISE used in the least squares cross-validation method, will go to $+\infty$ as $h \downarrow 0$. However, if there are ties, then the second term of (3.12) will blow up, and it might happen that (3.12) will go to $-\infty$ as $h \downarrow 0$. This would of course mean that the minimizer found in the least squares cross-validation method would be $h = 0$, which obviously doesn't make sense. With the forearm data we are in this last situation: Figure 3.4 shows the estimate for the MISE (Equation (3.12)) for different choices of h .

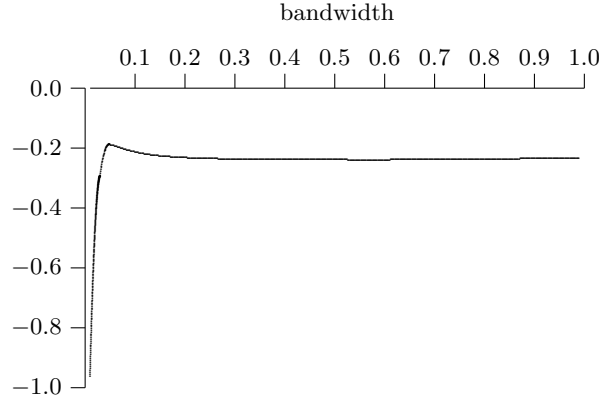


Figure 3.4: The estimated MISE for the forearm data.

One way to solve this problem is to find the minimum of the estimated MISE over $h > 0.1$. Another way is to slightly perturb the data, for example by defining a new data set $\tilde{x}$ as

$$\tilde{x}_i = x_i + \varepsilon_i,$$

where $\varepsilon_i \sim U(-0.05, 0.05)$. This way, the probability of having ties becomes 0. Both methods lead to almost the same value of the bandwidth, namely $h = 0.57$.

Testing for normality using kernel estimators

The least squares cross-validation method allows us to perform tests based on kernel density estimators, since it gives an automated way of choosing the bandwidth. As an example, we consider testing for normality. To be more precise, consider a sample $X_1, \dots, X_n$ from an unknown density $f \in C^2(\mathbb{R})$. Based on the data $x_1, \dots, x_n$ we would like to test the hypothesis

$$H_0 : f \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

against

$$H_1 : f \in C^2(\mathbb{R}) \setminus \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

We will choose our test statistic based on a kernel density estimator $\hat{f}_{n,h}$. If f is a normal density, $\hat{f}_{n,h}$ should be close to (the density of) $N(\hat{\mu}, \hat{\sigma}^2)$, where

$$\hat{\mu} = \bar{x}_n \quad \text{and} \quad \hat{\sigma}^2 = s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

Inspired by the likelihood ratio test, we look at the log-likelihood of the data as if it were a sample from the density $\hat{f}_{n,\hat{h}}$, where $\hat{h}$ is defined by the least squares cross validation method, and compare that to the likelihood of the data as if it were a sample from $N(\hat{\mu}, \hat{\sigma}^2)$. This gives us the following test statistic:

$$T_1(x) = T_1(x_1, \dots, x_n) = \sum_{i=1}^n \log \left(\hat{f}_{n,\hat{h}}(x_i) \right) - \sum_{i=1}^n \log \left(\phi \left(\frac{x_i - \hat{\mu}}{\hat{\sigma}} \right) / \hat{\sigma} \right).$$

Here ϕ is the standard normal density, so

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

If T_1 is too big, we would reject the null hypothesis that f is a normal density. In the case of the forearm data, we can calculate T_1 and get

$$T_1 = -1.0532.$$

The question is whether this is big or not. Before we consider this, we would also like to look at another possibility to compare $\hat{f}_{n,\hat{h}}$ to $N(\hat{\mu}, \hat{\sigma}^2)$, namely by defining the test statistic

$$T_2(x) = \int \left| \hat{f}_{n,\hat{h}}(t) - \frac{1}{\hat{\sigma}} \phi \left(\frac{t - \hat{\mu}}{\hat{\sigma}} \right) \right| dt.$$

This is known as the L^1 -distance between the two densities. Again we would reject H_0 when T_2 is big. For the forearm data we find

$$T_2 = 0.1334.$$

To see whether these two values are big enough to reject H_0 , we would have to find a critical value belonging to each of the two test statistics. The two critical values, $c_{1,\alpha}$ and $c_{2,\alpha}$, would have to satisfy the usual condition (1.5), which translates into

$$\sup_{\mu \in \mathbb{R}, \sigma > 0} P_{(\mu, \sigma^2)}(T_1(X) > c_{1,\alpha}) = \alpha \quad (3.13)$$

and

$$\sup_{\mu \in \mathbb{R}, \sigma > 0} P_{(\mu, \sigma^2)}(T_2(X) > c_{2,\alpha}) = \alpha. \quad (3.14)$$

Here, $X = (X_1, \dots, X_n)$, and as usual, when we denote $P_{(\mu, \sigma^2)}$, we mean that each X_i is $N(\mu, \sigma^2)$ distributed.

To calculate the critical values $c_{1,\alpha}$ and $c_{2,\alpha}$ we have the following useful result.

Proposition 3.2 Suppose $X_1, \dots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. Then the distributions of $T_1(X)$ and $T_2(X)$ do not depend on the parameters μ and σ .

This means that in order to calculate $c_{1,\alpha}$ and $c_{2,\alpha}$ using a bootstrap method, we only need to consider a standard normal sample.

Proof of Proposition 3.2: We will prove the result for $T_1(X)$; the result for $T_2(X)$ is left as an exercise (see Exercise 3.3). Define $Z_1, \dots, Z_n$ as a random sample from a $N(0, 1)$ distribution and define

$$X_i = \sigma Z_i + \mu.$$

Clearly, $X_1, \dots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. We will show that

$$T_1(X) = T_1(Z).$$

Define

$$\hat{f}_{n,h}^{(z)}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - Z_i}{h}\right)$$

and likewise $\hat{f}_{n,h}^{(x)}(t)$. Then

$$\begin{aligned} \hat{f}_{n,h}^{(x)}(t) &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - \sigma Z_i - \mu}{h}\right) \\ &= \frac{1}{\sigma} \frac{1}{nh/\sigma} \sum_{i=1}^n K\left(\frac{\frac{t-\mu}{\sigma} - Z_i}{h/\sigma}\right) \\ &= \frac{1}{\sigma} \hat{f}_{n,h/\sigma}^{(z)}\left(\frac{t - \mu}{\sigma}\right). \end{aligned}$$

Now define the estimated MISE (see Equation (3.11)) for the two cases:

$$\text{MISE}^{(z)}(h) = \int \hat{f}_{n,h}^{(z)}(t)^2 dt - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i} K\left(\frac{Z_i - Z_j}{h}\right)$$

and likewise $\text{MISE}^{(x)}$. Then

$$\begin{aligned} \text{MISE}^{(x)}(h) &= \int \hat{f}_{n,h}^{(x)}(t)^2 dt - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right) \\ &= \int \frac{1}{\sigma^2} \hat{f}_{n,h/\sigma}^{(z)}\left(\frac{t - \mu}{\sigma}\right)^2 dt - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i} K\left(\frac{Z_i - Z_j}{h/\sigma}\right) \\ &= \frac{1}{\sigma} \left(\int \hat{f}_{n,h/\sigma}^{(z)}(s)^2 ds - \frac{2}{n(n-1)h/\sigma} \sum_{i=1}^n \sum_{j \neq i} K\left(\frac{Z_i - Z_j}{h/\sigma}\right) \right) \\ &= \frac{1}{\sigma} \text{MISE}^{(z)}(h/\sigma). \end{aligned}$$

Now define the choice of bandwidth by the least squares cross-validation method:

$$\hat{h}^{(z)} = \arg \min_{h>0} \text{MISE}^{(z)}(h)$$

and likewise $\hat{h}^{(x)}$. Then

$$\begin{aligned} \hat{h}^{(x)} &= \arg \min_{h>0} \text{MISE}^{(x)}(h) \\ &= \arg \min_{h>0} \frac{1}{\sigma} \text{MISE}^{(z)}(h/\sigma) \\ &= \sigma \hat{h}^{(z)}. \end{aligned}$$

This means that the first term of our test statistic T_1 satisfies

$$\begin{aligned} \sum_{i=1}^n \log \left(\hat{f}_{n, \hat{h}^{(x)}}(X_i) \right) &= \sum_{i=1}^n \log \left(\frac{1}{\sigma} \hat{f}_{n, \hat{h}^{(x)}/\sigma}^{(z)} \left(\frac{X_i - \mu}{\sigma} \right) \right) \\ &= \sum_{i=1}^n \log \left(\frac{1}{\sigma} \hat{f}_{n, \hat{h}^{(z)}}^{(z)}(Z_i) \right). \end{aligned}$$

But since clearly

$$\hat{\mu}^{(x)} = \bar{X}_n = \sigma \bar{Z}_n + \mu = \sigma \hat{\mu}^{(z)} + \mu$$

and

$$\hat{\sigma}^{(x)} = \sigma \hat{\sigma}^{(z)},$$

we get for the second term of our test statistic T_1

$$\sum_{i=1}^n \log \left(\phi \left(\frac{X_i - \hat{\mu}^{(x)}}{\hat{\sigma}^{(x)}} \right) / \hat{\sigma}^{(x)} \right) = \sum_{i=1}^n \log \left(\frac{1}{\sigma} \phi \left(\frac{Z_i - \hat{\mu}^{(z)}}{\hat{\sigma}^{(z)}} \right) / \hat{\sigma}^{(z)} \right).$$

This indeed shows that

$$T_1(X) = T_1(Z).$$

□

Proposition 3.2 shows that if we find $c_{1,\alpha}$ such that

$$P_{(0,1)}(T_1(X) > c_{1,\alpha}) = \alpha,$$

then condition (3.13) is automatically satisfied. Similarly, for condition (3.14) we only need to find $c_{2,\alpha}$ such that

$$P_{(0,1)}(T_2(X) > c_{2,\alpha}) = \alpha.$$

Good approximations to $c_{1,\alpha}$ and $c_{2,\alpha}$ can be found relatively easily by using simulations. Our forearm data set consists of 140 measurements, so we simulated 10000 times a realization $x_1, \dots, x_{140}$ from a standard normal distribution, calculated for each of the 10000 samples the values of $T_1(x)$ and $T_2(x)$ and took the 95th percentile of these 10000 values for both statistics. We found

$$c_{1,0.05} = 10.78 \quad \text{and} \quad c_{2,0.05} = 0.2053.$$

Since for the forearm data

$$T_1 = -1.0532 < 10.78 \quad \text{and} \quad T_2 = 0.1334 < 0.2053,$$

both methods of testing the null hypothesis lead to not rejecting H_0 : we have found no evidence that the underlying distribution of the length of forearms differs significantly from a normal distribution, at significance level 0.05.

We could also use these simulation results to calculate the P-value of our two tests. According to Equation (1.6), the P-value of the two tests are given by

$$p_1 = \sup_{\mu \in \mathbb{R}, \sigma > 0} P_{(\mu, \sigma^2)}(T_1(X) > -1.0532) = P_{(0,1)}(T_1(X) > -1.0532)$$

and

$$p_2 = P_{(0,1)}(T_2(X) > 0.1334).$$

We have calculated $T_1(X)$ and $T_2(X)$ 10000 times, so we only have to count how many times these values were bigger than the values found for the forearm data. We found

$$p_1 = 0.8936 \quad \text{and} \quad p_2 = 0.3435.$$

We see that the test based on T_2 has a much lower P-value than the test based on T_1 . Apparently, T_2 is more sensitive to the kind of “deviation from normal” we see in the forearm data than T_1 ; see also the next paragraph on simulation results.

Some simulation results

We have done a simulation study to compare the two tests for normality defined in the previous paragraph, T_1 and T_2 . Before we mention how we did this, we would like to introduce two other tests for normality. The first one was (in a way) also introduced in [3]: a Kolomogorov-Smirnoff type of test, or KS-test. This KS-test is based on the following test statistic: we calculate $\hat{\mu} = \bar{x}_n$ and $\hat{\sigma}^2 = s_n^2$ and define

$$T_3 = \sup_{t \in \mathbb{R}} |F_n(t) - \Phi\left(\frac{t - \hat{\mu}}{\hat{\sigma}}\right)|.$$

Here, F_n denotes the empirical distribution function of the data $x_1, \dots, x_n$, so

$$F_n(t) = \frac{\#\{i : x_i \leq t\}}{n} = \frac{1}{n} \sum_{i=1}^n 1_{(-\infty, x_i]}(t)$$

and Φ is the distribution function of the standard normal distribution. The reason that we also look at this test is that KS-type of tests are well known and frequently used.

The second test we would like to introduce is the χ^2 -test (pronounce: chi-square test). This test is the most widely used model test, that is, a test to see whether we can reject some given parametric model for our data (for example, the normal distributions). The main reason for its popularity is its simplicity: we split up the range of the data in k pieces, usually between 3 and 5, but it

depends on how much data we have. Let us call these pieces $A_1, \dots, A_k$. These pieces are usually data-dependent, for example, when testing for normality, we have chosen 5 intervals such that each interval has a probability of 20% under the normal distribution with parameters $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2 = S_n^2$. Note that data-dependent means that if we had a different data set, we could use the same *method* to find the pieces $A_1, \dots, A_k$, but we will probably find different realizations of $A_1, \dots, A_k$. Then we define

$$N_i(X) = \#\{i : X_i \in A_i\},$$

the number of data elements in A_i . Suppose we want to test whether $X_1, \dots, X_n$ is normally distributed. Then we define

$$E_i = E_{(\hat{\mu}, \hat{\sigma}^2)}[N_i(X)] \quad \text{and} \quad O_i = N_i(x),$$

so E_i is the expected number of data points in A_i under the estimated normal distribution, whereas O_i is the actual number of points in A_i of the data $x_1, \dots, x_n$. The χ^2 -test statistic is now defined as

$$T_4 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}.$$

The reason that this is called a χ^2 -test, is that under the null hypothesis, the distribution of T_4 can be approximated by a so-called χ^2 -distribution, but we will not go in to this. We feel it is better to find the critical value for T_4 in the same way we found the critical values for T_1 and T_2 . At this point we would like to mention that if $X_1, \dots, X_n$ are $N(\mu, \sigma^2)$ distributed with unknown μ and σ^2 (this is the null hypothesis), then $T_3(X)$ and $T_4(X)$ have a distribution that is independent of μ and σ^2 (compare with Proposition 3.2). This means that we can find the two critical values $c_{3,0.05}$ and $c_{4,0.05}$ in the following way: we simulated 10000 samples of size 140 (the same size as the forearm data) from a $N(0, 1)$ distribution and calculated $T_3(X)$ and $T_4(X)$ for each of these samples. Then we took the 95th percentile of these 10000 realizations of the test statistics T_3 and T_4 . This gave us

$$c_{3,0.05} = 0.0757 \quad \text{and} \quad c_{4,0.05} = 6.9286.$$

For the forearm data we also calculated the two test statistics:

$$T_3 = 0.0485 \quad \text{and} \quad T_4 = 2.2857.$$

We can conclude that based on these two tests, we cannot reject the null hypothesis that the forearm data is a realization from a normal distribution.

Quick exercise 3.4 We know that for the forearm data, we have

$$\hat{\mu} = \bar{x}_n = 18.80 \quad \text{and} \quad \hat{\sigma}^2 = s_n^2 = 1.256.$$

Give the intervals $A_1, \dots, A_5$ used in the χ^2 -test.

As we have seen in Chapter 1, a good way to compare different tests is to look at the power of the tests. The problem is that the power has to be calculated in some fixed alternative distribution, and our set of alternative distributions is very big: basically any continuous distribution we can think of! We have chosen three alternatives, that differ from a normal distribution in three different ways. In Figure 3.5 we show the densities of the alternatives, together with the best fitting normal distribution.

The first alternative is a smooth, bimodal distribution (that is, the density of the distribution has two peaks), that is a combination of two normal distributions. It has a density given by

$$f_1(x) = 0.8\phi(x) + 0.2\phi\left(\frac{x-2}{\sqrt{0.5}}\right)/\sqrt{0.5}.$$

Remember that ϕ is defined as the density of the standard normal distribution. This alternative differs only locally from a normal distribution, and in a smooth way.

Quick exercise 3.5 Show that f_1 is indeed a density.

The second alternative is a uniform distribution on $[-1, 1]$. This is symmetric and unimodal (that is, it has one peak), but it differs from the normal distribution in a non-smooth way. Also, the tail behavior is quite different, since the uniform distribution has a bounded support.

The third alternative is a $t(5)$ -distribution. The difference with a normal distribution lies mainly in the tails: a t -distribution has a much higher probability of becoming quite big.

We have calculated the power of our four tests in these three alternatives in the following way: we chose a sample size equal to 140, the same size as the forearm data (this is of course not essential). We simulated a sample $X_1, \dots, X_{140}$ from each of the three alternative distributions, calculated the four test statistics $T_1, \dots, T_4$ for this sample and compared it to their respective critical values $c_{1,0.05}, \dots, c_{4,0.05}$ at significance level 5%. We repeat this procedure 10000 times, and then we just count the number of times a test statistic is bigger than its corresponding critical value (that is, given such a sample, the normality assumption can be rejected at significance level 5%, if we use that test statistic) and divide this by 10000. This is how we got the results in Table 3.2.

Table 3.2: Powers of the four test at the three alternatives ($n = 140$).

	Bimodal	Uniform	$t(5)$
Kernel density: likelihood ratio (T_1)	0.1304	0.8371	0.1915
Kernel density: L_1 -distance (T_2)	0.2594	0.9777	0.0930
KS-type test (T_3)	0.2588	0.7863	0.1678
χ^2 test (T_4)	0.2655	0.7375	0.1433

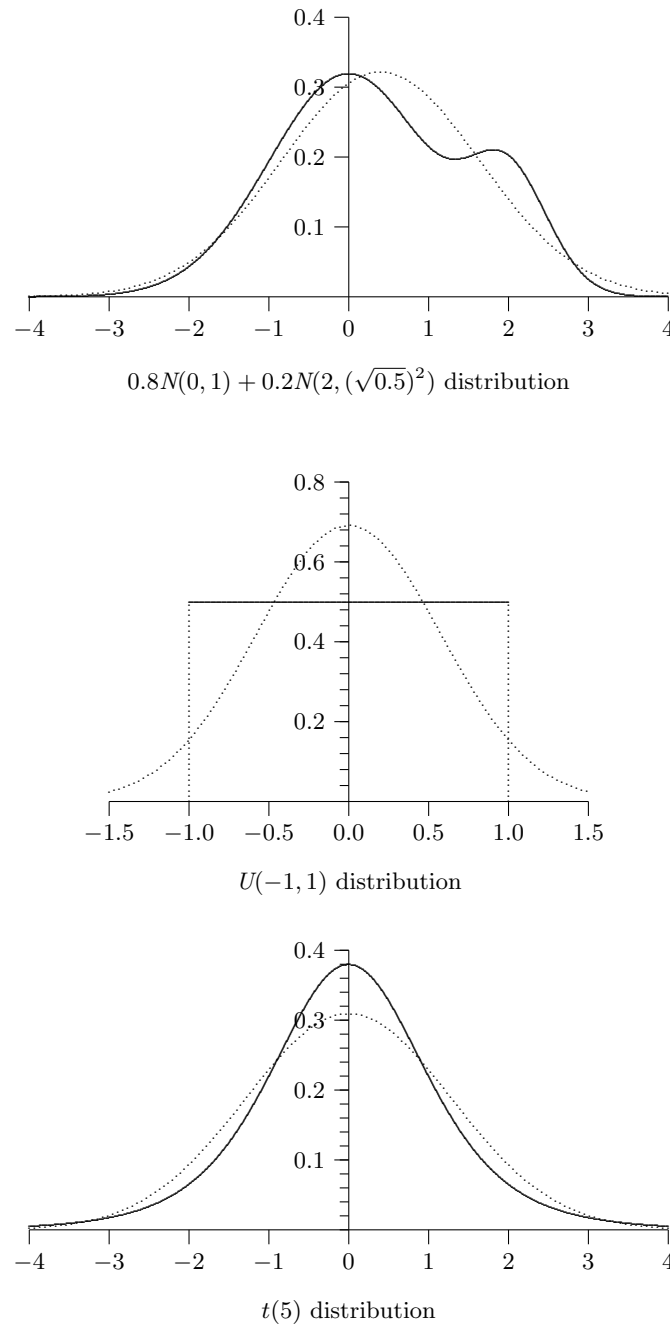


Figure 3.5: Three alternatives to a normal distribution.

The table clearly shows that different tests behave differently for different alternatives. The likelihood ratio test behaves well for the $t(5)$ distribution and quite poorly for the bimodal distribution. Apparently the test is sensitive for

changes in the tail of the distribution. The L_1 -distance based test behaves very well for the uniform alternative and poorly for the $t(5)$ distribution. In fact, it seems that the kernel density tests show a somewhat extreme behavior, compared to the KS-test and the χ^2 -test, which are more average.

How could we use these results? We cannot look at our data and then determine which test has the most power. This would be a data-based choice of test and that will influence the significance level of the test (we would then reject a normal distribution with more than 5% probability). However, if we would have some information about the underlying distribution beforehand (for example, we have reasons to suspect heavy tails), then it is a good idea to choose a test which is sensitive to the kind of deviation from the normal distribution we expect (in the case of heavy tails, one might prefer the likelihood ratio test).

3.3 Monotone densities

Consider the following data on the number of days between two mining disasters; see Table 3.3. The data are taken from [6] and are originally from [7].

Table 3.3: Days between two coal mining disasters.

157	123	2	124	12	4	10	216	80	12
33	66	232	826	40	12	29	190	97	65
186	23	92	197	431	16	154	95	25	19
78	202	36	110	276	16	88	225	53	17
538	187	34	101	41	139	42	1	250	80
3	324	56	31	96	70	41	93	24	91
143	16	27	144	45	6	208	29	112	43
193	134	420	95	125	34	127	218	2	0
378	36	15	31	215	11	137	4	15	72
96	124	50	120	203	176	55	93	59	315
59	61	1	13	189	345	20	81	286	114
108	188	233	28	22	61	78	99	326	275
54	217	113	32	388	151	361	312	354	307
275	78	17	1205	644	467	871	48	123	456
498	49	131	182	255	194	224	566	462	228
806	517	1643	54	326	1312	348	745	217	120
275	20	66	292	4	368	307	336	19	329
330	312	536	145	75	364	37	19	156	47
129	1630	29	217	7	18	1358	2366	952	632

A histogram of this data can be seen in Figure 3.6.

If the times of coal mining disasters are modelled by a Poisson process, the time between two disasters will have an exponential distribution, so it makes sense to test if the data is a realization of a sample from an exponential distribution. However, if we relax the assumption on the times of coal mining disasters (it should be a *renewal* process, but we will not go into the precise assumptions),

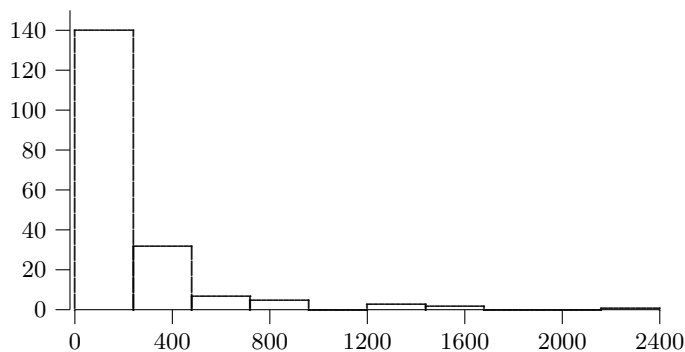


Figure 3.6: Histogram of the days between coal mining disasters.

we can show that the time between two disasters will have a *monotone density* on $[0, \infty)$, that is, the times in Table 3.3 can be modelled as a realization of a sample from a density f which is non-increasing:

$$t \leq s \Rightarrow f(t) \geq f(s).$$

Of course, the exponential density is an example of a monotone density on $[0, \infty)$.

If we take as our model

$$\Theta = \{f : [0, \infty) \rightarrow [0, \infty) : f \text{ monotone}\},$$

we might ask ourselves if we can construct the maximum likelihood estimator of f based on the data $x_1, \dots, x_n$. In the case of C^2 -densities, we have seen that we can make the likelihood of the data as big as we like, but in this case we cannot increase the likelihood of some data point x_i , without increasing f on the whole interval $[0, x_i]$. Since f has to be a density, this puts a limitation on the likelihood, suggesting that the maximum likelihood estimator can actually be found.

Quick exercise 3.6 Show that the likelihood of $x_1, \dots, x_n > 0$ is bounded by $1/(x_1 \cdots x_n)$ for all monotone densities f , so

$$l(x) = f(x_1) \cdots f(x_n) \leq \frac{1}{x_1 \cdots x_n}.$$

The maximum likelihood estimator of a monotone density

In this paragraph we will derive the maximum likelihood estimator for a monotone density. We consider a sample $X_1, \dots, X_n$ from a monotone density f on $[0, \infty)$. As usual, we define the empirical distribution function F_n by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{[X_i, \infty)}(t).$$

We call a function g *concave* on $[0, \infty)$ if

$$g(\lambda t + (1 - \lambda)s) \geq \lambda g(t) + (1 - \lambda)g(s) \quad \forall t, s \geq 0 \quad \forall 0 \leq \lambda \leq 1.$$

An example of a concave function and its defining property is given in Figure 3.7.

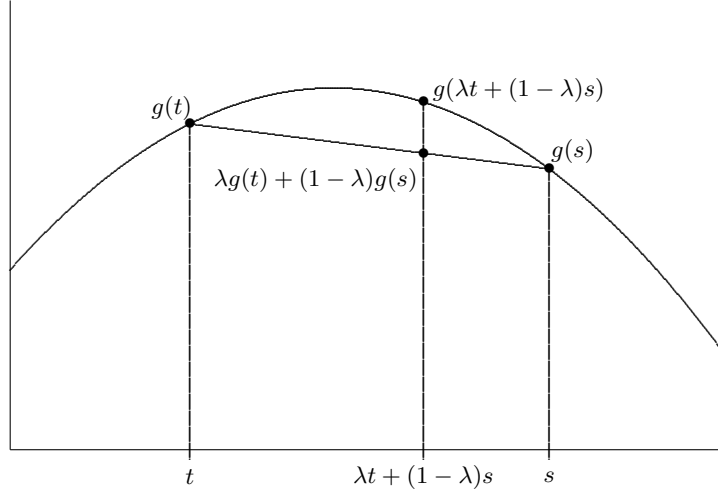


Figure 3.7: Concave function.

A property of concave functions is that if $\mathcal{G}$ is a family of positive concave functions, then

$$\hat{g}(t) = \inf_{g \in \mathcal{G}} g(t)$$

is also a (positive) concave function.

Quick exercise 3.7 Show that $\hat{g}$ is indeed a concave function.

This means that for any bounded positive function H we can define $\hat{H}$ as the *least concave majorant* of H by

$$\hat{H}(t) = \inf\{g(t) : g \geq H \text{ and } g \text{ is concave}\}.$$

Note that $\hat{H}$ is concave and $\hat{H} \geq H$. An example can be seen in Figure 3.8. The reason why concave functions play an important role when studying monotone densities is the following proposition:

Proposition 3.3 *Let f be the monotone density on $[0, \infty)$ of a random variable X . Then the distribution function F of X , which satisfies*

$$F(t) = \int_0^t f(s) ds,$$

is a concave function.

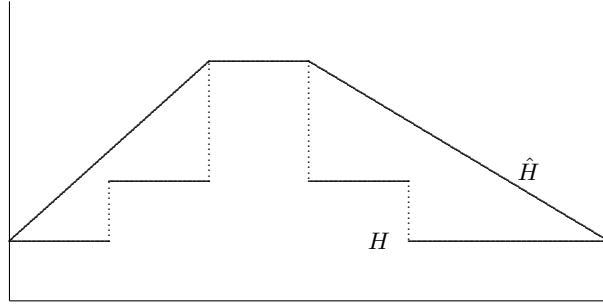


Figure 3.8: Least concave majorant.

Proof: We will show this in case f is continuous, but it is always true. Suppose $t < s$. Then

$$G(x) = \int_t^x f(u)du - \frac{F(s) - F(t)}{s - t}(x - t) \quad (t < x < s)$$

is a continuously differentiable function with $G(t) = G(s) = 0$ and

$$G'(x) = f(x) - \frac{F(s) - F(t)}{s - t}.$$

Since f is non-increasing, so is G' . Now suppose there exists $y \in (t, s)$ such that $G(y) < 0$. Then the mean value theorem states that there exists $z_1 \in (t, y)$ and $z_2 \in (y, s)$ such that

$$G'(z_1) = \frac{G(y) - G(t)}{y - t} < 0 \quad \text{and} \quad G'(z_2) = \frac{G(s) - G(y)}{s - y} > 0,$$

contradicting the fact that G' is non-increasing. Therefore, $G \geq 0$ on $[t, s]$, which implies that for $x = \lambda t + (1 - \lambda)s$ ($0 \leq \lambda \leq 1$) we have

$$F(x) = F(t) + \int_t^x f(u)du \geq F(t) + \frac{F(s) - F(t)}{s - t}(x - t) = \lambda F(t) + (1 - \lambda)F(s).$$

The general case may be proved by noting that a monotone density is the infimum of continuous monotone densities; the corresponding distribution function is therefore the infimum of concave distribution functions, and hence concave. $\square$

We are now ready to prove the following theorem, originally proved by Grenander in [4]:

Theorem 3.4 *Let $X_1, \dots, X_n$ be a sample from a monotone density on $[0, \infty)$. Denote by F_n the empirical distribution function and define $\hat{F}_n$ as the least concave majorant of F_n . The maximum likelihood estimator $\hat{f}_n$ over the class of monotone densities, also called the Grenander estimator, is the left-derivative of $\hat{F}_n$, that is,*

$$\hat{f}_n(t) = \lim_{h \downarrow 0} \frac{\hat{F}_n(t) - \hat{F}_n(t - h)}{h} \quad (\forall t > 0).$$

Equivalently, $\hat{f}_n$ is the (unique) left-continuous monotone density such that for all $t \geq 0$

$$\hat{F}_n(t) = \int_0^t \hat{f}_n(s) ds.$$

Exercise 3.4 explains why the left-derivative of $\hat{F}_n$ always exists. Before we prove this theorem, we give a small example of $\hat{F}_n$ and $\hat{f}_n$. Suppose the data is given by

$$\{0.5, 0.75, 0.8, 1.7, 1.8, 2.2, 2.3, 3\}.$$

Figure 3.9 shows $\hat{F}_n$ and F_n and Figure 3.10 shows $\hat{f}_n$ for this example.

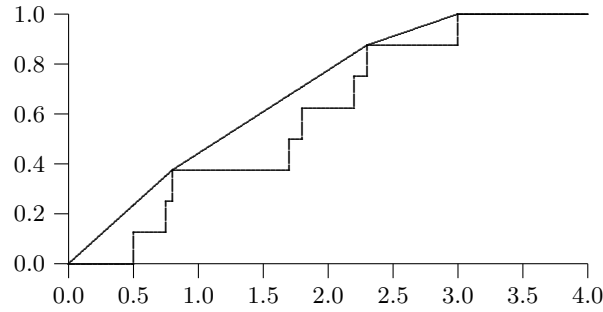


Figure 3.9: Estimated and empirical distribution function.

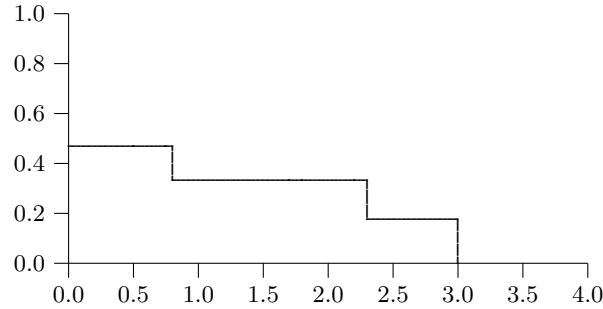


Figure 3.10: Estimated monotone density function.

Proof of Theorem 3.4: We follow the proof of [5]. We want to maximize the (averaged) log-likelihood

$$f \mapsto \frac{1}{n} \sum_{i=1}^n \log f(X_i),$$

over all monotone densities f . Consider the order statistics $X_{(1)} \leq \dots \leq X_{(n)}$. Note that if f is a monotone density that is not constant on the interval

$(X_{(i)}, X_{(i+1)}]$, then the monotone density $\tilde{f}$, defined by

$$\tilde{f}(t) = \begin{cases} f(t) + \frac{1}{X_{(i)}} \int_{X_{(i)}}^{X_{(i+1)}} f(s) - f(X_{(i+1)}) ds & \text{if } t \in [0, X_{(i)}] \\ f(X_{(i+1)}) & \text{if } t \in (X_{(i)}, X_{(i+1)}) \\ f(t) & \text{if } t \in [X_{(i+1)}, \infty) \end{cases}$$

has a bigger log-likelihood than f , that is,

$$\frac{1}{n} \sum_{i=1}^n \log \tilde{f}(X_i) \geq \frac{1}{n} \sum_{i=1}^n \log f(X_i),$$

since for each $1 \leq j \leq n$ we have $\tilde{f}(X_j) \geq f(X_j)$. In the same way we can make the likelihood bigger if f is not 0 on $(X_{(n)}, \infty)$. Finally, if $f(X_{(n)}) = 0$, then the log-likelihood will be $-\infty$, so we may also assume that $f(X_{(n)}) > 0$. This means that when we are maximizing the log-likelihood over all monotone densities, we only need to consider non-increasing densities that are constant on every interval $(X_{(i)}, X_{(i+1)}]$, greater than 0 in $X_{(n)}$ and 0 on $(X_{(n)}, \infty)$. Exercise 3.4 shows that $\hat{f}_n$ indeed satisfies these three conditions. These densities are all elements of the following class of non-increasing functions on $[0, X_{(n)}]$:

$$\mathcal{F} = \left\{ \sum_{i=1}^n a_i 1_{[0, X_{(i)}]}(t) : a_i \geq 0 \ (1 \leq i \leq n-1) \text{ and } a_n \in \mathbb{R} \right\}.$$

We have that for all $g \in \mathcal{F}$,

$$\int g(t) \hat{f}_n(t) dt \geq \frac{1}{n} \sum_{i=1}^n g(X_i). \quad (3.15)$$

Note that if f is a monotone density such that $f \in \mathcal{F}$ and $f(X_{(n)}) > 0$, then also $\log f \in \mathcal{F}$. To show (3.15), we remark that $\hat{F}_n \geq F_n$ and that $\hat{F}_n(t) = 1$ if $t \geq X_{(n)}$. This last statement follows from the fact that $1 \geq F_n$ and 1 is a concave function, so $1 \geq \hat{F}_n$; if $t \geq X_{(n)}$, then $\hat{F}_n(t) \geq F_n(t) = 1$, so $\hat{F}_n(t) = 1$. Then we have for $a_i \geq 0$

$$\int a_i 1_{[0, X_{(i)}]}(t) \hat{f}_n(t) dt = a_i \hat{F}_n(X_{(i)}) \geq a_i F_n(X_{(i)}) = \frac{1}{n} \sum_{j=1}^n a_i 1_{[0, X_{(i)}]}(X_j)$$

and for $a_n \in \mathbb{R}$

$$\int a_n 1_{[0, X_{(n)}]}(t) \hat{f}_n(t) dt = a_n \hat{F}_n(X_{(n)}) = a_n = \frac{1}{n} \sum_{j=1}^n a_n 1_{[0, X_{(n)}]}(X_j),$$

which implies Equation (3.15).

The next step is to show that

$$\int \log(\hat{f}_n(t)) \hat{f}_n(t) dt = \frac{1}{n} \sum_{i=1}^n \log \hat{f}_n(X_i). \quad (3.16)$$

Note that $\hat{f}_n(t) = 0$ for $t > X_{(n)}$ and we define $0 \cdot \log(0) = 0$, so we actually integrate over $[0, X_{(n)}]$. To show (3.16), we define m to be the number of sample points where F_n and $\hat{F}_n$ are equal. Furthermore, we define an increasing map $\pi : \{1, \dots, m\} \rightarrow \{1, \dots, n\}$ such that

$$\hat{F}_n(X_{(\pi(k))}) = F_n(X_{(\pi(k))}) \quad (\forall 1 \leq k \leq m).$$

For the example before this proof we would get $m = 3$ and

$$\pi(1) = 3, \pi(2) = 7 \text{ and } \pi(3) = 8.$$

Note that we always have $\pi(m) = n$ (and therefore $m \geq 1$). As can be seen from Figure 3.9 and Figure 3.10, we also have that $\hat{f}_n$ is constant on every interval $(X_{(\pi(k))}, X_{(\pi(k+1))}]$ ($k = 1, \dots, m-1$) and on the interval $[0, X_{(\pi(1))}]$. Therefore, if we define $X_{(0)} = 0$ and $\pi(0) = 0$,

$$\begin{aligned} \int_0^{X_{(n)}} \log(\hat{f}_n(t)) \hat{f}_n(t) dt &= \sum_{k=1}^m \log(\hat{f}_n(X_{(\pi(k))})) (\hat{F}_n(X_{(\pi(k))}) - \hat{F}_n(X_{(\pi(k-1))})) \\ &= \sum_{k=1}^m \log(\hat{f}_n(X_{(\pi(k))})) (F_n(X_{(\pi(k))}) - F_n(X_{(\pi(k-1))})) \\ &= \sum_{k=1}^m \sum_{i=\pi(k-1)+1}^{\pi(k)} \log(\hat{f}_n(X_{(i)})) \frac{1}{n} \\ &= \frac{1}{n} \sum_{i=1}^n \log(\hat{f}_n(X_{(i)})). \end{aligned}$$

The last step is to prove that for any density f on $[0, \infty)$

$$\int \log(\hat{f}_n(t)) \hat{f}_n(t) dt \geq \int \log(f(t)) \hat{f}_n(t) dt. \quad (3.17)$$

This follows from

$$\begin{aligned} \int \log\left(\frac{f(t)}{\hat{f}_n(t)}\right) \hat{f}_n(t) dt &= \int \log\left(1 + \frac{f(t) - \hat{f}_n(t)}{\hat{f}_n(t)}\right) \hat{f}_n(t) dt \\ &\leq \int (f(t) - \hat{f}_n(t)) dt \\ &= 0. \end{aligned}$$

Here we used that $\log(1+t) \leq t$ for $t > -1$. Now take any density $f \in \mathcal{F}$ such that $f(X_{(n)}) > 0$. Then $\log f \in \mathcal{F}$ and we can use (3.16), (3.17) and (3.15) respectively to see that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \log \hat{f}_n(X_i) &= \int \log(\hat{f}_n(t)) \hat{f}_n(t) dt \\ &\geq \int \log(f(t)) \hat{f}_n(t) dt \\ &\geq \frac{1}{n} \sum_{i=1}^n \log f(X_i). \end{aligned}$$

This proves that $\hat{f}_n$ is indeed the maximum likelihood estimator. $\square$

To illustrate this theorem, we have calculated $\hat{F}_n$ and $\hat{f}_n$ for the coal mining disasters data. The results can be seen in Figure 3.11 and Figure 3.12.

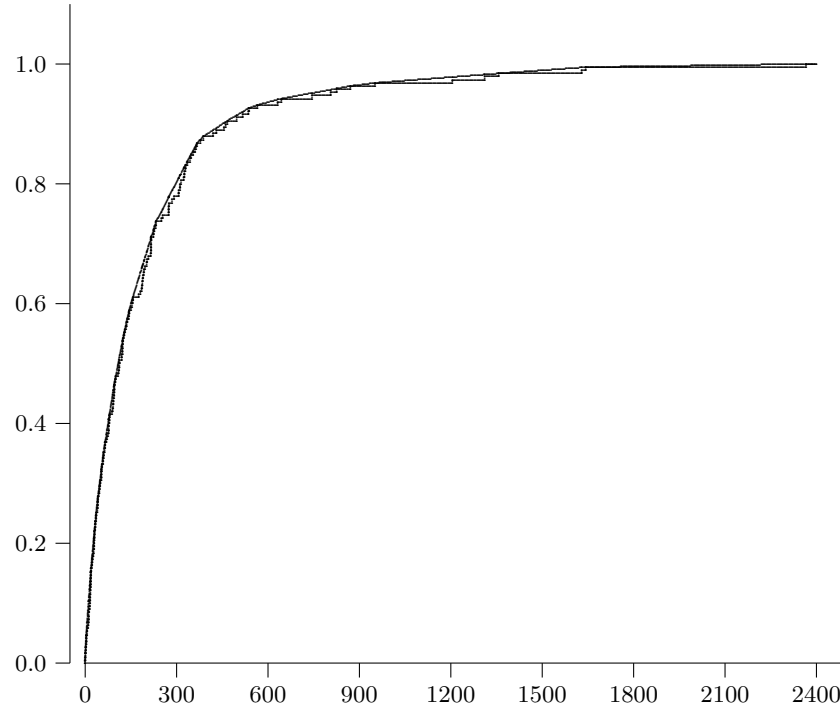


Figure 3.11: Estimated and empirical distribution function for coal mining disasters data.

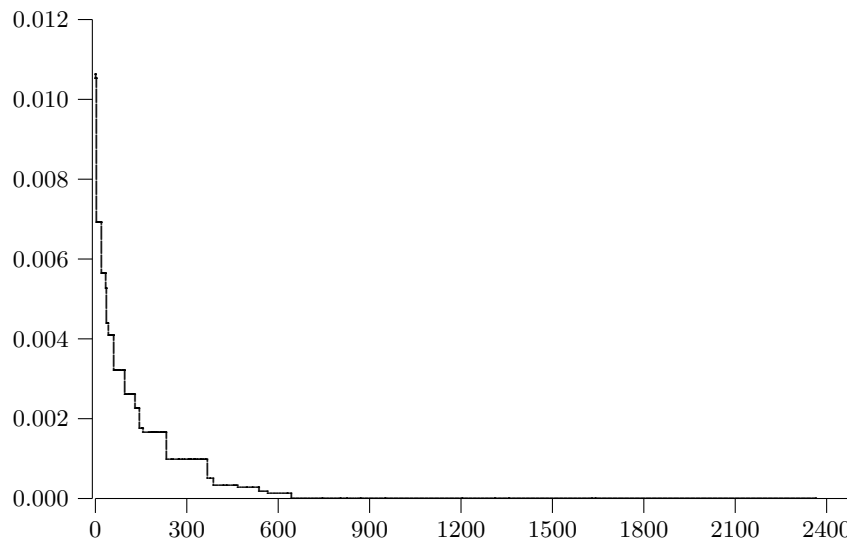


Figure 3.12: Estimated monotone density for coal mining disasters data.

Testing for exponentiality

We now have calculated the maximum likelihood estimator in the class of monotone densities. We can use this to perform likelihood ratio tests within this class. As an example, we will test the null hypothesis that the underlying density f of the coal mining disasters data is exponential against the alternative that f is a monotone density (but not exponential). So our null hypothesis is

$$H_0 = \{f(t) = \lambda e^{-\lambda t} : \lambda > 0, t \geq 0\},$$

and our alternative hypothesis is

$$H_1 = \Theta \setminus H_0 = \{f : [0, \infty) \rightarrow [0, \infty) : f \text{ monotone}\} \setminus H_0.$$

The likelihood ratio test statistic of a sample $X_1, \dots, X_n$ is then defined by

$$L(X) = \frac{\sup_{f \in H_1} f(X_1) \cdots f(X_n)}{\sup_{f \in H_0} f(X_1) \cdots f(X_n)} = \frac{\hat{f}_n(X_1) \cdots \hat{f}_n(X_n)}{(\bar{X}_n)^{-n} e^{-n}}.$$

Here we used that the maximum likelihood estimator for the parameter λ of the exponential distribution is given by

$$\hat{\lambda} = \frac{1}{\bar{X}_n}.$$

Quick exercise 3.8 Check this last statement.

We worked with the log-likelihood ratio test statistic, so

$$T(X) = \log L(X) = \sum_{i=1}^n \log \hat{f}_n(X_i) + n(1 + \log(\bar{X}_n)).$$

If we calculate this statistic for the coal mining disasters data, we get

$$T = 30.3.$$

To see whether this is a big value or not, we have to find c_α such that

$$\sup_{f \in H_0} P_f(T(X) > c_\alpha) = \alpha.$$

We will check in Exercise 3.4 that if we multiply our sample with $a > 0$, we get

$$\hat{f}_n^{(a)}(t) = \frac{1}{a} \hat{f}_n\left(\frac{t}{a}\right) \quad (\forall t > 0).$$

Here $\hat{f}_n^{(a)}$ denotes the maximum likelihood estimator based on the sample $aX_1, \dots, aX_n$. This implies that the test statistic T does not change if we multiply each sample point by a , and therefore we conclude that the distribution of $T(X)$, when $X_i \sim \text{Exp}(\lambda)$, is independent of λ . In other words,

$$\sup_{f \in H_0} P_f(T(X) > c_\alpha) = P_{\text{Exp}(1)}(T(X) > c_\alpha).$$

Thus to find c_α , we only need to simulate samples from an $Exp(1)$ distribution. We simulated 10000 samples of size 190 (the size of the coal mining disasters data set) from an $Exp(1)$ distribution, and calculated T for each of these samples. From those 10000 realizations of T we calculated the 95th percentile and found

$$c_\alpha = 19.46.$$

This means that the value $T = 30.3$ belonging to the coal mining disasters data is big enough to reject the hypothesis of exponentiality in favor of some other monotone density. In fact, the 10000 realizations we found of T from the $Exp(1)$ distribution were all smaller than 30.3, suggesting that the P-value of our test is smaller than 0.0005. See also Exercise 3.2.

Some simulation results

We have compared the likelihood ratio test to two other tests for exponentiality, namely a KS-type of test and a χ^2 -test, both described in the previous section. We found that the KS-type test rejected the hypothesis of exponentiality for the coal mining disaster data with a P-value of 0.0024, and that the χ^2 -test rejected this hypothesis with a P-value of 0.0192.

To compare the power of the three tests, we again had to choose (in this case, monotone) alternatives to the exponential distribution. The first two alternatives, together with the best fitting exponential (that is, with the same mean), are shown in Figure 3.13. The third alternative was the estimated monotone density for the coal mining disasters data, see Figure 3.12.

We took 10000 samples of size $n = 190$ (as in the coal example) from all three distributions, and performed the three tests for all samples. This resulted in the Table 3.4, containing the power of the three tests, at the three different alternatives.

Table 3.4: Powers of the three tests at the three alternatives ($n = 190$).

	$16(x+1)^{-15}$	$\frac{9}{100} (1 - \frac{x}{100})^8$	coal estimator
Grenander	0.1108	0.0527	0.9989
KS-type test	0.0916	0.1447	0.9528
χ^2 -test	0.0591	0.0766	0.8113

The first alternative and the estimated density for the coal mining disasters data both have heavy tails compared with an exponential distribution, that is, they have a much greater probability to be far away from, for example, their mean. Apparently, the maximum likelihood estimator is sensitive for this, since the likelihood ratio test has higher power than the KS-type test and the χ^2 -test for these two alternatives. On the other hand, the second alternative has a very thin tail compared to the exponential (it is actually 0 when $x > 100$). For this alternative the likelihood ratio test performs rather poorly.

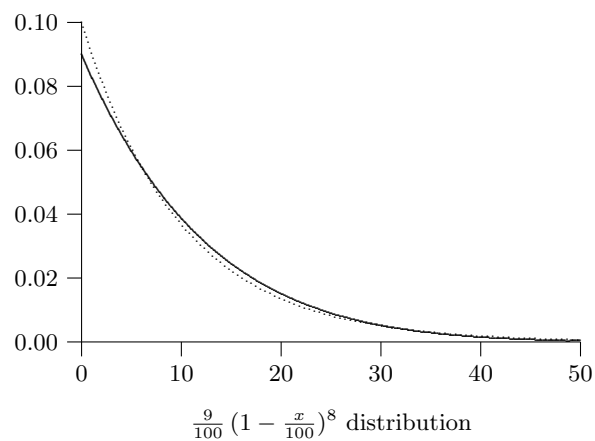
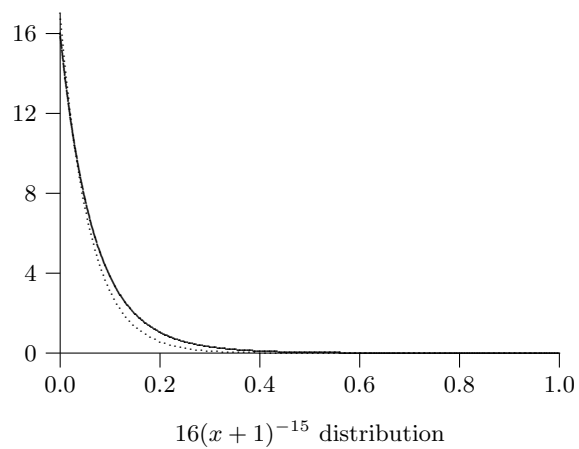


Figure 3.13: Two alternatives to an exponential distribution.

3.4 Solutions to the quick exercises

3.1 Define

$$k(x) = 1_{[-0.5, 0.5]}(x)(1 + \cos(2\pi x))$$

and check that this is a density. Define

$$f_n(x) = \frac{n}{3}k(n(x-1)) + \frac{n}{3}k(n(x-3)) + \frac{n}{3}k(n(x-4)).$$

This is a density for all $n \geq 1$ and

$$f_n(1) \cdot f_n(3) \cdot f_n(4) = \frac{n^3}{9}.$$

3.2 Since K is a positive function, $\hat{f}_{n,h}$ will also be positive. Furthermore,

$$\begin{aligned}\int \hat{f}_{n,h}(t)dt &= \frac{1}{nh} \sum_{i=1}^n \int K\left(\frac{t-x_i}{h}\right) dt \\ &= \frac{1}{n} \sum_{i=1}^n \int K(s)ds \\ &= 1.\end{aligned}$$

Here we used the substitution $s = (t - x_i)/h$.

3.3 We only have to rewrite Equation (3.3):

$$\begin{aligned}\text{Var}(\hat{f}_{n,h}(t)) &= \frac{1}{nh} \int K(s)^2 f(t-hs)ds - \frac{1}{n} \left(\int K(s)f(t-hs)ds \right)^2 \\ &= \frac{1}{nh} \left(f(t) \int K(s)^2 ds + \int K(s)^2 (f(t-hs) - f(t))ds + \right. \\ &\quad \left. + h \left(\int K(s)f(t-hs)ds \right)^2 \right).\end{aligned}$$

3.4 The intervals are chosen such that each interval has a 20% probability under the $N(\hat{\mu}, \hat{\sigma}^2)$ distribution. We use Table 3.5 to see that

$$P(N(0, 1) > 0.84) \approx 0.2 \quad \text{and} \quad P(N(0, 1) > 0.25) \approx 0.4.$$

This means that

$$\begin{aligned}A_1 &= (-\infty, \hat{\mu} - 0.84\hat{\sigma}] = (-\infty, 17.86] \\ A_2 &= (\hat{\mu} - 0.84\hat{\sigma}, \hat{\mu} - 0.25\hat{\sigma}] = (17.86, 18.52] \\ A_3 &= (\hat{\mu} - 0.25\hat{\sigma}, \hat{\mu} + 0.25\hat{\sigma}] = (18.52, 19.08] \\ A_4 &= (\hat{\mu} + 0.25\hat{\sigma}, \hat{\mu} + 0.84\hat{\sigma}] = (19.08, 19.74] \\ A_5 &= (\hat{\mu} + 0.84\hat{\sigma}, \infty) = (19.74, \infty).\end{aligned}$$

3.5

$$\begin{aligned}\int f_1(x)dx &= 0.8 \int \phi(x)dx + 0.2 \int \frac{1}{\sqrt{0.5}} \phi\left(\frac{x-2}{\sqrt{0.5}}\right) dx \\ &= 0.8 + 0.2 \int \phi(z)dz \\ &= 1.\end{aligned}$$

3.6 Since f non-increasing, we get that $f(t) \geq f(s)$ if $s \geq t$. This means that

$$1 \geq \int_0^s f(t)dt \geq sf(s),$$

so $f(s) \leq 1/s$, for all $s \in \mathbb{R}$.

3.7 For all $g \in \mathcal{G}$ we have:

$$g(\lambda t + (1 - \lambda)s) \geq \lambda g(t) + (1 - \lambda)g(s) \geq \lambda \hat{g}(t) + (1 - \lambda)\hat{g}(s).$$

Taking the infimum over $\mathcal{G}$ gives

$$\hat{g}(\lambda t + (1 - \lambda)s) \geq \lambda \hat{g}(t) + (1 - \lambda)\hat{g}(s).$$

3.8 The log-likelihood is given by

$$l(\lambda) = \sum_{i=1}^n (\log(\lambda) - \lambda X_i) = n \log(\lambda) - n \bar{X}_n \lambda.$$

Taking the derivative immediately gives a maximum of l at

$$\hat{\lambda} = \frac{1}{\bar{X}_n}.$$

3.5 Exercises

3.1 Consider the data set

$$\{6.07, 6.92, 7.78, 7.83, 7.94, 8.43, 9.05, 9.29, 9.57, 9.80\}.$$

We consider this to be a realization of a random sample $X_1, \dots, X_{10}$.

- a. Suppose that $X_i \sim U(a, b)$. Give the maximum likelihood estimates of a and b for the given data set.
- b. We want to test the uniformity of data, that is, if indeed $X_i \sim U(a, b)$, using a χ^2 -test. We split up the real line in three intervals A_1, A_2, A_3 such that under the estimated uniform distribution, $P(X \in A_j) = 1/3$, for $j = 1, 2, 3$. Give the χ^2 test statistic T for our data set.
- c. Show that if $X_i \sim U(a, b)$, the distribution of $T(X)$ does not depend on a or b .
- d. We simulated 10000 samples of size 10 from the $U(0, 1)$ distribution, and calculated T for each sample. After we ordered them, we found that

$$T(490) = T(491) = \dots = T(510) = 0.2$$

and that

$$T(9490) = T(9491) = \dots = T(9510) = 5.6$$

Can you explain how we could have found the same value for T several times, even though we sampled from a continuous distribution?

- e. Test the hypothesis that our data is a realization of a uniform distribution with a significance of 5%.

3.2 Consider Table 3.2.

- a. Based on how the numbers in the table were found, give a 95% confidence interval of the *true* power of the test T_1 at the second alternative, that is, the uniform distribution on $[-1, 1]$. You may use the normal approximation.
- b. In the case of the coal mining disaster data, when testing for exponentiality, of 10000 realizations of $T(X)$, where $X_i \sim \text{Exp}(1)$, not one was found to be bigger than 30.3, the value for T when using the actual data. Give a 95% upper confidence interval for the true P-value of the test.

3.3 [Proof of Proposition 3.2]

- a. Proof that if $X_i \sim N(\mu, \sigma)$, the distribution of $T_2(X)$ does not depend on the parameters μ and σ , using the results of the proof for $T_1(X)$.
- b. Proof the same result for the KS-type test $T_3(X)$ and the χ^2 -test $T_4(X)$.

3.4 Let g be a concave function on $[0, \infty)$ and let $t > 0$. Define the function ϕ_t as

$$\phi_t : (0, t) \rightarrow \mathbb{R} : h \mapsto \frac{g(t) - g(t-h)}{h}$$

and for $h \in (0, t)$

$$\psi_{t,h} : (0, \infty) \rightarrow \mathbb{R} : s \mapsto \phi_t(h)(s-t) + g(t). \quad (3.18)$$

So $\psi_{t,h}$ is the linear function such that $\psi_{t,h}(t) = g(t)$ and $\psi_{t,h}(t-h) = g(t-h)$.

- a. Show that $\psi_{t,h} \leq g$ on $[t-h, t]$ and that $\psi_{t,h} \geq g$ on $[0, t-h] \cup [t, \infty)$.
- b. Show that

$$\phi_t(h) \geq \frac{g(s) - g(s-a)}{a} = \phi_s(a)$$

for $s \geq t$ and $0 < a \leq s-t+h$. Hint: make a picture!

- c. Show that the function

$$g'_-(t) = \lim_{h \downarrow 0} \frac{g(t) - g(t-h)}{h}$$

is well-defined for all $t > 0$ and that g'_- is non-increasing.

- d. Now take $g = \hat{F}_n$, the least concave majorant of the empirical distribution function F_n belonging to the data set $0 < x_1 \leq x_2 \leq \dots \leq x_n$. Show that $\hat{f}_n = g'_-$ is constant on every interval $(x_{i-1}, x_i]$ ($i = 1, \dots, n$), where $x_0 = 0$.
- e. Define $y_i = ax_i$ ($\forall 1 \leq i \leq n$), where $a > 0$. Define $F_n^{(y)}$ as the empirical distribution function belonging to the sample $y_1, \dots, y_n$, so

$$F_n^{(y)}(t) = \frac{1}{n} \sum_{i=1}^n 1_{[y_i, \infty)}(t).$$

Likewise, define $\hat{F}_n^{(y)}$ as the least concave majorant of $F_n^{(y)}$, and $\hat{f}_n^{(y)}$ as its left-derivative. Show that

$$F_n^{(y)}(t) = F_n(t/a)$$

and that

$$\hat{F}_n^{(y)}(t) = \hat{F}_n(t/a).$$

Hint: show that if H is a concave function such that $H \geq F_n^{(y)}$, then $\tilde{H}(t) = H(t/a)$ is a concave function such that $\tilde{H} \geq F_n$.

f. Show that

$$\hat{f}_n^{(y)}(t) = \frac{1}{a} \hat{f}_n(t/a).$$

Bibliography

- [1] Patrick Billingsley. *Probability and measure*. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.
- [2] O.L. Davies and P.L. Goldsmith. *Statistical Methods in Research and Production*. Oliver and Boyd, Edinburgh, 1972.
- [3] F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä, and Meester. L.E. *Kanstat*. College dictaat, Delft, 2003.
- [4] U. Grenander. On the theory of mortality measurement, part ii. *Skandinavisk Aktuarietidskrift*, 39:125–153, 1956.
- [5] P. Groeneboom and H.P. Lopuhaä. Isotonic estimators of monotone densities and distribution functions: basic facts. *Statistica Neerlandica*, 47(3):175–183, 1993.
- [6] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *Small Data Sets*. Chapman and Hall, London, 1994.
- [7] R.G. Jarrett. A note on the intervals between coal mining disasters. *Biometrika*, 66:191–193, 1979.
- [8] K. Pearson and A. Lee. On the laws of inheritance in man. i. inheritance of physical characters. *Biometrika*, 2:357–462, 1903.
- [9] B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall, London, 1986.
- [10] M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman & Hall, London, 1995.
- [11] E. J. Williams. *Regression analysis*. John Wiley & Sons Inc., New York, 1959.

Index

- X -matrix, 35
- χ^2 -test, 75
- abrasion loss
 - coefficients, 46
 - data, 33
- alternative hypothesis, 2, 6
- bandwidth, 63
- bias-variance trade-off, 67
- Coal mining diisasters
 - data, 79
- confidence region, 19
- confidence region for $\rho(\theta)$, 24
- corrected total sum of squares, 38
- data, 1, 5
- data space, 1, 5
- decision variable, 5
- dependent variable , 34
- design-matrix, 35
- equivalent test statistics, 6
- estimated MISE, 69
- experiment, 5
- Gauss-Markov conditions, 43
- Grenander estimator, 82
- heteroscedasticity, 35
- homoscedasticity, 35
- hypothesis testing, 1
- independent variable, 34
- intercept, 35
- Janka hardness example
 - data, 52
- kernel, 63
- least concave majorant, 81
- least squares cross-validation method,
 - 69
 - with ties, 70
- least squares estimator, 36
- leave-one-out estimator, 68
- Length of forearms
 - data, 62
- likelihood ratio test, 12
- linear model, 35, 36
- lower confidence region for $\rho(\theta)$, 26
- Mean Integrated Squared Error, 67
- Mean Squared Error, 66
- MISE, 67
- monotone density, 80
- monotone family of test functions,
 - 15
- MSE, 66
- Multiple R-squared, 38
- multivariate normal distribution, 41
- nonparametric models, 61
- normal equations, 36
- Normality assumption, 46
- null hypothesis, 1, 6
- P-value, 15
- power, 9
- R-squared, 38
- random sample, 5
- regression function, 34
- regression model, 34
- regression-matrix, 35
- regressor, 34
- residual, 36
- residual sum of squares, 36
- sample space, 5

significance level, 6
simple hypothesis, 10
standard error, 46
standard normal multivariate distribution, 41
statistical model, 1, 5

test function, 2, 6
test statistic, 6
total sum of squares, 38

unbiased test, 14
uniformly most powerful, 10
upper confidence region for $\rho(\theta)$, 25

Tables

Table 3.5: Right tail probabilities $1 - \Phi(a) = P(Z \geq a)$ for a $N(0,1)$ distributed random variable Z .

a	0	1	2	3	4	5	6	7	8	9
0.0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0.1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0.2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0.3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0.4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
0.5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0.6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
0.7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
0.8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
0.9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1.0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1.1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1.2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1.3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1.4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1.5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1.6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1.7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1.8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1.9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2.0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
2.1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2.2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2.3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2.4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
2.5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
2.6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
2.7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
2.8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
2.9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
3.0	0013	0013	0013	0012	0012	0011	0011	0011	0010	0010
3.1	0010	0009	0009	0009	0008	0008	0008	0008	0007	0007
3.2	0007	0007	0006	0006	0006	0006	0006	0005	0005	0005
3.3	0005	0005	0005	0004	0004	0004	0004	0004	0004	0003
3.4	0003	0003	0003	0003	0003	0003	0003	0003	0003	0002

Table 3.6: Right critical values $t_{m,p}$ of the t -distribution with m degrees of freedom corresponding to right tail probability p : $P(T_m \geq t_{m,p}) = p$. The last row in the table are right critical values of the $N(0, 1)$ distribution: $t_{\infty,p} = z_p$.

m	Right tail probability p							
	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Table 3.7: Right critical values $F_{0.05}(m, n)$ of the F -distribution with m and n degrees of freedom corresponding to right tail probability 0.05: $P(F(m, n) \geq F_{0.05}(m, n)) = 0.05$.

n	m								
	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
110	3.93	3.08	2.69	2.45	2.30	2.18	2.09	2.02	1.97
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96