CHAPTER 4 ESTIMATION

4.1 Introduction

The application of the methods of probability to the analysis and interpretation of data is known as statistical inference. In particular, we wish to make an inference about a *population* based on information contained in a *sample*. Since populations are characterized by numerical descriptive measures called parameters, the objective of many statistical investigations is to make an inference about one or more population parameters. There are two broad areas of inference: estimation (the subject of this chapter) and hypothesis-testing (the subject of the next chapter).

When we say that we have a random sample $X_1, X_2, \ldots, X_n$ "from a random variable X" or "from a population with distribution function $F(x;\theta)$," we mean that $X_1, X_2, \ldots, X_n$ are identically and independently distributed random variables each with c.d.f. $F(x;\theta)$, that is, depending on some parameter θ. We usually assume that the form of the distribution, e.g., binomial, Poisson, Normal, etc. is known but the parameter is unknown. We wish to obtain information from the data (sample) to enable us to make some statement about the parameter. Note that, θ may be a vector, e.g., $\theta = (\mu, \sigma^2)$.

The general problem of estimation is to find out something about θ using the information in the observed values of the sample, $x_1, x_2, \ldots, x_n$. That is, we want to choose a function $H(x_1, x_2, \ldots, x_n)$ that will give us a good estimate of the parameter θ in $F(x;\theta)$.

Next we will consider some general methods of estimation. Since different methods may lead to different estimators for the same parameter, we will then need to consider criteria for deciding whether one estimate is better than another.

4.2 Statistical philosophies

The dominant philosophy of inference is based on the *frequentist theory of probability.* According to the frequentist theory probability statements can only be made regarding events associated with a random experiment. Recall that events are subsets of the sample space assocatiated with the experiment. A random experiment is an experiment which has a well defined set of possible outcomes Ω. In addition, we must be able to envisage an infinite sequence of independent repetitions of the experiment with the actual outcome of each repetition being some unpredictable element of Ω. A random vector is a collection of numerical quantities associated with each possible outcome in Ω. In performing the experiment we determine which element of Ω has occurred and thereby the observed values of all random variables or random vectors of interest. Since the outcome of the experiment is unpredictable so too is the value of any random variable or random vector. Since we can envisage an infinite sequence of independent repetitions of the experiment, we can envisage an infinite sequence of independent determinations of the value of a random variable (or vector). The purpose of a statistical model is to describe the unpredictability of such a sequence of determinations.

Consider the random experiment which consists of picking someone at random from the 2007 electoral register for Limerick. The outcome of such an experiment will be a human being and the set Ω consists of all human beings whose names are on the register. We can clearly envisage an infinite sequence of independent repetitions of such an experiment. Consider the random variable X where $X = 0$ if the outcome of the experiment is a male and $X = 1$ if the outcome of the experiment is a female. When we say that $P(X = 1) = 0.54$ we are taken to mean that in an infinite sequence of independent repetitions of the experiment exactly 54% of the outcomes will produce a value of $X = 1$.

Now consider the random experiment which consists of picking 3 people at random from the 1997 electoral register for Limerick. The outcome of such an experiment will be a collection of 3 human beings and the set Ω consists of all subsets of 3 human beings which may be formed from the set of all human beings whose names are on the register. We can clearly envisage an infinite sequence of independent repetitions of such an experiment. Consider the random vector $\mathbf{X} = (X_1, X_2, X_3)$ where for $i = 1, 2, 3$, $X_i = 0$ if the ith person chosen is a male and $X_i = 1$ if the ith person chosen is a female. When we say that X_1, X_2, X_3 are *independent and identically distributed* or i.i.d. with $P(X_i = 1) = \theta$ we are taken to mean that in an infinite sequence of independent repetitions of the experiment the proportion of outcomes which produce, for instance, a value of $\mathbf{X} = (1, 1, 0)$ is given by $\theta^2(1 - \theta)$.

Suppose that the value of θ is unknown and we propose to estimate it by the estimator $\hat{\theta}$ whose value is given by the proportion of females in the sample of size 3. Since $\hat{\theta}$ depends on the value of $\mathbf{X}$ we sometimes write $\hat{\theta}(\mathbf{X})$ to emphasise this fact. We can work out the probability distribution of $\hat{\theta}$ as follows :

$\mathbf{X}$	$P(\mathbf{X}=\mathbf{x})$	$\hat{\theta}(\mathbf{x})$
$(0,0,0)$	$(1-\theta)^3$	0
$(0,0,1)$	$\theta(1-\theta)^2$	$1/3$
$(0,1,0)$	$\theta(1-\theta)^2$	$1/3$
$(1,0,0)$	$\theta(1-\theta)^2$	$1/3$
$(0,1,1)$	$\theta^2(1-\theta)$	$2/3$
$(1,0,1)$	$\theta^2(1-\theta)$	$2/3$
$(1,1,0)$	$\theta^2(1-\theta)$	$2/3$
$(1,1,1)$	θ^3	1

Thus $P(\hat{\theta}=0)=(1-\theta)^3, P(\hat{\theta}=1/3)=3\theta(1-\theta)^2, P(\hat{\theta}=2/3)=3\theta^2(1-\theta)$ and $P(\hat{\theta}=1)=\theta^3$.

We now ask whether $\hat{\theta}$ is a **good** estimator of θ? Clearly if $\theta=0$ we have that $P(\hat{\theta}=\theta)=P(\hat{\theta}=0)=1$ which is good. Likewise if $\theta=1$ we also have that $P(\hat{\theta}=\theta)=P(\hat{\theta}=1)=1$. If $\theta=1/3$ then $P(\hat{\theta}=\theta)=P(\hat{\theta}=1/3)=3(1/3)(1-1/3)^2=4/9$. Likewise if $\theta=2/3$ we have that $P(\hat{\theta}=\theta)=P(\hat{\theta}=2/3)=3(2/3)^2(1-2/3)=4/9$. However if the value of θ lies outside the set $\{0,1/3,2/3,1\}$ we have that $P(\hat{\theta}=\theta)=0$.

Since $\hat{\theta}$ is a random variable we might try to calculate its expected value $\mathrm{E}(\hat{\theta})$ i.e. the average value we would get if we carried out an infinite number of independent repetitions of the experiment. We have that

$$\begin{aligned}\mathrm{E}(\hat{\theta}) &= 0P(\hat{\theta}=0)+(1/3)P(\hat{\theta}=1/3)+(2/3)P(\hat{\theta}=2/3)+1P(\hat{\theta}=1)\ ,\\ &= 0(1-\theta)^3+(1/3)3\theta(1-\theta)^2+(2/3)3\theta^2(1-\theta)+1\theta^3\ ,\\ &= \theta.\end{aligned}$$

Thus if we carried out an infinite number of independent repetitions of the experiment and calculate the value of $\hat{\theta}$ for each repetition the average of the $\hat{\theta}$ values would be exactly θ,the true value of the parameter! This is true no matter what the actual value of θ is. Such an estimator is said to be **unbiased**.

Consider the quantity $L=(\hat{\theta}-\theta)^2$ which might be regarded as a measure of the error or loss involved in using $\hat{\theta}$ to estimate θ. The possible values for L are $(0-\theta)^2,\ (1/3-\theta)^2,\ (2/3-\theta)^2$ and $(1-\theta)^2$. We can calculate the expected value of L as follows:

$$\begin{aligned}\mathrm{E}(L) &= (0-\theta)^2P(\hat{\theta}=0)+(1/3-\theta)^2P(\hat{\theta}=1/3)\\ &\qquad +(2/3-\theta)^2P(\hat{\theta}=2/3)+(1-\theta)^2P(\hat{\theta}=1)\\ &= \theta^2(1-\theta)^3+(1/3-\theta)^2 3\theta(1-\theta)^2+(2/3-\theta)^2 3\theta^2(1-\theta)+(1-\theta)^2\theta^3\ ,\\ &= \theta(1-\theta)/3\ .\end{aligned}$$

The quantity $\mathrm{E}(L)$ is called the **mean squared error** (MSE) of the estimator $\hat{\theta}$. Since the quantity $\theta(1-\theta)$ attains its maximum value of $1/4$ for $\theta=1/2$, the largest value $\mathrm{E}(L)$ can attain is $1/12$ which occurs if the true value of the parameter θ happens to be equal to $1/2$; for all other values of θ the quantity $\mathrm{E}(L)$ is less than $1/12$. If somebody could invent a different estimator $\tilde{\theta}$ of θ whose MSE was less than that of $\hat{\theta}$ for *all* values of θ then we would prefer $\tilde{\theta}$ to $\hat{\theta}$.

This trivial example gives some idea of the kinds of calculations that we will be performing. The basic frequentist principle is that statistical procedures should be judged in terms of their *average* performance in an infinite series of independent repetitions of the experiment which produced the data. An important point to note is that the parameter values are treated as fixed (although unknown) throughout this infinite series of repetitions. We should be happy to use a procedure which performs well on the average and should not be concerned with how it performs on any one particular occasion.

4.3 The frequentist approach to estimation

Suppose that we are going to observe a value of a random vector $\mathbf{X}$. Let $\mathcal{X}$ denote the set of possible values $\mathbf{X}$ can take and, for $\mathbf{x} \in \mathcal{X}$, let $f(\mathbf{x}|\theta)$ denote the probability that $\mathbf{X}$ takes the value $\mathbf{x}$ where the parameter θ is some unknown element of the set Θ.

The problem we face is that of estimating θ. An estimator $\hat{\theta}$ is a procedure which for each possible value $\mathbf{x} \in \mathcal{X}$ specifies which element of Θ we should quote as an estimate of θ. When we observe $\mathbf{X} = \mathbf{x}$ we quote $\hat{\theta}(\mathbf{x})$ as our estimate of θ. Thus $\hat{\theta}$ is a function of the random vector $\mathbf{X}$. Sometimes we write $\hat{\theta}(\mathbf{X})$ to emphasise this point.

Given any estimator $\hat{\theta}$ we can calculate its expected value for each possible value of $\theta \in \Theta$. An estimator is said to be unbiased if this expected value is identically equal to θ. If an estimator is unbiased then we can conclude that if we repeat the experiment an infinite number of times with θ fixed and calculate the value of the estimator each time then the average of the estimator values will be exactly equal to θ. From the frequentist viewpoint this is a desireable property and so, where possible, frequentists use unbiased estimators.

Definition 4.1 (The Frequentist philosophy)**.** To evaluate the usefulness of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{x})$ of θ, examine the properties of the random variable $\hat{\theta} = \hat{\theta}(\mathbf{X})$.

Using different methods of estimation can lead to different estimators. Criteria for deciding which are good estimators are required. Before listing the qualities of a good estimator, it is important to understand that they are random variables. For example, suppose that we take a sample of size 5 from a uniform distribution and calculate x. Each time we repeat the experiment we will probably get a different sample of 5 and therefore a different $\overline{x}$. The behaviour of an estimator for different random samples will be described by a probability distribution. The actual distribution of the estimator is not a concern here and only its mean and variance will be considered. As a first condition it seems reasonable to ask that the distribution of the estimator be centered around the parameter it is estimating. If not it will tend to overestimate or underestimate θ. A second property an estimator should possess is precision. An estimator is precise if the dispersion of its distribution is small. These two concepts are incorporated in the definitions of *unbiasedness* and *efficiency* below. [5]

Definition 4.2 (Unbiased estimators)**.** An estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is said to be unbiased for a parameter θ if it equals θ in expectation:

$$\mathrm{E}[\hat{\theta}(\mathbf{X})] = \mathrm{E}(\hat{\theta}) = \theta.$$

Intuitively, an unbiased estimator is ‘right on target’.

Definition 4.3 (Bias of an estimator)**.** The bias of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ of θ is defined as $\text{bias}(\hat{\theta}) = \text{E}[\hat{\theta}(\mathbf{X}) - \theta]$.

There may be large number of unbiased estimators of a parameter for any given distribution and a further criterion for choosing between all the unbiased estimators is needed. [5]

Definition 4.4 (Bias corrected estimators)**.** If $\text{bias}(\hat{\theta})$ is of the form $c\theta$, then (obviously) $\tilde{\theta} = \hat{\theta}/(1+c)$ is unbiased for θ. Likewise, if $\text{bias}(\hat{\theta}) = \theta + c$, then $\tilde{\theta} = \hat{\theta} - c$ is unbiased for θ. In such situations we say that $\tilde{\theta}$ is a biased corrected version of $\hat{\theta}$.

Definition 4.5 (Unbiased functions)**.** More generally $\hat{g}(\mathbf{X})$ is said to be unbiased for a function $g(\theta)$ if $\text{E}[\hat{g}(\mathbf{X})] = g(\theta)$.

Note that even if $\hat{\theta}$ is an unbiased estimator of θ, $g(\hat{\theta})$ will generally not be an unbiased estimator of $g(\theta)$ unless g is linear or affine. This limits the importance of the notion of unbiasedness. It might be at least as important that an estimator is accurate in the sense that its distribution is highly concentrated around θ.

Is unbiasedness a good thing? Unbiasedness is important when combining estimates, as averages of unbiased estimators are unbiased (see the review exercises at the end of this chapter). For example, when combining standard deviations $s_1, s_2, \ldots, s_k$ with degrees of freedom $\text{df}_1, \ldots, \text{df}_k$ we always average their squares

$$\bar{s} = \sqrt{\frac{\text{df}_1 s_1^2 + \cdots + \text{df}_k s_k^2}{\text{df}_1 + \cdots + \text{df}_k}}$$

as s_i^2 are unbiased estimators of the variance σ^2, whereas s_i are not unbiased estimators of σ (see the review exercises). Be careful when averaging biased estimators! It may well be appropriate to make a bias-correction before averaging.

4.4 Estimation by the method of moments

Now that we have a better understanding about some factors that may play a role in determining the quality of an estimator, we would like to know about some methods of finding estimators. We will focus on finding *point estimators* first, i.e. for which the true value of a (function of a) parameter is assumed to be a point. Several methods exist to compute point estimators, including the 'methods of moments' and 'maximul likelihood' (which we will discuss in more detail in this course), but also the 'method of least squares' (see the Chapter 9 on 'Regression'), the 'Bayes' method, the 'minimum-chi-square' method and the 'minimum-distance' method.

Point estimates should preferentially be accompanied by some interval about the point estimate together with some measure of accurance that the true value of the parameter lies within the interval. Hence, instead of making the inference of estimating the true value of the parameter to be a point, we might make the inference of estimating that the true value of the parameter is contained in some interval. We then speak of *interval estimation*, which will be the subject of Chapter 5.

4.4.1 Traditional methods of moments

The Method of Moments was first proposed near the turn of the century by the British statistician Karl Pearson. The Method of Maximum Likelihood though goes back much further and will be dealt with later. Both Gauss and Daniel Bernoulli made use of the technique, the latter as early as 1777. Fisher though, in the early years of the twentieth century, was the first to make a thorough study of the method's properties and the procedure is often credited to him. [5]

Recall that, for a random variable X, the rth moment about the origin is $\mu'_r = E(X^r)$ and that for a random sample $X_1, X_2, \ldots, X_n$, the rth sample moment about the origin is defined by

$$M_r = \sum_{i=1}^{n} X_i^r / n, r = 1, 2, 3, \ldots$$

and its observed value is denoted by

$$m_r = \sum_{i=1}^{n} x_i^r / n.$$

Note that the first sample moment is just the sample mean, $\overline{X}$.

We will first prove a property of sample moments.

Theorem 4.1. *Let $X_1, X_2, \ldots, X_n$ be a random sample of X. Then*

$$E(M_r) = \mu'_r, r = 1, 2, 3, \ldots$$

Proof.

$$E(M_r) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^r\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^r) = \frac{1}{n}\sum_{i=1}^{n} \mu'_r = \mu'_r.$$

□

This theorem provides the motivation for estimation by the method of moments (with the estimator being referred to as the method of moments estimator or MME). The sample moments, $M_1, M_2, \ldots$, are random variables whose means are $\mu'_1, \mu'_2, \ldots$ Since the population moments depend on the parameters of the distribution, estimating them by the sample moments leads to estimation of the parameters.

We will consider this method of estimation by means of 2 examples, then state the general procedure.

Example 4.1.
In this example, the distribution only has one parameter. Given $X_1, X_2, \ldots, X_n$ is a random sample from a $U(0, \theta)$ distribution, find the method of moments estimator (MME) of θ.

Solution of Example 4.1. *Now, for the uniform distribution ($f(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$),*

$$\begin{aligned} \mu = E(X) &= \int_0^\theta x \times \frac{1}{\theta} dx \\ &= \frac{\theta}{2} \end{aligned}$$

Using the Method of Moments we proceed to estimate $\mu = \frac{\theta}{2}$ *by* m_1*. Thus since* $m_1 = \overline{x}$ *we have*

$$\frac{\overline{\theta}}{2} = \overline{x}$$

and

$$\overline{\theta} = 2\overline{x}.$$

Then, $\overline{\theta} = 2\overline{X}$ *and the MME of* θ *is* $2\overline{X}$*.*

Computer Exercise 4.1. *Generate 100 samples of size 10 from a uniform distribution,* $U(0,\theta)$ *with* $\theta = 10$*. Estimate the value of* θ *from your samples using the method of moments and plot the results. Comment on the results.*

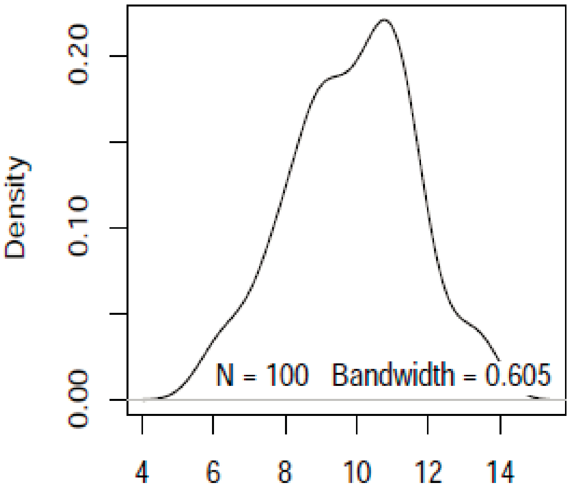
In this exercise, we know a priori that $\theta = 10$ *and have generated random samples. The samples are analysed as if* θ *is unknown and estimated by the method of moments. Then we can compare the estimates with the known value.*

Solution of Computer Exercise 4.1.

```
#_________ UniformMoment.R _____
theta <- 10
sampsz <- 10
nsimulations <- 100
theta.estimates <- numeric(nsimulations)

for (i in 1:nsimulations){
ru <- runif(n=sampsz,min=0,max=theta)
Xbar <- mean(ru)
 theta.estimates[i] <- 2*Xbar
} # end of the i loop

plot(density(theta.estimates))
```



(You should do the exercise and obtain a plot for yourself.)

It should be clear from the plot that about 50% of the estimates are greater than 10 which is outside the parameter space for a $U(0,10)$ *distribution. This is undesirable.*

Example 4.2.
In this example the distribution has two parameters. Given $X_1,\ldots,X_n$ *is a random sample from the* $N(\mu,\sigma^2)$ *distribution, find the method of moments estimates of* μ *and* σ^2*.*

Solution of Example 4.2. *For the normal distribution,* $E(X) = \mu$ *and* $E(X^2) = \sigma^2 + \mu^2$ *(Thm. 2.2 of [5]).*

Using the Method of Moments: Equate $E(X)$ to m_1 and $E(X^2)$ to m_2 so that, $\overline{\mu} = \overline{x}$ and $\overline{\sigma}^2 + \overline{\mu}^2 = m_2$. that is, estimate μ by $\overline{x}$ and estimate σ^2 by $m_2 - \overline{x}^2$. Then,

$$\overline{\mu} = \overline{x}, \text{ and } \overline{\sigma}^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2.$$

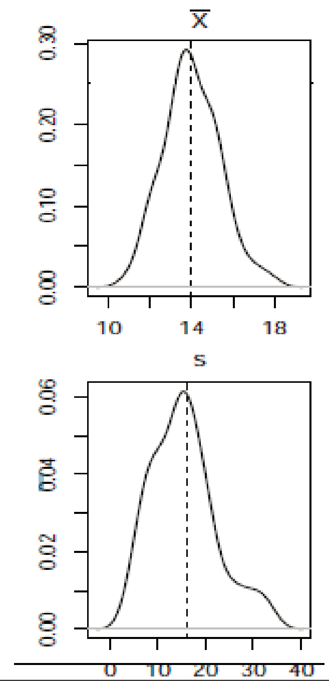
The latter can also be written as $\overline{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Computer Exercise 4.2. *Generate 100 samples of size 10 from a normal distribution with $\mu = 14$ and $\sigma = 4$. Estimate μ and σ^2 from your samples using the method of moments. Plot the estimated values of μ and σ^2. Comment on your results.*

Solution of Computer Exercise 4.2.

```
#_____________NormalMoments.R ___________
mu <- 14
sigma <- 4
sampsz <- 10
nsimulations <- 100
mu.estimates <- numeric(nsimulations)
var.estimates <- numeric(nsimulations)
for (i in 1:nsimulations){
rn <- rnorm(mean=mu,sd=sigma,n=sampsz)
mu.estimates[i] <- mean(rn)
var.estimates[i] <- mean( (rn -mean(rn))^2 )
} # end of i loop

plot(density(mu.estimates))
plot(density(var.estimates))
```



The plot you obtain for the row means should be centred around the true mean of 14. However, you will notice that the plot of the variances is not centred about the true variance of 16 as you would like. Rather it will appear to be centred about a value less than 16. The reason for this will become evident when we study the properties of MME estimators in more detail.

General Procedure

Let $X_1, X_2, \dots, X_n$ be a random sample from $F(x : \theta_1, \dots, \theta_k)$. That is, suppose that there are k parameters to be estimated. Let μ'_r, m_r $(r = 1, 2, \dots, k)$ denote the first k population and sample moments respectively, and suppose that each of these population moments are certain known functions of the parameters. That is,

$$\begin{aligned} \mu'_1 &= g_1(\theta_1, \dots, \theta_k), \\ \mu'_2 &= g_2(\theta_1, \dots, \theta_k), \\ &\vdots \\ \mu'_k &= g_k(\theta_1, \dots, \theta_k). \end{aligned}$$

Solving simultaneously the set of equations,

$$\mu_r' = g_r(\overline{\theta_1}, \ldots, \overline{\theta_k}) = m_r, r = 1, 2, \ldots, k$$

gives the required estimates $\overline{\theta_1}, \ldots, \overline{\theta_k}$.

4.4.2 Generalized methods of moments

Sometimes, methods of moments estimation in non-conclusive in deriving estimates. We will show this in the context of estimating the slope of a line through the origin (Chapter 9). The Generalized Methods of Moments provides an alternative. It offers an estimation strategy when the number of restricted moments exceeds the number of parameters to be estimated.

4.5 Properties of an estimator

4.5.1 Unbiasedness

Recall the definition of an unbiased estimator: Definition 4.2. From a frequentist point of view, this is a very desirable property for an estimator.

Problem 4.2. Let X have a binomial distribution with parameters n and θ. Show that the sample proportion $\hat{\theta} = X/n$ is an unbiased estimate of θ.

Solution. $X \sim \text{Bin}(n, \theta) \Rightarrow \text{E}(X) = n\theta$. Then $\text{E}(\hat{\theta}) = \text{E}(X/n) = \text{E}(X)/n = n\theta/n = \theta$. As $\text{E}(\hat{\theta}) = \theta$, the estimator $\hat{\theta}$ is unbiased. □

Problem 4.3. Let $X_1, \ldots, X_n$ be independent and identically distributed with density

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)}, & \text{for } x > \theta; \\ 0, & \text{otherwise.} \end{cases}$$

Show that $\hat{\theta} = \bar{X} = (X_1 + \cdots + X_n)/n$ is a biased estimator of θ. Propose an unbiased estimator $\tilde{\theta}$ of the form $\tilde{\theta} = \hat{\theta} + c$.

Solution. $\text{E}(X) = \int_\theta^\infty xe^{-(x-\theta)}\text{d}x = \left[-xe^{-(x-\theta)} + \int e^{-(x-\theta)}\text{d}x\right]_\theta^\infty = -(x+1)e^{-(x-\theta)}|_\theta^\infty = \theta + 1$. Next, $\text{E}(\hat{\theta}) = \text{E}(\bar{X}) = \frac{1}{n}\text{E}(X_1 + X_2 + \cdots + X_n) = \theta + 1 \neq \theta \Rightarrow \hat{\theta}$ is biased. Propose $\tilde{\theta} = \bar{X} - 1$. Then $\text{E}(\tilde{\theta}) = \text{E}(\bar{X}) - 1 = \theta + 1 - 1 = \theta$ and $\tilde{\theta}$ is unbiased. □

4.5.2 Trading off Bias and Variance

4.5.2.1 Mean-Squared Error

Although desirable from a frequentist standpoint, unbiasedness is not a property that helps us choose between estimators. To do this we must examine some measure of loss like the mean squared error.

Definition 4.6 (Mean squared error)**.** The mean squared error of the estimator $\hat{\theta}$ is defined as $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2$. Given the same set of data, $\hat{\theta}_1$ is "better than $\hat{\theta}_2$ if $MSE(\hat{\theta}_1) \leq \mathrm{MSE}(\hat{\theta}_2)$ (uniformly better if true $\forall\, \theta$).

Lemma 4.4 (The MSE variance-bias tradeoff)**.** The MSE decomposes as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2.$$

Proof. The problem of finding minimum MSE estimators cannot be solved uniquely:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}(\hat{\theta} - \theta)^2 \\
&= \mathrm{E}\{\ [\ \hat{\theta} - \mathrm{E}(\hat{\theta})\ ]\ + [\ \mathrm{E}(\hat{\theta}) - \theta\ ]\}^2 \\
&= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]^2 + \mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta]^2 \\
&\qquad + 2\underbrace{\mathrm{E}\left\{[\hat{\theta} - \mathrm{E}(\hat{\theta})][\mathrm{E}(\hat{\theta}) - \theta]\right\}}_{=0} \\
&= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]^2 + \mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta]^2 \\
&= \mathrm{Var}(\hat{\theta}) + \underbrace{[\mathrm{E}(\hat{\theta}) - \theta]^2}_{\mathrm{Bias}(\hat{\theta})^2}
\end{aligned}$$

□

Note : This lemma implies that the mean squared error of an unbiased estimator is equal to the variance of the estimator.

Problem 4.5. Consider $X_1, \ldots, X_n$ where $X_i \sim \mathrm{N}(\theta, \sigma^2)$ and σ is known. Three estimators of θ are $\hat{\theta}_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\hat{\theta}_2 = X_1$, and $\hat{\theta}_3 = (X_1 + \bar{X})/2$. Pick one.

Solution. $\mathrm{E}(\hat{\theta}_1) = \frac{1}{n}[\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)] = \frac{1}{n}[\theta + \cdots + \theta] = \frac{1}{n}[n\theta] = \theta$, (unbiased). Next $\mathrm{E}(\hat{\theta}_2) = \mathrm{E}(X_1) = \theta$, (unbiased). Finally $\mathrm{E}(\hat{\theta}_3) = \frac{1}{2}\mathrm{E}\left\{\frac{n+1}{n}X_1 + \frac{1}{n}[X_2 + \cdots + X_n]\right\} = \frac{1}{2}\left\{\frac{n+1}{n}\mathrm{E}(X_1) + \frac{1}{n}[\mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)]\right\} = \frac{1}{2}\{\frac{n+1}{n}\theta + \frac{n-1}{n}\theta\} = \theta$, (unbiased). All three estimators are unbiased. For a class of estimators that are unbiased, the mean squared error will be equal to the estimation variance. Calculate $\mathrm{Var}(\hat{\theta}_1) = \frac{1}{n^2}[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)] = \frac{1}{n^2}[\sigma^2 + \cdots + \sigma^2] = \frac{1}{n^2}[n\sigma^2] = \frac{1}{n}\sigma^2$. Trivially $\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}(X_1) = \sigma^2$. Finally $\mathrm{Var}(\hat{\theta}_3) = (\sigma^2/n + \sigma^2)/4 + 2\mathrm{Cov}(\bar{X}, X_1)$. So $\bar{X}$ appears "best" in the sense that $\mathrm{Var}(\hat{\theta})$ is smallest among these three unbiased estimators. □

Problem 4.6. Consider $X_1, \ldots, X_n$ to be independent random variables with means $\mathrm{E}(X_i) = \mu + \beta_i$ and variances $\mathrm{Var}(X_i) = \sigma_i^2$. Such a situation could arise when X_i are estimators of μ obtained from independent sources and β_i is the bias of the estimator X_i. Consider pooling the estimators of μ into a common estimator using the linear combination $\hat{\mu} = w_1X_1 + w_2X_2 + \cdots + w_nX_n$.

(i) If the estimators are unbiased, show that $\hat{\mu}$ is unbiased if and only if $\sum w_i = 1$.

(ii) In the case when the estimators are unbiased, show that $\hat{\mu}$ has minimum variance when the weights are inversely proportional to the variances σ_i^2.

(iii) Show that the variance of $\hat{\mu}$ for optimal weights w_i is $\mathrm{Var}(\hat{\mu}) = 1/\sum_i \sigma_i^{-2}$.

(iv) Consider the case when estimators may be biased. Find the mean square error of the optimal linear combination obtained above, and compare its behaviour as $n \to \infty$ in the biased and unbiased case, when $\sigma_i^2 = \sigma^2,\ i = 1, \dots, n$.

Solution. $\mathrm{E}(\hat{\mu}) = \mathrm{E}(w_1 X_1 + \cdots + w_n X_n) = \sum_i w_i \mathrm{E}(X_i) = \sum_i w_i \mu = \mu \sum_i w_i$ so $\hat{\mu}$ is unbiased if and only if $\sum_i w_i = 1$. The variance of our estimator is $\mathrm{Var}(\hat{\mu}) = \sum_i w_i^2 \sigma_i^2$, which should be minimized subject to the constraint $\sum_i w_i = 1$. Differentiating the Lagrangian $\mathcal{L} = \sum_i w_i^2 \sigma_i^2 - \lambda \left(\sum_i w_i - 1 \right)$ with respect to w_i and setting equal to zero yields $2 w_i \sigma_i^2 = \lambda \Rightarrow w_i \propto \sigma_i^{-2}$ so that $w_i = \sigma_i^{-2} / (\sum_j \sigma_j^{-2})$. Then, for optimal weights we get $\mathrm{Var}(\hat{\mu}) = \sum_i w_i^2 \sigma_i^2 = (\sum_i \sigma_i^{-4} \sigma_i^2)/(\sum_i \sigma_i^{-2})^2 = 1/(\sum_i \sigma_i^{-2})$. When $\sigma_i^2 = \sigma^2$ we have that $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$ which tends to zero for $n \to \infty$ whereas $\mathrm{bias}(\hat{\mu}) = \sum \beta_i / n = \bar{\beta}$ is equal to the average bias and $\mathrm{MSE}(\hat{\mu}) = \sigma^2/n + \bar{\beta}^2$. Therefore the bias tends to dominate the variance as n gets larger, which is very unfortunate. □

Problem 4.7. Let $X_1, \dots, X_n$ be an independent sample of size n from the *uniform* distribution on the interval $(0, \theta)$, with density for a single observation being $f(x|\theta) = \theta^{-1}$ for $0 < x < \theta$ and 0 otherwise, and consider $\theta > 0$ unknown.

(i) Find the expected value and variance of the estimator $\hat{\theta} = 2\bar{X}$.

(ii) Find the expected value of the estimator $\tilde{\theta} = X_{(n)}$, i.e. the largest observation.

(iii) Find an unbiased estimator of the form $\check{\theta} = cX_{(n)}$ and calculate its variance.

(iv) Compare the mean square error of $\hat{\theta}$ and $\check{\theta}$.

Solution. $\hat{\theta}$ has $\mathrm{E}(\hat{\theta}) = \mathrm{E}(2\bar{X}) = \frac{2}{n}[\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)] = \frac{2}{n}[(\theta/2) + \cdots + (\theta/2)] = \frac{2}{n}[n(\theta/2)] = \theta$ (unbiased), and $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(2\bar{X}) = \frac{4}{n^2}[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)] = \frac{4}{n^2}[(\theta^2/12) + \cdots + (\theta^2/12)] = \frac{4}{n^2}\frac{n}{12}\theta^2 = \frac{1}{3n}\theta^2$. Let $U = X_{(n)}$, we then have $P(U \le u) = \prod_i^n P(X_i \le u) = (u/\theta)^n$ for $0 < u < \theta$ so differentiation yields that U has density $f(u|\theta) = n u^{n-1} \theta^{-n}$ for $0 < u < \theta$. Direct integration now yields $\mathrm{E}(\tilde{\theta}) = \mathrm{E}(U) = \frac{n}{n+1}\theta$ (a biased estimator). The estimator $\check{\theta} = \frac{n+1}{n} U$ is unbiased. Direct integration gives $\mathrm{E}(U^2) = \frac{n}{n+2}\theta^2$ so $\mathrm{Var}(\tilde{\theta}) = \mathrm{Var}(U) = \frac{n}{(n+2)(n+1)^2}\theta^2$ and $\mathrm{Var}(\check{\theta}) = \frac{1}{n(n+2)}\theta^2$. As $\hat{\theta}$ and $\check{\theta}$ are both unbiased estimators $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$ and $\mathrm{MSE}(\check{\theta}) = \mathrm{Var}(\check{\theta})$. Clearly the mean square error of $\hat{\theta}$ is very large compared to the mean square error of $\check{\theta}$. □

4.5.2.2 Minimum-Variance Unbiased

Getting a small MSE often involves a tradeoff between variance and bias. By not insisting on $\hat{\theta}$ being unbiased, the variance can sometimes be drastically reduced. For unbiased estimators, the MSE obviously equals the variance, $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$, so no tradeoff can be made. One approach is to restrict ourselves to the subclass of estimators that are *unbiased* and *minimum variance.*

Definition 4.7 (Minimum-variance unbiased estimator)**.** If an unbiased estimator of $g(\theta)$ has minimum variance among all unbiased estimators of $g(\theta)$ it is called a minimum variance unbiased estimator (MVUE).

We will develop a method of finding the MVUE when it exists. When such an estimator does not exist we will be able to find a lower bound for the variance of an unbiased estimator in the class of unbiased estimators, and compare the variance of our unbiased estimator with this lower bound.

Definition 4.8 (Score function)**.** For the (possibly vector valued) observation $X = x$ to be informative about θ, the density must vary with θ. If $f(x|\theta)$ is smooth and differentiable, this change is quantified to first order by the score function

$$S(\theta) = \frac{\partial}{\partial\theta} \ln f(x|\theta) \equiv \frac{f'(x|\theta)}{f(x|\theta)}.$$

Under suitable regularity conditions (differentiation wrt θ and integration wrt x can be interchanged), we have

$$\begin{aligned} \mathrm{E}\{S(\theta)\} &= \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)\mathrm{d}x \;=\; \int f'(x|\theta)\mathrm{d}x \;, \\ &= \frac{\partial}{\partial\theta}\left\{\int f(x|\theta)\mathrm{d}x\right\} \;=\; \frac{\partial}{\partial\theta}\,1 \;=\; 0. \end{aligned}$$

Thus the score function has expectation zero. Using the central limit theorem, the score statistic has asymptotically a normal distribution, with mean zero. The variance will be derived below.

True frequentism evaluates the properties of estimators based on their "long-run" behaviour. The value of x will vary from sample to sample so we have treated the score function as a random variable and looked at its average across all possible samples.

Lemma 4.8 (Fisher information)**.** The variance of $S(\theta)$ is the expected Fisher information about θ

$$\mathrm{E}(\mathcal{I}(\theta)) = \mathrm{E}\{S(\theta)^2\} \equiv \mathrm{E}\left\{\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2\right\}$$

Proof. Using the chain rule

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2}\ln f &= \frac{\partial}{\partial\theta}\left[\frac{1}{f}\frac{\partial f}{\partial\theta}\right] \\ &= -\frac{1}{f^2}\left[\frac{\partial f}{\partial\theta}\right]^2 + \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2} \\ &= -\left[\frac{\partial \ln f}{\partial\theta}\right]^2 + \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2} \end{aligned}$$

If integration and differentiation can be interchanged

$$\mathrm{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\right] = \int_X \frac{\partial^2 f}{\partial\theta^2}\mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\int_X \mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\,1 \;=\; 0,$$

thus

$$-\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2\right] = \mathrm{E}(\mathcal{I}(\theta)). \tag{4.5.1}$$

Variance measures lack of knowledge. Reasonable that the reciprocal of the variance should be defined as the amount of information carried by the (possibly vector valued) observation x about θ. □

Theorem 4.9 (Cramér Rao lower bound)**.** Let $\hat{\theta}$ be an unbiased estimator of θ. Then

$$\mathrm{Var}(\hat{\theta}) \;\geq\; \{\; \mathrm{E}(\mathcal{I}(\theta)) \;\}^{-1}.$$

Proof. Unbiasedness, $\mathrm{E}(\hat{\theta}) = \theta$, implies

$$\int \hat{\theta}(x) f(x|\theta) \mathrm{d}x = \theta.$$

Assume we can differentiate wrt θ under the integral, then

$$\int \frac{\partial}{\partial\theta} \left\{ \hat{\theta}(x) f(x|\theta) \mathrm{d}x \right\} = 1.$$

The estimator $\hat{\theta}(x)$ can't depend on θ, so

$$\int \hat{\theta}(x) \frac{\partial}{\partial\theta} \{ f(x|\theta) \mathrm{d}x \} = 1.$$

For any pdf f,

$$\frac{\partial f}{\partial\theta} = f \frac{\partial}{\partial\theta} (\ln f)\,,$$

so that now

$$\int \hat{\theta}(x) f \frac{\partial}{\partial\theta} (\ln f)\, \mathrm{d}x = 1.$$

Thus

$$\mathrm{E}\left[\hat{\theta}(x) \frac{\partial}{\partial\theta} (\ln f) \right] = 1.$$

Define random variables

$$U = \hat{\theta}(x),$$

and

$$S = \frac{\partial}{\partial\theta} (\ln f)\,.$$

Then $\mathrm{E}\,(US) = 1$. We already know that the score function has expectation zero, $\mathrm{E}\,(S) = 0$. Consequently $\mathrm{Cov}(U,S) = \mathrm{E}(US) - \mathrm{E}(U)\mathrm{E}(S) = \mathrm{E}(US) = 1$.

$$\{\mathrm{Corr}(U,S)\}^2 \;=\; \frac{\{\mathrm{Cov}(U,S)\}^2}{\mathrm{Var}(U)\mathrm{Var}(S)} \;\leq\; 1$$

Setting $\mathrm{Cov}(U,S) = 1$ we get

$$\mathrm{Var}(U)\mathrm{Var}(S) \;\geq\; 1$$

This implies

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathrm{E}(\mathcal{I}(\theta))}$$

which is our main result. We call $\{\; \mathrm{E}(\mathcal{I}(\theta)) \;\}^{-1}$ the Cramér Rao lower bound (CRLB). □

Sufficient conditions for the proof of CRLB are that all the integrands are finite, within the range of x. We also require that the limits of the integrals do not depend on θ. That is, the range of x, here $f(x|\theta)$, cannot depend on θ. This second condition is violated for many density functions, i.e. the CRLB is not valid for the uniform distribution. We can have absolute assessment for unbiased estimators by comparing their variances to the CRLB. We can also assess biased estimators. If its variance is lower than CRLB then it is indeed a very good estimate, although it is biased.

Example 4.3.
Consider i.i.d. random variables $X_i,\ i = 1, \dots, n$, *with*

$$f_{X_i}(x_i|\mu) = \frac{1}{\mu}\exp\left(-\frac{1}{\mu}x_i\right).$$

Denote the joint distribution of $X_1, \dots, X_n$ *by*

$$f = \prod_{i=1}^{n} f_{X_i}(x_i|\mu) = \left(\frac{1}{\mu}\right)^n \exp\left(-\frac{1}{\mu}\sum_{i=1}^{n} x_i\right),$$

so that

$$\ln f = -n\ln(\mu) - \frac{1}{\mu}\sum_{i=1}^{n} x_i.$$

The score function is the partial derivative of $\ln f$ *wrt the unknown parameter* μ,

$$S(\mu) = \frac{\partial}{\partial\mu}\ln f = -\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} x_i$$

and

$$\mathrm{E}\left\{S(\mu)\right\} = \mathrm{E}\left\{-\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} X_i\right\} = -\frac{n}{\mu} + \frac{1}{\mu^2}\mathrm{E}\left\{\sum_{i=1}^{n} X_i\right\}$$

For $X \sim \mathrm{Exp}(1/\mu)$, *we have* $\mathrm{E}(X) = \mu$ *implying* $\mathrm{E}(X_1 + \cdots + X_n) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = n\mu$ *and* $\mathrm{E}\left\{S(\mu)\right\} = 0$ *as required.*

$$\begin{aligned}
\mathrm{E}(\mathcal{I}(\mu)) &= -\mathrm{E}\left\{\frac{\partial}{\partial\mu}\left(-\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} X_i\right)\right\} \\
&= -\mathrm{E}\left\{\frac{n}{\mu^2} - \frac{2}{\mu^3}\sum_{i=1}^{n} X_i\right\} \\
&= -\frac{n}{\mu^2} + \frac{2}{\mu^3}\mathrm{E}\left\{\sum_{i=1}^{n} X_i\right\} \\
&= -\frac{n}{\mu^2} + \frac{2n\mu}{\mu^3} = \frac{n}{\mu^2}
\end{aligned}$$

Hence

$$\mathrm{CRLB} = \frac{\mu^2}{n}.$$

Let us propose $\hat{\mu} = \bar{X}$ *as an estimator of* μ. *Then*

$$\mathrm{E}(\hat{\mu}) = \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} = \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n} X_i\right\} = \mu,$$

verifying that $\hat{\mu} = \bar{X}$ *is indeed an unbiased estimator of* μ. *For* $X \sim \mathrm{Exp}(1/\mu)$, *we have* $\mathrm{E}(X) = \mu$ *and* $\mathrm{Var}(X) = \mu^2$, *implying*

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\mu^2}{n^2} = \frac{\mu^2}{n}.$$

We have therefore shown that $\mathrm{Var}(\hat{\mu}) = \{\ \mathrm{E}(\mathcal{I}(\theta))\ \}^{-1}$, *and therefore conclude that the unbiased estimator* $\hat{\mu} = \bar{x}$ *achieves its CRLB.*

4.5.3 Efficiency

Definition 4.9 (Efficiency)**.** Define the efficiency of the unbiased estimator $\hat{\theta}$ as

$$\mathrm{eff}(\hat{\theta}) \ = \ \frac{\mathrm{CRLB}}{\mathrm{Var}(\hat{\theta})}\ ,$$

where $\mathrm{CRLB} = \{\ \mathrm{E}(\mathcal{I}(\theta))\ \}^{-1}$. Clearly $0 < \mathrm{eff}(\hat{\theta}) \leq 1$. An unbiased estimator $\hat{\theta}$ is said to be efficient if $\mathrm{eff}(\hat{\theta}) = 1$.

Definition 4.10 (Asymptotic efficiency)**.** The *asymptotic efficiency* of an unbiased estimator $\hat{\theta}$ is the limit of the efficiency as $n \to \infty$. An unbiased estimator $\hat{\theta}$ is said to be asymptotically efficient if its asymptotic efficiency is equal to 1.

Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be 2 unbiased estimators of θ with variances $Var(\widehat{\theta_1})$, $Var(\widehat{\theta_2})$ respectively. We can then say that $\widehat{\theta_1}$ is **more efficient** than $\widehat{\theta_2}$ if

$$Var(\widehat{\theta_1}) < Var(\widehat{\theta_2}).$$

That is, $\widehat{\theta_1}$ is more efficient than $\widehat{\theta_2}$ if it has a smaller variance.

Definition 4.11 (Relative Efficiency)**.** The **relative efficiency** of $\widehat{\theta_2}$ with respect to $\widehat{\theta_1}$ is defined as $Var(\widehat{\theta_1})/Var(\widehat{\theta_2})$

It will now be useful to indicate that the estimator is based on a sample of size n by denoting it by $\widehat{\theta_n}$.

4.5.4 Consistency

In the previous subsections, we defined unbiasedness and mean-squared error of an estimator. Both concepts were defined on a fixed sample size. In this subsection we will define two concepts that are defined for increasing sample size.

Definition 4.12 (Mean-squared-error Consistency)**.** Let $\widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_n}, \ldots$ be a sequence of estimators of θ, where each estimator $\widehat{\theta_n}$ based on a sample of size n. This sequence of estimators is defined to be a *mean-squared-error* consistent sequence of estimators of θ if and only if

$$\lim_{n\to\infty} E[(\widehat{theta_n} - \theta)^2] = 0, \text{for all } \theta \tag{4.5.2}$$

Mean-squared-error consistency implies that both the bias and the variance of $\widehat{\theta_n}$ approaches 0, since the mean-squared-error in the definition can be decomposed in a variance and squared-bias component.

Definition 4.13 (Simple (weakly) consistency)**.** $\widehat{\theta_n}$ is a **consistent** estimator of θ if

$$\lim_{n\to\infty} P(\left|\widehat{\theta_n} - \theta\right| > \epsilon) = 0\text{for all } \epsilon > 0. \tag{4.5.3}$$

We then say that $\widehat{\theta_n}$ **converges in probability** to θ as $n \to \infty$. Equivalently,

$$\lim_{n\to\infty} P(\left|\widehat{\theta_n} - \theta\right| < \epsilon) = \lim_{n\to\infty} P(\left|\widehat{\theta_n} - \epsilon\right| < \theta < \widehat{\theta_n} + \epsilon) = 1. \tag{4.5.4}$$

If an estimator is a mean-squared-error consistent estimator, it is also a simple consistent estimator, but not necessarily vice versa.

These are large-sample or *asymptotic* properties. Consistency has to do only with the limiting behaviour of an estimator as the sample size increases without limit and does not imply that the observed value of $\widehat{\theta}$ is necessarily close to θ for any specific size of sample n. if only a relatively small sample is available, it would seem immaterial whether a consistent estimator is used or not.

The following theorem (which will not be proven) gives a method of testing for consistency.

Theorem 4.10. *If,* $\lim_{n\to\infty} E(\widehat{\theta_n}) = \theta$ *and* $\lim_{n\to\infty} Var(\widehat{\theta_n}) = 0$, *then* $\widehat{\theta_n}$ *is a consistent estimator of* θ.

4.5.5 Loss and Risk Functions

The concept of a mean-squared-error can be seen as a measure of how 'close' an estimate is to the 'truth'. Using the language of 'decision theory', in which one wants to make a 'decision' about an estimate, on e might call the value of some estimator $\widehat{\theta} = H(X_1, \ldots, H_n)$ a *decision* and call the estimator itself a *decision function*, since it tells us what decision to make. The estimate itself may be in error. If so, some measure of severity of the error seems appropriate. The word 'loss' is used in place of 'error', and 'loss function' is used as a measure of 'error'. A formal definition follows.

Definition 4.14 (Loss function)**.** Consider estimating θ. Let $H(x_1, \ldots, x_n)$ denote an estimate of θ. The *loss function* denoted by $l(H(x_1, \ldots, x_n); \theta)$ is defined to be a real-valued function satisfying

(i) $l(H(x_1, \ldots, x_n); \theta) \geq 0$ for all possible estimates $H(x_1, \ldots, x_n)$ and all allowable θ

(ii) $l(H(x_1, \ldots, x_n); \theta) = 0$ for $H(x_1, \ldots, x_n) = \theta$

The function $l(H(x_1, \ldots, x_n); \theta)$ equals the *loss* incurred if one estimates the true parameter to be $\widehat{\theta}$

One popular loss function is the squared error loss function, defined as $(\widehat{\theta} - \theta)^2$.

Definition 4.15 (Risk function)**.** For a given loss function $l(.;.)$ the *risk function* denoted by $\mathcal{R}_l(\theta)$ of an estimator $\widehat{\theta}$ is defined to be $\mathcal{R}_l(\theta) = E(l(\widehat{\theta}; \theta))$

The risk function is the average loss. For the squared error loss function, the risk function is the familiar mean-squared-error.

4.6 Sufficiency

The final concept of **sufficiency** requires some explanation before a formal definition is given. The random sample $X_1, X_2, \ldots, X_n$ drawn from the distribution with $F(x;\theta)$ contains information about the parameter θ. To estimate θ, this sample is first condensed to a single random variable by use of a statistic $\theta^* = H(X_1, X_2, \ldots, X_n)$. The question of interest is whether any information about θ has been lost by this condensing process. For example, a possible choice of θ^* is $H(X_1, \ldots, X_n) = X_1$ in which case it seems that some of the information in the sample has been lost since the observations $X_2, \ldots, X_n$ have been ignored. In many cases, the statistic θ^* does contain all the relevant information about the parameter θ that the sample contains.

Definition 4.16 (Sufficiency)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $F(x;\theta)$ and let $\theta^* = H(X_1, X_2, \ldots, X_n)$ be a statistic (a function of the X_i only). Let $\theta' = H'(X_1, X_2, \ldots, X_n)$ be any other statistic which is not a function of θ^*. If for each of the statistics θ', the conditional density of θ' given θ^* does not involve θ, then θ^* is called a **sufficient statistic** for θ. That is, $f(\theta'|\theta^*)$ does not contain θ, then θ^* is sufficient for θ.

You should think of sufficiency in the sense of using all the relevant information in the sample. For example, to say that $\overline{x}$ is sufficient for μ in a particular distribution means that knowledge of the actual observations $x_1, x_2, \ldots, x_n$ gives us no more information about μ than does only knowing the average of the n observations.

Example 4.4.
$T = t(X) = \bar{X}$ *is sufficient for* μ *when* $X_i \sim$ *i.i.d.* $\mathrm{N}(\mu, \sigma^2)$.

To better understand the motivation behind the concept of sufficiency consider three independent Binomial trials where $\theta = P(X = 1)$.

Event	Probability	Set
0 0 0	$(1-\theta)^3$	A_0

1 0 0
0 1 0
0 0 1
	$\theta(1-\theta)^2$	A_1

0 1 1
1 0 1
1 1 0
	$\theta^2(1-\theta)$	A_2

1 1 1	θ^3	A_3

Knowing which A_i the sample is in carries all the information about θ. Which particular sample within A_i gives us no extra information about θ. Extra information about other aspect of the model maybe, but not about θ. Here $T = t(X) = \sum X_i$ equals the number of "successes, and identifies A_i. Mathematically we can express the above concept by saying that the probability

$P(X=x|A_i)$ does not depend on θ. i.e. $P(010|A_1;\theta)=1/3$. More generally, a statistic $T=t(X)$ is said to be sufficient for the parameter θ if $\Pr(X=x|T=t)$ does not depend on θ. Sufficient statistics are most easily recognized through the following fundamental result:

Theorem 4.11 (Neyman's Factorization Criterion)**.** A statistic $T=t(X)$ is sufficient for θ if and only if the family of densities can be factorized as

$$f(x;\theta)=h(x)k\left\{t(x);\theta\right\},\ x\in\mathcal{X},\ \theta\in\Theta. \tag{4.6.1}$$

i.e. into a function which does not depend on θ and one which only depends on x through $t(x)$. This is true in general. We will prove it in the case where $\mathcal{X}$ is discrete.

Example 4.5 (Poisson).
Let $X=(X_1,\dots,X_n)$ be independent and Poisson distributed with mean λ so that the joint density

$$f(x;\lambda)\ =\ \prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}\,e^{-\lambda}\ =\ \frac{\lambda^{\Sigma x_i}}{\prod_i x_i!}e^{-n\lambda}.$$

Take $k\left\{\sum x_i;\lambda\right\}=\lambda^{\Sigma x_i}\ e^{-n\lambda}$ and $h(x)=\left(\prod x_i!\right)^{-1}$, then $t(x)=\sum_i x_i$ is sufficient.

Example 4.6 (Binomial).
Let $X=(X_1,\dots,X_n)$ be independent and Bernoulli distributed with parameter θ so that

$$f(x;\theta)\ =\ \prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}\ =\ \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

Take $k\left\{\sum x_i;\theta\right\}=\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$ and $h(x)=1$, then $t(x)=\sum_i x_i$ is sufficient.

Example 4.7 (Uniform).
Factorization criterion works in general but can mess up if the pdf depends on θ. Let $X=(X_1,\dots,X_n)$ be independent and Uniform distributed with parameter θ so that $X_1,X_2,\dots,X_n\sim\mathrm{Unif}(0,\theta)$. Then

$$f(x;\theta)=\frac{1}{\theta^n}\qquad 0\le x_i\le\theta\ \ \forall\ i.$$

It is not at all obvious but $t(x)=\max(x_i)$ is a sufficient statistic. We have to show that $f(x|t)$ is independent of θ. Well

$$f(x|t)=\frac{f(x,t)}{f_T(t)}.$$

Then

$$\begin{aligned} P\left(T\le t\right) &= P\left(X_1\le t,\dots,X_n\le t\right)\\ &= \prod_{i=1}^{n}P\left(X_i\le t\right)\ =\ \left(\frac{t}{\theta}\right)^n \end{aligned}$$

So

$$F_T(t)\ =\ \frac{t^n}{\theta^n}\qquad\Rightarrow\qquad f_T(t)\ =\ \frac{nt^{n-1}}{\theta^n}$$

Also

$$f(x,t)=\frac{1}{\theta^n}\ \equiv\ f(x;\theta).$$

Hence

$$f(x|t)\ =\ \frac{1}{nt^{n-1}},$$

and is independent of θ.

4.7 The Likelihood approach

4.7.1 Maximum likelihood estimation

Unless stated otherwise, the material in this section is mostly extracted from [8].

Let x be a realization of the random variable X with probability density $f_X(x|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$ is a vector of m unknown parameters to be estimated. The set of allowable values for $\boldsymbol{\theta}$, denoted by Φ, is called the *parameter space.* Define the likelihood function

$$L(\boldsymbol{\theta}|x) = f_X(x|\boldsymbol{\theta}). \tag{4.7.1}$$

It is crucial to stress that the argument of $f_X(x|\boldsymbol{\theta})$ is x, but the argument of $L(\boldsymbol{\theta}|x)$ is θ. It is therefore convenient to view the likelihood function $L(\theta)$ as the probability of the observed data x considered as a function of θ.

As explained in [5], the term **likelihood of the sample** needs to be defined, and this has to be done separately for discrete and continuous distributions.

Definition 4.17 ([5]). Let $x_1, x_2, \ldots, x_n$ be sample observations taken on the random variables $X_1, X_2, \ldots, X_n$. Then the likelihood of the sample, $L(\theta|x_1, x_2, \ldots, x_n)$, is defined as:

1. the joint probability of $x_1, x_2, \ldots, x_n$ if $X_1, X_2, \ldots, X_n$ are discrete, and
2. the joint probability density function of $X_1, \ldots, X_n$ evaluated at $x_1, x_2, \ldots, x_n$ if the random variables are continuous.

The **likelihood function** for a set of n identically and independently distributed (i.i.d.) random variables, $X_1, X_2, \ldots, X_n$, can thus be written as:

$$Ł(\theta; x_1, \ldots, x_n) = \begin{cases} P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n) & \text{for } X \text{ discrete,} \\ f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta) & \text{for } X \text{ continuous} \end{cases}. \tag{4.7.2}$$

For the discrete case, $L(\theta; x_1, \ldots, x_n)$ is the probability (or likelihood) of observing $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. It would then seem that a sensible approach to selecting an estimate of θ would be to find the value of θ which maximizes the probability of observing $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, (the event which occured). [5]

More generally, when the random continuous variables $X_1, \ldots, X_n$ are mutually independent we can write the joint density as

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{n} f_{X_j}(x_j)$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$ is a realization of the random vector $\mathbf{X} = (X_1, \ldots, X_n)'$, and the likelihood function becomes

$$L_{\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}) = \prod_{j=1}^{n} f_{X_j}(x_j|\boldsymbol{\theta}).$$

Note that we have used a bold θ to indicate that it may be a vector, although we will not be consistent in using different notations for vectors or single parameters. When the densities $f_{X_j}(x_j)$ are identical, we can unambiguously write $f(x_j)$.

Usually it is convenient to work with the natural logarithm of the likelihood called the log-likelihood, denoted by

$$\ell(\boldsymbol{\theta}|x) = \ln L(\boldsymbol{\theta}|x).$$

When $\theta \in \mathbb{R}^1$ we can define the *score function* as the first derivative of the log-likelihood

$$S(\theta) = \frac{\partial}{\partial\theta} \ln L(\theta),$$

as seen before.

The *maximum likelihood estimate* (MLE) $\hat{\theta}$ of θ is the solution to the score equation

$$S(\theta) = 0.$$

Put simply [5], the **maximum likelihood estimate** (MLE) of θ is that value of θ which maximizes the likelihood. To state it more mathematically, the MLE of θ is that value of θ, say $\widehat{\theta}$ such that

$$L(\widehat{\theta}; x_1, \dots, x_n) > L(\theta'; x_1, \dots, x_n)$$

where θ' is any other value of θ.

At the maximum [8], the second partial derivative of the log-likelihood is negative, so we define the curvature at $\hat{\theta}$ as $I(\hat{\theta})$ where

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2} \ln L(\theta).$$

We can check that a solution $\hat{\theta}$ of the equation $S(\theta) = 0$ is actually a maximum by checking that $I(\hat{\theta}) > 0$. A large curvature $I(\hat{\theta})$ is associated with a tight or strong peak, intuitively indicating less uncertainty about θ. In likelihood theory $I(\theta)$ is a key quantity called the observed Fisher information, and $I(\hat{\theta})$ is the observed Fisher information evaluated at the MLE $\hat{\theta}$. Although $I(\theta)$ is a function, $I(\hat{\theta})$ is a scalar. Note the difference with the concept of 'expected Fisher information' (See 4.6).

The likelihood function $L(\theta|x)$ supplies an order of preference or plausibility among possible values of θ based on the observed x. It ranks the plausibility of possible values of θ by how probable they make the observed x. If $P(x|\theta = \theta_1) > P(x|\theta = \theta_2)$ then the observed x makes $\theta = \theta_1$ more plausible than $\theta = \theta_2$, and consequently from (4.7.1), $L(\theta_1|x) > L(\theta_2|x)$. The likelihood ratio $L(\theta_1|x)/L(\theta_2|x) = f(\theta_1|x)/f(\theta_2|x)$ is a measure of the plausibility of θ_1 relative to θ_2 based on the observed data. The relative likelihood $L(\theta_1|x)/L(\theta_2|x) = k$ means that the observed value x will occur k times more frequently in repeated samples from the population defined by the value θ_1 than from the population defined by θ_2. Since only ratios of likelihoods are meaningful, it is convenient to standardize the likelihood with respect to its maximum.

Define the *relative likelihood* as $R(\theta|x) = L(\theta|x)/L(\hat{\theta}|x)$. The relative likelihood varies between 0 and 1. The MLE $\hat{\theta}$ is most plausible value of θ in that it makes the observed sample most probable. The relative likelihood measures the plausibility of any particular value of θ relative to that of $\hat{\theta}$.

Comments [5]

1. It is customary to use $\widehat{\theta}$ to denote both estimator (random variable) and estimate (its observed value). Recall that we used $\overline{\theta}$ for the MME.

2. Since $L(\theta; x_1, x_2, \ldots, x_n)$ is a product, and sums are usually more convenient to deal with than products, it is customary to maximize $\log L(\theta; x_1, \ldots, x_n)$ which we usually abbreviate to $l(\theta)$. This has the same effect. Since $\log L$ is a strictly increasing function of L, it will take on its maximum at the same point.

3. In some problems, θ will be a vector in which case $L(\theta)$ has to be maximized by differentiating with respect to 2 (or more) variables and solving simultaneously 2 (or more) equations.

4. The method of differentiation to find a maximum only works if the function concerned actually has a turning point.

4.7.2 Properties of MLE

The following four properties are the main reasons for recommending the use of Maximum Likelihood Estimators.

1. The MLE is consistent.

2. The MLE has a distribution that tends to normality as $n \to \infty$.

3. If a sufficient statistic for θ exists, then the MLE is sufficient.

4. The MLE is **invariant** under functional transformations. That is, if $\widehat{\theta} = H(X_1, X_2, \ldots, X_n)$ is the MLE of θ and if $u(\theta)$ is a continuous monotone function of θ, then $u(\widehat{\theta})$ is the MLE of $u(\theta)$. This is known as the **invariance property** of MLEs (see also 4.7.3).

   For example, in the normal distribution where the mean is μ and the variance is σ^2, $(n-1)S^2/n$ is the MLE of σ^2, so the MLE of σ is $\sqrt{(n-1)S^2/n}$.

Furthermore, suppose that an experiment consists of measuring random variables $x_1, x_2, \ldots, x_n$ which are i.i.d. with probability distribution depending on a parameter θ. Let $\hat{\theta}$ be the MLE of θ. Define

$$\begin{aligned} W_1 &= \sqrt{\mathrm{E}[I(\theta)]}(\hat{\theta} - \theta) \\ W_2 &= \sqrt{I(\theta)}(\hat{\theta} - \theta) \\ W_3 &= \sqrt{\mathrm{E}[I(\hat{\theta})]}(\hat{\theta} - \theta) \\ W_4 &= \sqrt{I(\hat{\theta})}(\hat{\theta} - \theta). \end{aligned}$$

Then, W_1, W_2, W_3, and W_4 are all random variables and, as $n \to \infty$, the probabilistic behaviour of each of W_1, W_2, W_3, and W_4 is well approximated by that of a $N(0,1)$ random variable (see further below to see a hint of the proof). Then, since $\mathrm{E}[W_1] \approx 0$, we have that $\mathrm{E}[\hat{\theta}] \approx \theta$ and so $\hat{\theta}$ is approximately unbiased. Also $\mathrm{Var}[W_1] \approx 1$ implies that $\mathrm{Var}[\hat{\theta}] \approx (\mathrm{E}[I(\theta)])^{-1}$ and so $\hat{\theta}$ is approximately efficient.

Let the data $\mathbf{X}$ have probability distribution $g(\mathbf{X}; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ is a vector of m unknown parameters. Let $\mathbf{I}(\boldsymbol{\theta})$ be the $m \times m$ information matrix as defined above and let $\mathrm{E}[\mathbf{I}(\boldsymbol{\theta})]$ be the $m \times m$ matrix obtained by replacing the elements of $\mathbf{I}(\boldsymbol{\theta})$ by their expected values. Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$. Let $CRLB_r$ be the rth diagonal element of $[\mathbf{E}[\mathbf{I}(\boldsymbol{\theta})]]^{-1}$. For $r = 1, 2, \ldots, m$, define $W_{1r} = (\hat{\theta}_r - \theta_r)/\sqrt{CRLB_r}$. Then, as $n \to \infty, W_{1r}$ behaves like a standard normal random variable.

Suppose we define W_{2r} by replacing $CRLB_r$ by the rth diagonal element of the matrix $[I(\boldsymbol{\theta})]^{-1}, W_{3r}$ by replacing $CRLB_r$ by the rth diagonal element of the matrix $[\mathrm{E}I(\hat{\boldsymbol{\theta}})]^{-1}$ and W_{4r} by replacing $CRLB_r$ by the rth diagonal element of the matrix $[I(\hat{\boldsymbol{\theta}})]^{-1}$. Then it can be shown that as $n \to \infty, W_{2r}, W_{3r}$, and W_{4r} all behave like standard normal random variables. These results will become very handy when developing test statistics (Chapter refchap:testing)

A detailed proof of the assertion that W_i above follow approximately a standard normal distribution is beyond the scope of this course. However, in what follows, we give a hint of how the proof would go. Suppose that $\widehat{\theta}$ is the MLE of θ. We then apply the Taylor expansion of $S(\widehat{\theta})$ at the point θ_0, yielding

$$0 = l'(\widehat{\theta}) \approx l^{'}(\theta_0) + (\widehat{\theta} - \theta_0)l^{''}(\theta_0).$$

Therefore,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \approx \frac{-n^{-1/2}l^{'}(\theta_0)}{n^{-1}l^{''}(\theta_0)}$$

. Based on properties we already saw, it can then be shown that the mean of the numerator is 0 and its variance is $\mathrm{E}(I(\theta_0))$. By the law of large numbers, the expectation of the denomenator converges to $-\mathrm{E}(I(\theta_0))$. We thus have

$$\sqrt{n}(\widehat{\theta} - \theta_0) \approx \frac{n^{-1/2}l^{'}(\theta_0)}{\mathrm{E}(I(\theta_0))}.$$

Consequently,

$$\mathrm{E}(\sqrt{n}(\widehat{\theta} - \theta_0)) \approx 0,$$

and

$$\mathrm{Var}(\sqrt{n}(\widehat{\theta} - \theta_0)) \approx \frac{1}{\mathrm{E}(I(\theta_0))}.$$

4.7.3 The Invariance principle

How do we deal with parameter transformation? We will assume a one-to-one transformation, but the idea applied generally. Consider a binomial sample with $n = 10$ independent trials resulting in data $x = 8$ successes. The likelihood ratio of $\theta_1 = 0.8$ versus $\theta_2 = 0.3$ is

$$\frac{L(\theta_1 = 0.8)}{L(\theta_2 = 0.3)} = \frac{\theta_1^8(1-\theta_1)^2}{\theta_2^8(1-\theta_2)^2} = 208.7 \ ,$$

that is, given the data $\theta = 0.8$ is about 200 times more likely than $\theta = 0.3$.

Suppose we are interested in expressing θ on the logit scale as

$$\psi \equiv \ln\{\theta/(1-\theta)\} \ ,$$

then 'intuitively' our relative information about $\psi_1 = \ln(0.8/0.2) = 1.29$ versus $\psi_2 = \ln(0.3/0.7) = -0.85$ should be

$$\frac{L^*(\psi_1)}{L^*(\psi_2)} = \frac{L(\theta_1)}{L(\theta_2)} = 208.7 \ .$$

That is, our information should be *invariant* to the choice of parameterization. (For the purposes of this example we are not too concerned about how to calculate $L^*(\psi)$.)

Theorem 4.12 (Invariance of the MLE)**.** *If g is a one-to-one function, and $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.*

Proof. This is trivially true as we let $\theta = g^{-1}(\mu)$ then $f\{y|g^{-1}(\mu)\}$ is maximized in μ exactly when $\mu = g(\hat{\theta})$. When g is not one-to-one the discussion becomes more subtle, but we simply choose to define $\hat{g}_{\text{MLE}}(\theta) = g(\hat{\theta})$ □

It seems intuitive that if $\hat{\theta}$ is most likely for θ and our knowledge (data) remains unchanged then $g(\hat{\theta})$ is most likely for $g(\theta)$. In fact, we would find it strange if $\hat{\theta}$ is an estimate of θ, but $\hat{\theta}^2$ is not an estimate of θ^2. In the binomial example with $n = 10$ and $x = 8$ we get $\hat{\theta} = 0.8$, so the MLE of $g(\theta) = \theta/(1-\theta)$ is

$$g(\hat{\theta}) = \hat{\theta}/(1-\hat{\theta}) = 0.8/0.2 = 4.$$

This convenient property is not necessarily true of other estimators. For example, if $\hat{\theta}$ is the MVUE of θ, then $g(\hat{\theta})$ is generally not MVUE for $g(\theta)$.

Frequentists generally accept the invariance principle without question. This is not the case for Bayesians. The invariance property of the likeihood ratio is incompatible with the Bayesian habit of assigning a probability distribution to a parameter.

Examples

All the examples are coming from [8], unless stated otherwise.

Example 4.8 ([5]).
Given X is distributed $\text{bin}(1, p)$ *where $p \in (0, 1)$, and a random sample $x_1, x_2, \ldots, x_n$, find the maximum likelihood estimate of p.*

Solution of Example 4.3 ([5])**.** *The likelihood is,*

$$\begin{aligned} L(p; x_1, x_2, \ldots, x_n) &= P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n) \\ &= \prod_{i=1}^{n} \binom{1}{x_i} p^{x_i}(1-p)^{1-x_i} \\ &= p^{x_1+x_2+\cdots+x_n}(1-p)^{n-x_1-x_2-\cdots-x_n} \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i} \end{aligned}$$

So

$$\log L(p) = \sum x_i \log p + \left(n - \sum x_i\right)\log(1-p)$$

Differentiating with respect to p, we have

$$d\frac{\log L(p)}{dp} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p}$$

This is equal to zero when $\sum x_i(1-p) = p\,(n - \sum x_i)$, that is, when $p = \sum x_i/n$. This estimate is denoted by $\widehat{p}$. Thus, if the random variable X is distributed $\text{bin}(1,p)$, the MLE of p derived from a sample of size n is

$$\widehat{p} = \overline{X}. \tag{4.7.3}$$

Example 4.9 (Binomial sampling).
The number of successes in n Bernoulli trials is a random variable R taking on values $r = 0, 1, \dots, n$ with probability mass function

$$P(R = r) = \binom{n}{r}\theta^r(1-\theta)^{n-r}.$$

This is the exact same sampling scheme as in the previous example except that instead of observing the sequence $\mathbf{y}$ we only observe the total number of successes r. Hence the likelihood function has the form

$$L_R\,(\theta|r) = \binom{n}{r}\theta^r(1-\theta)^{n-r}.$$

The relevant mathematical calculations are as follows:

$$\begin{aligned}
\ell_R\,(\theta|r) &= \ln\binom{n}{r} + r\ln(\theta) + (n-r)\ln(1-\theta) \\
S\,(\theta) &= \frac{r}{n} + \frac{n-r}{1-\theta} \qquad\qquad \Rightarrow\ \hat{\theta} = \frac{r}{n} \\
I\,(\theta) &= \frac{r}{\theta^2} + \frac{n-r}{(1-\theta)^2}\ > 0 \qquad \forall\ \theta \\
\mathrm{E}(\hat{\theta}) &= \frac{\mathrm{E}(r)}{n} = \frac{n\theta}{n} = \theta \qquad \Rightarrow\ \hat{\theta}\ \textit{unbiased} \\
\mathrm{Var}(\hat{\theta}) &= \frac{\mathrm{Var}(r)}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} \\
\mathrm{E}\left[I\,(\theta)\right] &= \frac{\mathrm{E}(r)}{\theta^2} + \frac{n-\mathrm{E}(r)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} \\
&= \frac{n}{\theta(1-\theta)} = \left(\mathrm{Var}[\hat{\theta}]\right)^{-1}
\end{aligned}$$

and $\hat{\theta}$ attains the Cramer-Rao lower bound (CRLB).

Example 4.10 ([5]).
Given random variable X is distributed uniformly on $[0,\theta]$, find the MLE of θ based on a sample of size n.

Solution of Example 4.4 ([5]). *Now $f(x_i;\theta) = 1/\theta$, $x_i \in [0,\theta]$, $i = 1, 2, \dots, n$. So the likelihood is*

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^{n}(1/\theta) = 1/\theta^n.$$

When we come to find the maximum of this function we note that the slope is not zero anywhere, so there is no use finding $\frac{dL(\theta)}{d\theta}$ or $\frac{d\log L(\theta)}{d\theta}$.

Note however that $L(\theta)$ increases as $\theta \to 0$. So $L(\theta)$ is maximized by setting θ equal to the smallest value it can take. If the observed values are $x_1, \dots, x_n$ then θ can be no smaller than the largest of these. This is because $x_i \in [0,\theta]$ for $i = 1, \dots, n$. That is, each $x_i \le \theta$ or $\theta \ge$ each x_i.

Thus, if X is distributed $U(0,\theta)$, the MLE of θ is

$$\widehat{\theta} = \max(X_i). \tag{4.7.4}$$

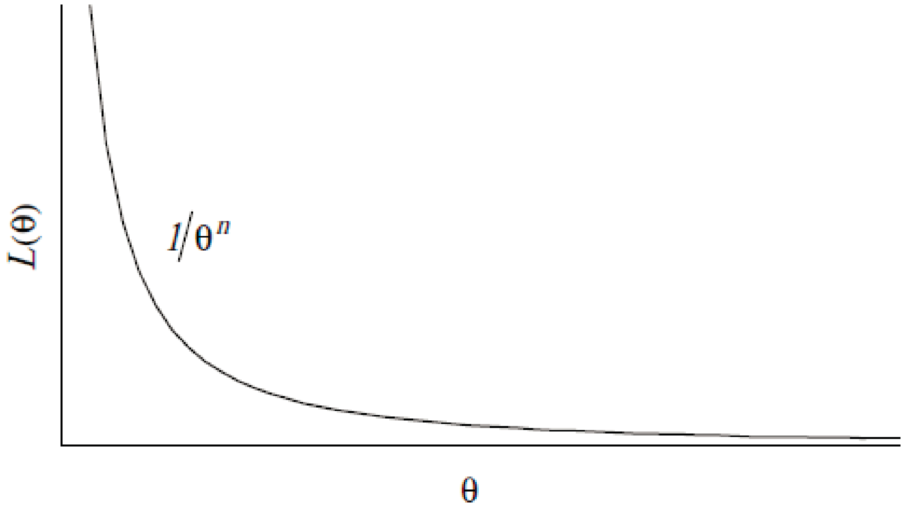


Figure 4.7.1: $L(\theta) = 1/\theta^n$

Computer Exercise 4.3. *Generate 100 random samples of size 10 from a* $U(0, 10)$ *distribution. For each of the 100 samples generated calculate the MME and MLE for* μ *and graph the results.*

1. *From the graphs does it appear that the estimators are biased or unbiased? Explain.*
2. *Estimate the variance of the two estimators by finding the sample variance of the 100 estimates (for each estimator). Which estimator appears more efficient?*

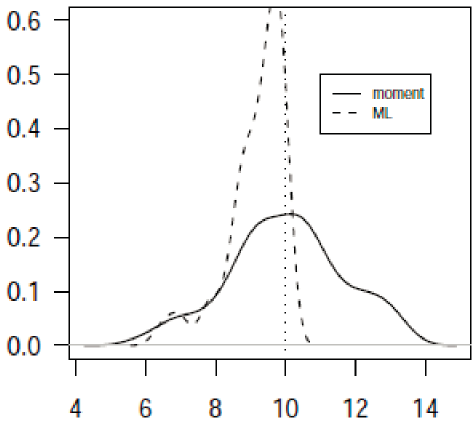
Solution of Computer Exercise 4.3.

```
theta <- 10
sampsz <- 10
nsimulations <- 100
moment.estimates <- numeric(nsimulations)
ML.estimates <- numeric(nsimulations)

for (i in 1:nsimulations){
ru <- runif(n=sampsz,min=0,max=theta)

 moment.estimates[i] <- 2*mean(ru)
 ML.estimates[i] <- max(ru)
}

plot(density(moment.estimates),
 xlab=" ",ylab=" ",main=" ",ylim=c(0,0.6),
las=1)
abline(v=theta,lty=3)
lines(density(ML.estimates),lty=2)
legend(11,0.5,legend=c("moment","ML"),lty=1:2,
cex=0.6)
```



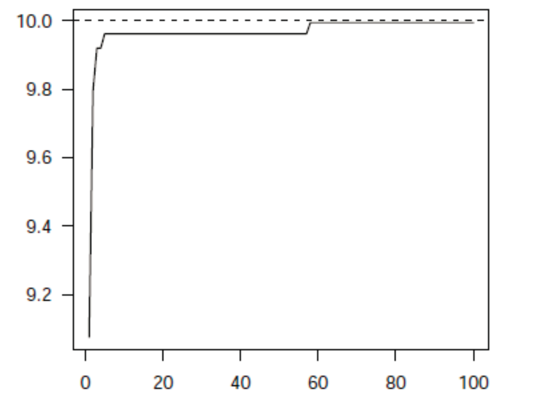
You should see that the Method of Moments gives unbiased estimates of which many are not in the range space as noted in Computer Example 4.2. The maximum likelihood estimates are all less than 10 and so are biased.

Computer Exercise 4.4. *Demonstrate that the MLE is consistent for estimating θ for a $U(0,\theta)$ distribution.*

Method: Generate the random variables one at a time. After each is generated calculate the MLE of θ for the sample of size n generated to that point and so obtain a sequence of estimators, $\{\theta_n\}$. Plot the sequence.

Solution of Computer Exercise 4.4. *Uniform random variables are generated one at a time and $\widehat{\theta}_n$ is found as the maximum of $\widehat{\theta_{n-1}}$ and nth uniform rv generated. The estimates are plotted in order.*

```
#_________ UniformConsistency.R _____
theta <- 10
sampsz <- 10
nsimulations <- 100
ML.est <- numeric(nsimulations)
for (i in 1:nsimulations){
  ru <- runif(n=sampsz,min=0,max=theta)
 if(i==1) ML.est[i] <- max(ru)
else ML.est[i] <- max(ML.est[i-1],max(ru) )
}
plot(ML.est,type='l')
abline(h=theta,lty=2)
```



We can see that as n increases, $\widehat{\theta} \to \theta$.

Example 4.11 (Bernoulli Trials).
Consider n independent Bernoulli trials. The jth observation is either a "success" or "failure" coded $x_j = 1$ and $x_j = 0$ respectively, and

$$P(X_j = x_j) = \theta^{x_j}(1-\theta)^{1-x_j}$$

for $j = 1, \ldots, n$. The vector of observations $\mathbf{y} = (x_1, x_2, \ldots, x_n)^T$ is a sequence of ones and zeros, and is a realization of the random vector $\mathbf{Y} = (X_1, X_2, \ldots, X_n)^T$. As the Bernoulli outcomes are assumed to be independent we can write the joint probability mass function of $\mathbf{Y}$ as the product of the marginal probabilities, that is

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n} P(X_j = x_j) \\ &= \prod_{j=1}^{n} \theta^{x_j}(1-\theta)^{1-x_j} \\ &= \theta^{\sum x_j}(1-\theta)^{n-\sum x_j} \\ &= \theta^{r}(1-\theta)^{n-r} \end{aligned}$$

where $r = \sum_{i=1}^{n} x_j$ is the number of observed successes (1's) in the vector $\mathbf{y}$. The log-likelihood function is then

$$\ell(\theta) = r \ln\theta + (n-r)\ln(1-\theta),$$

and the score function is

$$S(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \frac{r}{\theta} - \frac{(n-r)}{1-\theta}.$$

Figure 4.7.2: Relative likelihood functions for seed germinating probabilities.

Solving for $S(\hat{\theta}) = 0$ *we get* $\hat{\theta} = r/n$. *The information function is*

$$I(\theta) = \frac{r}{\theta^2} + \frac{n-r}{(1-\theta)^2} > 0 \quad \forall \ \theta,$$

guaranteeing that $\hat{\theta}$ *is the MLE. Each* X_i *is a Bernoulli random variable and has expected value* $E(X_i) = \theta$, *and variance* $Var(X_i) = \theta(1-\theta)$. *The MLE* $\hat{\theta}(\mathbf{y})$ *is itself a random variable and has expected value*

$$E(\hat{\theta}) = E\left(\frac{r}{n}\right) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}\sum_{i=1}^n \theta = \theta,$$

implying that $\hat{\theta}(\mathbf{y})$ *is an unbiased estimator of* θ. *The variance of* $\hat{\theta}(\mathbf{y})$ *is*

$$Var(\hat{\theta}) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{1}{n^2}\sum_{i=1}^n (1-\theta)\theta = \frac{(1-\theta)\theta}{n}.$$

Finally, note that

$$\mathcal{I}(\theta) = \mathrm{E}\left[I(\theta)\right] = \frac{\mathrm{E}(r)}{\theta^2} + \frac{n-\mathrm{E}(r)}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)} = \left(\mathrm{Var}[\hat{\theta}]\right)^{-1},$$

and $\hat{\theta}$ *attains the Cramer-Rao lower bound (CRLB).*

Example 4.12 (Germinating seeds).
Suppose 25 seeds were planted and $r = 5$ *seeds germinated. Then* $\hat{\theta} = r/n = 0.2$ *and* $\mathrm{Var}(\hat{\theta}) = 0.2 \times 0.8/25 = 0.0064$. *The relative likelihood is*

$$R_1(\theta) = \left(\frac{\theta}{0.2}\right)^5 \left(\frac{1-\theta}{0.8}\right)^{20}.$$

Suppose 100 seeds were planted and $r = 20$ *seeds germinated. Then* $\hat{\theta} = r/n = 0.2$ *but* $\mathrm{Var}(\hat{\theta}) = 0.2 \times 0.8/100 = 0.0016$. *The relative likelihood is*

$$R_2(\theta) = \left(\frac{\theta}{0.2}\right)^{20} \left(\frac{1-\theta}{0.8}\right)^{80}.$$

Suppose 25 seeds were planted and it is known only that $r \leq 5$ *seeds germinated. In this case the exact number of germinating seeds is unknown. The information about* θ *is given by the likelihood function*

$$L(\theta) \ = \ P(R \leq 5) \ = \ \sum_{r=0}^{5} \binom{25}{r} \theta^r (1-\theta)^{25-r}.$$

Here, the most plausible value for θ *is* $\hat{\theta} = 0$, *implying* $L(\hat{\theta}) = 1$. *The relative likelihood is* $R_3(\theta) = L(\theta)/L(\hat{\theta}) = L(\theta)$. $R_1(\theta)$ *is plotted as the dashed curve in figure 4.7.2.* $R_2(\theta)$ *is plotted as the dotted curve in figure 4.7.2.* $R_3(\theta)$ *is plotted as a solid curve in figure 4.7.2.*

Example 4.13 (Prevalence of a Genotype).
Geneticists interested in the prevalence of a certain genotype, observe that the genotype makes its first appearance in the 22nd subject analysed. If we assume that the subjects are independent, the likelihood function can be computed based on the geometric distribution, as $L(\theta) = (1-\theta)^{n-1}\theta$. *The score function is then* $S(\theta) = \theta^{-1} - (n-1)(1-\theta)^{-1}$. *Setting* $S(\hat{\theta}) = 0$ *we get* $\hat{\theta} = n^{-1} = 22^{-1}$. *The observed Fisher information equals* $I(\theta) = \theta^{-2} + (n-1)(1-\theta)^{-2}$ *and is greater than zero for all* θ, *implying that* $\hat{\theta}$ *is MLE.*

Suppose that the geneticists had planned to stop sampling once they observed $r = 10$ *subjects with the specified genotype, and the tenth subject with the genotype was the 100th subject anaylsed overall. The likelihood of* θ *can be computed based on the negative binomial distribution, as*

$$L(\theta) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}$$

for $n = 100,\ r = 5$. *The usual calculation will confirm that* $\hat{\theta} = r/n$ *is MLE.*

Example 4.14 (Radioactive Decay).
In this classic set of data Rutherford and Geiger counted the number of scintillations in 72 second intervals caused by radioactive decay of a quantity of the element polonium. Altogether there were 10097 scintillations during 2608 such intervals:

Count	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Observed	*57*	*203*	*383*	*525*	*532*	*408*	*573*	*139*
Count	*8*	*9*	*10*	*11*	*12*	*13*	*14*	
Observed	*45*	*27*	*10*	*4*	*1*	*0*	*1*	

The Poisson probability mass function with mean parameter θ *is*

$$f_X(x|\theta) = \frac{\theta^x \exp(-\theta)}{x!}.$$

The likelihood function equals

$$L(\theta) = \prod \frac{\theta^{x_i}\exp(-\theta)}{x_i!} = \frac{\theta^{\sum x_i}\exp(-n\theta)}{\prod x_i!}.$$

The relevant mathematical calculations are

$$\begin{aligned}
\ell(\theta) &= (\Sigma x_i)\ln(\theta) - n\theta - \ln[\Pi(x_i!)] \\
S(\theta) &= \frac{\sum x_i}{\theta} - n \\
\Rightarrow\ \hat{\theta} &= \frac{\sum x_i}{n} = \bar{x} \\
I(\theta) &= \frac{\Sigma x_i}{\theta^2} > 0, \qquad \forall\ \theta
\end{aligned}$$

implying $\hat{\theta}$ *is MLE. Also* $\mathrm{E}(\hat{\theta}) = \sum \mathrm{E}(x_i) = \frac{1}{n}\sum\theta = \theta$, *so* $\hat{\theta}$ *is an unbiased estimator. Next* $\mathrm{Var}(\hat{\theta}) = \frac{1}{n^2}\sum \mathrm{Var}(x_i) = \frac{1}{n}\theta$ *and* $\mathcal{I}(\theta) = \mathrm{E}[I(\theta)] = n/\theta = (\mathrm{Var}[\hat{\theta}])^{-1}$ *implying that* $\hat{\theta}$ *attains the theoretical CRLB. It is always useful to compare the fitted values from a model against the observed values.*

i	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
O_i	*57*	*203*	*383*	*525*	*532*	*408*	*573*	*139*	*45*	*27*	*10*	*4*	*1*	*0*	*1*
E_i	*54*	*211*	*407*	*525*	*508*	*393*	*254*	*140*	*68*	*29*	*11*	*4*	*1*	*0*	*0*
	$+3$	-8	-24	0	$+24$	$+15$	$+19$	-1	-23	-2	-1	0	-1	$+1$	$+1$

The Poisson law agrees with the observed variation within about one-twentieth of its range.

Example 4.15 (Exponential distribution).
Suppose random variables $X_1, \ldots, X_n$ are i.i.d. as Exp(θ). Then

$$\begin{array}{rrcl} & L(\theta) & = & \prod_{i=1}^{n} \theta \exp\left(-\theta x_i\right) \\ & & = & \theta^n \exp\left(-\theta \sum x_i\right) \\ & \ell(\theta) & = & n \ln \theta - \theta \sum x_i \\ & S(\theta) & = & \frac{n}{\theta} - \sum_{i=1}^{n} x_i \\ \Rightarrow & \hat{\theta} & = & \frac{n}{\sum x_i} \\ & I(\theta) & = & \frac{n}{\theta^2} > 0 \qquad \forall\ \theta. \end{array}$$

In order to work out the expectation and variance of $\hat{\theta}$ we need to work out the probability distribution of $Z = \sum_{i=1}^{n} X_i$, where $X_i \sim$ Exp(θ). From the appendix on probability theory we have $Z \sim$ Ga(θ, n). Then

$$\begin{array}{rcl} \mathrm{E}\left[\frac{1}{Z}\right] & = & \int\limits_0^\infty \frac{1}{z} \frac{\theta^n z^{n-1} \exp\left(-\theta z\right)}{\Gamma(n)} dz \\ & = & \frac{\theta^2}{\Gamma(n)} \int\limits_0^\infty (\theta z)^{n-2} \exp\left(-\theta z\right) dz \\ & = & \frac{\theta}{\Gamma(n)} \int\limits_0^\infty u^{n-2} \exp\left(-u\right) du \\ & = & \frac{\theta\ \Gamma(n-1)}{\Gamma(n)} = \frac{\theta}{n-1}. \end{array}$$

We now return to our estimator

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{n}{Z}$$

implies

$$\mathrm{E}[\hat{\theta}] = \mathrm{E}\left[\frac{n}{Z}\right] = n\mathrm{E}\left[\frac{1}{Z}\right] = \frac{n}{n-1}\theta$$

which turns out to be biased. Propose the alternative estimator $\tilde{\theta} = \frac{n-1}{n}\hat{\theta}$. Then

$$\mathrm{E}[\tilde{\theta}] = \mathrm{E}\left[\frac{(n-1)}{n}\hat{\theta}\right] = \frac{(n-1)}{n}\mathrm{E}[\hat{\theta}] = \frac{(n-1)}{n}\left(\frac{n}{n-1}\right)\theta = \theta$$

shows $\tilde{\theta}$ is an unbiased estimator.

As this example demonstrates, maximum likelihood estimation does not automatically produce unbiased estimates. If it is thought that this property is (in some sense) desirable, then some adjustments to the MLEs, usually in the form of scaling, may be required. We conclude this example with the following tedious (but *straightforward) calculations.*

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{Z^2}\right] &= \frac{1}{\Gamma(n)}\int_0^\infty \theta^n z^{n-3}\exp -\theta z\, dz \\
&= \frac{\theta^2}{\Gamma(n)}\int_0^\infty u^{n-3}\exp -\theta u\, du \\
&= \frac{\theta^2\Gamma(n-2)}{\Gamma(n)} = \frac{\theta^2}{(n-1)(n-2)} \\
\Rightarrow\ \mathrm{Var}[\tilde{\theta}] &= \mathrm{E}[\tilde{\theta}^2] - \Big(\mathrm{E}[\tilde{\theta}]\Big)^2 = \mathrm{E}\left[\frac{(n-1)^2}{Z^2}\right] - \theta^2 \\
&= \frac{(n-1)^2\theta^2}{(n-1)(n-2)} - \theta^2 = \frac{\theta^2}{n-2}.
\end{aligned}$$

We have already calculated that

$$I(\theta) = \frac{n}{\theta^2} \qquad \Rightarrow \qquad \mathrm{E}\left[I(\theta)\right] = \frac{n}{\theta^2} \neq \Big(\mathrm{Var}[\tilde{\theta}]\Big)^{-1}.$$

However,

$$\mathit{eff}(\tilde{\theta}) = \frac{\left(\mathrm{E}\left[I(\theta)\right]\right)^{-1}}{\mathrm{Var}[\tilde{\theta}]} = \frac{\theta^2}{n} \div \frac{\theta^2}{n-2} = \frac{n-2}{n}$$

which although not equal to 1, converges to 1 as $n \to \infty$, and $\tilde{\theta}$ is asymptotically efficient.

Example 4.16 (Lifetime of a component).
The time to failure T of components has an exponential distributed with mean μ. Suppose n components are tested for 100 hours and that m components failed at times $t_1, \ldots, t_m$, with $n-m$ components surviving the 100 hour test. The likelihood function can be written

$$L(\theta) = \underbrace{\prod_{i=1}^{m} \frac{1}{\mu} e^{-t_i/\mu}}_{\textit{components failed}}\ \underbrace{\prod_{j=m+1}^{n} P(T_j > 100)}_{\textit{components survived}}.$$

Clearly $P(T \leq t) = 1 - e^{-t/\mu}$ implies $P(T > 100) = e^{-100/\mu}$ is the probability of a component surviving the 100 hour test. Then

$$\begin{aligned}
L(\mu) &= \left(\prod_{i=1}^{m} \frac{1}{\mu} e^{-t_i/\mu}\right)\Big(e^{-100/\mu}\Big)^{n-m}, \\
\ell(\mu) &= -m\ln\mu - \frac{1}{\mu}\sum_{i=1}^{m} t_i - \frac{1}{\mu}100(n-m), \\
S(\mu) &= -\frac{m}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{m} t_i + \frac{1}{\mu^2}100(n-m).
\end{aligned}$$

Setting $S(\hat{\mu}) = 0$ suggests the estimator $\hat{\mu} = \left[\sum_{i=1}^{m} t_i + 100\ (n-m)\right]/m$. Also, $I(\hat{\mu}) = m/\hat{\mu}^2 > 0$, and $\hat{\mu}$ is indeed the MLE. Although failure times were recorded for just m components, this example usefully demonstrates that all n components contribute to the estimation of the mean failure parameter μ. The $n-m$ surviving components are often referred to as right censored.

Example 4.17 (Gaussian Distribution).
Consider data $X_1, X_2 \ldots, X_n$ *distributed as* $N(\mu, \upsilon)$. *Then the likelihood function is*

$$L(\mu,\upsilon) = \left(\frac{1}{\sqrt{\pi\upsilon}}\right)^n \exp\left\{-\frac{\sum\limits_{i=1}^n (x_i-\mu)^2}{2\upsilon}\right\}$$

and the log-likelihood function is

$$\ell(\mu,\upsilon) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\upsilon) - \frac{1}{2\upsilon}\sum_{i=1}^n (x_i-\mu)^2 \qquad (4.7.5)$$

Unknown mean and known variance: *As* υ *is known we treat this parameter as a constant when differentiating wrt* μ. *Then*

$$S(\mu) = \frac{1}{\upsilon}\sum_{i=1}^n (x_i-\mu), \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \textit{and} \qquad I(\theta) = \frac{n}{\upsilon} \ > 0 \ \forall \ \mu.$$

Also, $\mathrm{E}[\hat{\mu}] = n\mu/n = \mu$, *and so the MLE of* μ *is unbiased. Finally*

$$\mathrm{Var}[\hat{\mu}] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^n x_i\right] = \frac{\upsilon}{n} = (\mathrm{E}[I(\theta)])^{-1}.$$

Known mean and unknown variance: *Differentiating (4.7.5) wrt* υ *returns*

$$S(\upsilon) = -\frac{n}{2\upsilon} + \frac{1}{2\upsilon^2}\sum_{i=1}^n (x_i-\mu)^2,$$

and setting $S(\upsilon) = 0$ *implies*

$$\hat{\upsilon} = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2.$$

Differentiating again, and multiplying by -1 *yields the information function*

$$I(\upsilon) = -\frac{n}{2\upsilon^2} + \frac{1}{\upsilon^3}\sum_{i=1}^n (x_i-\mu)^2.$$

Clearly $\hat{\upsilon}$ *is the MLE since*

$$I(\hat{\upsilon}) = \frac{n}{2\upsilon^2} \ > \ 0.$$

Define

$$Z_i = (X_i-\mu)^2/\sqrt{\upsilon},$$

so that $Z_i \sim N(0,1)$. *From the appendix on probability*

$$\sum_{i=1}^n Z_i^2 \sim \chi^2_n,$$

implying $\mathrm{E}[\sum Z_i^2] = n$, *and* $\mathrm{Var}[\sum Z_i^2] = 2n$. *The MLE*

$$\hat{\upsilon} = (\upsilon/n)\sum_{i=1}^n Z_i^2.$$

Then

$$\mathrm{E}[\hat{\upsilon}] = \mathrm{E}\left[\frac{\upsilon}{n}\sum_{i=1}^n Z_i^2\right] = \upsilon,$$

and

$$\mathrm{Var}[\hat{\upsilon}] = \left(\frac{\upsilon}{n}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} Z_i^2\right] = \frac{2\upsilon^2}{n}.$$

Finally,

$$\begin{aligned} \mathrm{E}\left[I(\upsilon)\right] &= -\frac{n}{2\upsilon^2} + \frac{1}{\upsilon^3}\sum_{i=1}^{n}\mathrm{E}\left[(x_i-\mu)^2\right] \\ &= -\frac{n}{2\upsilon^2} + \frac{n\upsilon}{\upsilon^3} \\ &= \frac{n}{2\upsilon^2}. \end{aligned}$$

Hence the CRLB $= 2\upsilon^2/n$*, and so* $\hat{\upsilon}$ *has efficiency 1.*

Our treatment of the two parameters of the Gaussian distribution in the last example was to (i) fix the variance and estimate the mean using maximum likelihood; and then (ii) fix the mean and estimate the variance using maximum likelihood.

In practice we would like to consider the simultaneous estimation of these parameters. In the next section of these notes we extend MLE to multiple parameter estimation.

4.8 Properties of Sample Mean and Sample Variance

In this section we will consider the sample mean $\overline{X}$ and the sample variance S^2 and examine which of the above properties they have.

Theorem 4.13. *Let X be a random variable with mean μ and variance σ^2. Let $\overline{X}$ be the sample mean based on a random sample of size n. Then $\overline{X}$ is an unbiased and consistent estimator of μ.*

Proof. Now $E(\overline{X}) = \mu$, no matter what the sample size is, and $Var(\overline{X}) = \sigma^2/n$. The latter approaches 0 as $n \to \infty$, satisfying Theorem 4.10. □

It can also be shown that of all linear functions of $X_1, X_2, \ldots, X_n$, $\overline{X}$ has minimum variance. Note that the above theorem is true no matter what distribution is sampled. Some applications are given below.

For a random sample $X_1, X_2, \ldots, X_n$, $\overline{X}$ is an unbiased and consistent estimator of:

1. μ when the X_i are distributed $N(\mu, \sigma^2)$;

2. p when the X_i are distributed bin$(1, p)$;

3. λ when the X_i are distributed Poisson(λ);

4. $1/\alpha$ when the X_i have p.d.f. $f(x) = \alpha e^{-\alpha x}$, $x > 0$.

Sample Variance

Recall that the sample variance is defined by

$$S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1).$$

Theorem 4.14. *Given $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and variance σ^2, then S^2 is an unbiased and consistent estimator of σ^2.*

Proof.

$$\begin{aligned}
(n-1)E(S^2) &= E\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= E\sum_{i=1}^{n}[X_i - \mu - (\overline{X} - \mu)[^2 \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\overline{X} - \mu)^2 + n(\overline{X} - \mu)^2\right] \\
&= E\sum_{i=1}^{n}(X_i - \mu)^2 - nE(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}Var(X_i) - nVar(\overline{X}) \\
&= n\sigma^2 - n\frac{\sigma^2}{n} \\
&= (n-1)\sigma^2.
\end{aligned}$$

So

$$E(S^2) = \sigma^2, \tag{4.8.1}$$

which proves the unbiasedness.

For any distribution that has a fourth moment,

$$Var(S^2) = \frac{\mu_4 - 3\mu_2^2}{n} - \frac{2\mu_2^2}{n-1} \tag{4.8.2}$$

Clearly $\lim_{n\to\infty} Var(S^2) = 0$, so from Theorem 4.10, S^2 is a consistent estimator of σ^2.

□

We make the following comments.

1. In the special case of theorem 4.14 where the X_i are distributed $N(\mu, \sigma^2)$ with both μ and σ^2 unknown, the MLE of σ^2 is $\sum_{i=1}^{n}(X_i - \overline{X})^2/n$ which is $(n-1)S^2/n$. So in this case the MLE is biased.

2. The number in the denominator of S^2, that is, $n-1$, is called the **number of degrees of freedom**. The numerator is the sum of n deviations (from the mean) squared but the deviations are not independent. There is one constraint on them, namely the fact that $\sum(X_i - \overline{X}) = 0$. As soon as $n-1$ of the $X_i - \overline{X}$ are known, the nth one is determined.

3. In calculating the observed value of S^2, s^2, the following form is usually convenient.

$$s^2 = \left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right] \Big/ (n-1) \tag{4.8.3}$$

or, equivalently,

$$s^2 = \frac{\sum x_i^2 - n\overline{x}^2}{n-1}$$

The equivalence of the two forms is easily seen:

$$\sum(x_i - \overline{x})^2 = \sum(x_i^2 - 2\overline{x}x_i + \overline{x}^2) = \sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2$$

where the RHS can readily be seen to be $\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}$.

4.9 Multi-parameter Estimation

Suppose that a statistical model specifies that the data $\mathbf{y}$ has a probability distribution $f(\mathbf{y}; \alpha, \beta)$ depending on two unknown parameters α and β. In this case the likelihood function is a function of the two variables α and β and having observed the value $\mathbf{y}$ is defined as $L(\alpha, \beta) = f(\mathbf{y}; \alpha, \beta)$ with $\ell(\alpha, \beta) = \ln L(\alpha, \beta)$. The MLE of (α, β) is a value $(\hat{\alpha}, \hat{\beta})$ for which $L(\alpha, \beta)$, or equivalently $\ell(\alpha, \beta)$, attains its maximum value.

Define $S_1(\alpha, \beta) = \partial\ell/\partial\alpha$ and $S_2(\alpha, \beta) = \partial\ell/\partial\beta$. The MLEs $(\hat{\alpha}, \hat{\beta})$ can be obtained by solving the pair of simultaneous equations :

$$\begin{aligned} S_1(\alpha, \beta) &= 0 \\ S_2(\alpha, \beta) &= 0 \end{aligned}$$

The observed information matrix $\mathbf{I}(\alpha, \beta)$ is defined to be the matrix :

$$\mathbf{I}(\alpha, \beta) = \begin{pmatrix} I_{11}(\alpha, \beta) & I_{12}(\alpha, \beta) \\ I_{21}(\alpha, \beta) & I_{22}(\alpha, \beta) \end{pmatrix} = -\begin{pmatrix} \frac{\partial^2}{\partial\alpha^2}\ell & \frac{\partial^2}{\partial\alpha\partial\beta}\ell \\ \frac{\partial^2}{\partial\beta\partial\alpha}\ell & \frac{\partial^2}{\partial\beta^2}\ell \end{pmatrix}$$

The conditions for a value (α_0, β_0) satisfying $S_1(\alpha_0, \beta_0) = 0$ and $S_2(\alpha_0, \beta_0) = 0$ to be a MLE are that

$$I_{11}(\alpha_0, \beta_0) > 0,\ I_{22}(\alpha_0, \beta_0) > 0,$$

and

$$\det(\mathbf{I}(\alpha_0, \beta_0) = I_{11}(\alpha_0, \beta_0)I_{22}(\alpha_0, \beta_0) - I_{12}(\alpha_0, \beta_0)^2 > 0.$$

This is equivalent to requiring that both eigenvalues of the matrix $\mathbf{I}(\alpha_0, \beta_0)$ be positive.

Example 4.18 (Gaussian distribution).
Let $X_1, X_2 \ldots, X_n$ be i.i.d. observations from a $\mathcal{N}(\mu, v)$ density in which both μ and v are unknown. The log likelihood is

$$\begin{aligned}\ell(\mu, v) &= \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{1}{2v}(x_i-\mu)^2\right]\right] \\ &= \sum_{i=1}^{n}\left[-\frac{1}{2}\ln[2\pi] - \frac{1}{2}\ln[v] - \frac{1}{2v}(x_i-\mu)^2\right] \\ &= -\frac{n}{2}\ln[2\pi] - \frac{n}{2}\ln[v] - \frac{1}{2v}\sum_{i=1}^{n}(x_i-\mu)^2.\end{aligned}$$

Hence

$$S_1(\mu, v) = \frac{\partial \ell}{\partial \mu} = \frac{1}{v}\sum_{i=1}^{n}(x_i-\mu) = 0$$

implies that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}. \tag{4.9.1}$$

Also

$$S_2(\mu, v) = \frac{\partial \ell}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

implies that

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{4.9.2}$$

Calculating second derivatives and multiplying by -1 gives that the information matrix $\mathbf{I}(\mu, v)$ equals

$$\mathbf{I}(\mu, v) = \begin{pmatrix} \frac{n}{v} & \frac{1}{v^2}\sum_{i=1}^{n}(x_i-\mu) \\ \frac{1}{v^2}\sum_{i=1}^{n}(x_i-\mu) & -\frac{n}{2v^2} + \frac{1}{v^3}\sum_{i=1}^{n}(x_i-\mu)^2 \end{pmatrix}$$

Hence $\mathbf{I}(\hat{\mu}, \hat{v})$ is given by :

$$\begin{pmatrix} \frac{n}{\hat{v}} & 0 \\ 0 & \frac{n}{2v^2} \end{pmatrix}$$

Clearly both diagonal terms are positive and the determinant is positive and so $(\hat{\mu}, \hat{v})$ are, indeed, the MLEs of (μ, v).

Lemma 4.15 (Joint distribution of the sample mean and sample variance)**.** If $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, v)$ then the sample mean $\bar{X}$ and sample variance $(n-1)S^2$ are independent. Also $\bar{X}$ is distributed $\mathcal{N}(\mu, v/n)$ and $(n-1)S^2/v$ is a chi-squared random variable with $n-1$ degrees of freedom.

Proof. Define

$$\begin{aligned} W &= \sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2 \\ \Rightarrow \quad \frac{W}{v} + \frac{(\bar{X}-\mu)^2}{v/n} &= \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{v} \end{aligned}$$

The RHS is the sum of n independent standard normal random variables squared, and so is distributed $\chi^2_{n\text{df}}$ Also, $\bar{X} \sim \mathcal{N}(\mu, v/n)$, therefore $(\bar{X}-\mu)^2/(v/n)$ is the square of a standard normal and so is distributed $\chi^2_{1\text{df}}$ These Chi-Squared random variables have moment generating functions $(1-2t)^{-n/2}$ and $(1-2t)^{-1/2}$ respectively. Next, W/v and $(\bar{X}-\mu)^2/(v/n)$ are independent:

$$\begin{aligned}
\mathrm{Cov}(X_i-\bar{X},\bar{X}) &= \mathrm{Cov}(X_i,\bar{X})-\mathrm{Cov}(\bar{X},\bar{X}) \\
&= \mathrm{Cov}\left(X_i,\frac{1}{n}\sum X_j\right)-\mathrm{Var}(\bar{X}) \\
&= \frac{1}{n}\sum_j \mathrm{Cov}(X_i,X_j)-\frac{v}{n} \\
&= \frac{v}{n}-\frac{v}{n} \;=\; 0
\end{aligned}$$

But, $\mathrm{Cov}(X_i-\bar{X},\bar{X}-\mu)=\mathrm{Cov}(X_i-\bar{X},\bar{X})=0$, hence

$$\sum_i \mathrm{Cov}(X_i-\bar{X},\bar{X}-\mu)=\mathrm{Cov}\left(\sum_i (X_i-\bar{X}),\bar{X}-\mu\right)=0$$

As the moment generating function of the sum of independent random variables is equal to the product of their individual moment generating functions, we see

$$\begin{aligned}
\mathrm{E}\left[e^{t(W/v)}\right](1-2t)^{-1/2} &= (1-2t)^{-n/2} \\
\Rightarrow \qquad \mathrm{E}\left[e^{t(W/v)}\right] &= (1-2t)^{-(n-1)/2}
\end{aligned}$$

But $(1-2t)^{-(n-1)/2}$ is the moment generating function of a χ^2 random variables with $(n-1)$ degrees of freedom, and the moment generating function **uniquely** characterizes the random variable W/v. □

Example 4.19 ((Chi-squared distribibution).
Go back to equation (4.9.1), and $\bar{X}\sim\mathcal{N}(\mu,v/n)$. *Clearly* $\mathrm{E}(\bar{X})=\mu$ *(unbiased) and* $\mathrm{Var}(\bar{X})=v/n$, *so* $\bar{X}$ *achieved the CRLB. Go back to equation (4.9.2). Then from lemma 4.15 we have*

$$\frac{n\hat{v}}{v}\sim\chi^2_{n-1}$$

so that

$$\begin{aligned}
\mathrm{E}\left(\frac{n\hat{v}}{v}\right) &= n-1 \\
\Rightarrow \qquad \mathrm{E}(\hat{v}) &= \left(\frac{n-1}{n}\right)v
\end{aligned}$$

Instead, propose the (unbiased) estimator

$$\tilde{v} \;=\; \frac{n}{n-1}\hat{v} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{4.9.3}$$

Observe that

$$\mathrm{E}(\tilde{v}) \;=\; \left(\frac{n}{n-1}\right)\mathrm{E}(\hat{v}) \;=\; \left(\frac{n}{n-1}\right)\left(\frac{n-1}{n}\right)v \;=\; v$$

and $\tilde{v}$ *is unbiased as suggested. We can easily show that*

$$\mathrm{Var}(\tilde{v}) \;=\; \frac{2v^2}{(n-1)}$$

Hence

$$\text{eff}(\tilde{v}) = \frac{2v^2}{n} \div \frac{2v^2}{(n-1)} = 1 - \frac{1}{n}$$

Clearly $\tilde{v}$ is not efficient, but is asymptotically efficient.

4.10 Newton-Raphson optimization

4.10.1 One-paramter scenario

Example 4.20 (Radioactive Scatter).
A radioactive source emits particles intermittently and at various angles. Let X denote the cosine of the angle of emission. The angle of emission can range from 0 degrees to 180 degrees and so X takes values in $[-1,1]$. Assume that X has density

$$f(x|\theta) = \frac{1+\theta x}{2}$$

for $-1 \le x \le 1$ where $\theta \in [-1,1]$ is unknown. Suppose the data consist of n independently identically distributed measures of X yielding values $x_1, x_2, ..., x_n$. Here

$$\begin{aligned} L(\theta) &= \frac{1}{2^n}\prod_{i=1}^{n}(1+\theta x_i) \\ l(\theta) &= -n\ln[2] + \sum_{i=1}^{n}\ln[1+\theta x_i] \\ S(\theta) &= \sum_{i=1}^{n}\frac{x_i}{1+\theta x_i} \\ I(\theta) &= \sum_{i=1}^{n}\frac{x_i^2}{(1+\theta x_i)^2} \end{aligned}$$

Since $I(\theta) > 0$ for all θ, the MLE may be found by solving the equation $S(\theta) = 0$. It is not immediately obvious how to solve this equation.

By Taylor's Theorem we have

$$\begin{aligned} 0 &= S(\hat{\theta}) \\ &= S(\theta_0) + (\hat{\theta}-\theta_0)S'(\theta_0) + (\hat{\theta}-\theta_0)^2 S''(\theta_0)/2 + ... \end{aligned}$$

So, if $|\hat{\theta}-\theta_0|$ is small, we have that

$$0 \approx S(\theta_0) + (\hat{\theta}-\theta_0)S'(\theta_0),$$

and hence

$$\begin{aligned} \hat{\theta} &\approx \theta_0 - S(\theta_0)/S'(\theta_0) \\ &= \theta_0 + S(\theta_0)/I(\theta_0) \end{aligned}$$

We now replace θ_0 by this improved approximation $\theta_0 + S(\theta_0)/I(\theta_0)$ and keep repeating the process until we find a value $\hat{\theta}$ for which $|S(\hat{\theta})| < \epsilon$ where ϵ is some prechosen small number such as 0.000001. This method is called Newton's method for solving a non-linear equation.

If a unique solution to $S(\theta) = 0$ exists, Newton's method works well regardless of the choice of θ_0. When there are multiple solutions, the solution to which the algorithm converges depends crucially on the choice of θ_0. In many instances it is a good idea to try various starting values just to be sure that the method is not sensitive to the choice of θ_0.

4.10.2 Two-paramter scenario

Suppose that m parameters are involved, organized in the vector θ. When the m equations $S_r(\boldsymbol{\theta}) = 0,\ r = 1, \dots, m$ in this scenario cannot be solved directly numerical optimization is required. Let $\mathbf{S}(\boldsymbol{\theta})$ be the $m \times 1$ vector whose rth element is $S_r(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ be the solution to the set of equations $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{0}$ and let $\boldsymbol{\theta}_0$ be an initial guess at $\hat{\boldsymbol{\theta}}$. Then a first order Taylor's series approximation to the function S about the point $\boldsymbol{\theta}_0$ is given by

$$S_r(\hat{\boldsymbol{\theta}}) \approx S_r(\boldsymbol{\theta}_0) + \sum_{j=1}^{m} (\hat{\theta}_j - \theta_{0j}) \frac{\partial S_r}{\partial \theta_j}(\boldsymbol{\theta}_0)$$

for $r = 1, 2, \dots, m$ which may be written in matrix notation as

$$\mathbf{S}(\hat{\boldsymbol{\theta}}) \approx \mathbf{S}(\boldsymbol{\theta}_0) - \mathbf{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Requiring $\mathbf{S}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, this last equation can be reorganized to give

$$\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_0 + \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{S}(\boldsymbol{\theta}_0). \tag{4.10.1}$$

Thus given $\boldsymbol{\theta}_0$ this is a method for finding an improved guess at $\hat{\boldsymbol{\theta}}$. We then replace $\boldsymbol{\theta}_0$ by this improved guess and repeat the process. We keep repeating the process until we obtain a value $\boldsymbol{\theta}^*$ for which $|S_r(\boldsymbol{\theta}^*)|$ is less than ϵ for $r = 1, 2, \dots, m$ where ϵ is some small number like 0.0001. $\boldsymbol{\theta}^*$ will be an approximate solution to the set of equations $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{0}$. We then evaluate the matrix $\mathbf{I}(\boldsymbol{\theta}^*)$ and if all m of its eigenvalues are positive we set $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$.

4.10.3 Initial values

The Newton-Raphson method depends on the initial guess being close to the true value. If this requirement is not satisfied the procedure might convergence to a minimum instead of a maximum, or just simply diverge and fail to produce any estimates at all. Methods of finding good initial estimates depend very much on the problem at hand and may require some ingenuity.

A more general method for finding initial values is the method of moments. Retaking the example on 'Radioactive Scatter' (Example 4.20), using the Methods of Moments to find an intial value, involves solving the equation $\mathrm{E}(X) = \bar{x}$ for θ. For the previous example

$$\mathrm{E}(X) = \int_{-1}^{1} \frac{x(1+\theta x)}{2}\, dx = \frac{\theta}{3}$$

and so $\theta_0 = 3\bar{x}$ might yield a good choice for a starting value.

Suppose the following 10 values were recorded:

$$0.00,\ 0.23,\ -0.05,\ 0.01,\ -0.89,\ 0.19,\ 0.28,\ 0.51,\ -0.25 \textit{ and } 0.27.$$

Then $\bar{x} = 0.03$, and we substitute $\theta_0 = .09$ into the updating formula

$$\begin{aligned} \hat{\theta}_{new} &= \theta_{old} + \left(\sum_{i=1}^{n} \frac{x_i}{1 + \theta_{old} x_i} \right) \left(\sum_{i=1}^{n} \frac{x_i^2}{(1 + \theta_{old} x_i)^2} \right)^{-1} \\ \Rightarrow \quad \theta_1 &= 0.2160061 \\ \theta_2 &= 0.2005475 \\ \theta_3 &= 0.2003788 \\ \theta_4 &= 0.2003788 \end{aligned}$$

The relative likelihood function is plotted in Figure 4.10.1.

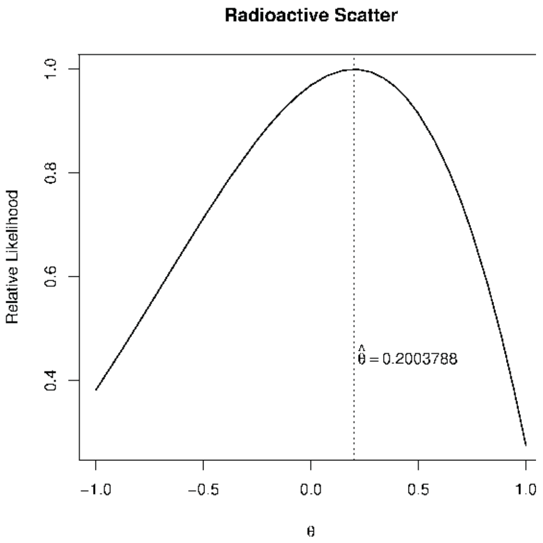


Figure 4.10.1: Relative likelihood for the radioactive scatter, solved by Newton Raphson.

4.10.4 Fisher's method of scoring

Several commonly applied modifications of the Newton-Raphson method exist. One such modification to Newton's method is to replace the matrix $\mathbf{I}(\boldsymbol{\theta})$ in (4.10.1) with $\bar{\mathbf{I}}(\boldsymbol{\theta}) = \mathrm{E}\left[\mathbf{I}(\boldsymbol{\theta})\right].$ The matrix $\bar{\mathbf{I}}(\boldsymbol{\theta})$ is positive definite, thus overcoming many of the problems regarding matrix inversion. Like Newton-Raphson, there is no guarantee that Fisher's method of scoring (as it is then called) will avoid producing negative parameter estimates or converging to local minima. Unfortunately, calculating $\bar{\mathbf{I}}(\boldsymbol{\theta})$ often can be mathematically difficult.

4.10.5 The method of profiling

Especially when several parameters are involved, a simplification of the estimaton process may be obtained by looking at profile likelihoods instead.

Recall that a likelihood function for a low-dimensional parameter can be conveniently visualized by its graph. If the likelihood is smooth, then this is roughly, at least locally, a reversed parabola with its top at the maximum likelihood estimator (MLE). The (negative) curvature of the graph at the MLE is known as the observed information and provides an estimate for the inverse of the variance of the MLE; steep likelihoods yield accurate estimates.

Definition 4.18 (Profile likelihood)**.** Consider a vector of parameters (θ_1, θ_2), with likelihood function $l(\theta_1, \theta_2; x)$. Suppose that $\widehat{\theta_{2.1}}$ is the MLE of θ_2 for a given value of θ_1. Then the profile likelihood for θ_1 is $l(\theta_1, \theta_{2.1}; x)$.

The profile likelihood may be used to a considerable extent as a full likelihood (but not always!). It is customary to use the curvature of the profile likelihood function as an estimate of the variability of $\widehat{\theta}$. For more information about profile likelihoods, we refer to [1].

Example 4.21 (Weibull).
A Weibull random variable with 'shape' parameter $a > 0$ and 'scale' parameter $b > 0$ has density

$$f_T(t) = (a/b)(t/b)^{a-1}\exp\{-(t/b)^a\}$$

for $t \geq 0$. The (cumulative) distribution function is

$$F_T(t) = 1 - \exp\{-(t/b)^a\}$$

on $t \geq 0$. Suppose that the time to failure T of components has a Weibull distribution and after testing n components for 100 hours, m components fail at times $t_1, \dots, t_m$, with $n - m$ components surviving the 100 hour test. The likelihood function can be written

$$L(\theta) = \underbrace{\prod_{i=1}^{m} \frac{a}{b}\left(\frac{t_i}{b}\right)^{a-1} \exp\left\{-\left(\frac{t_i}{b}\right)^a\right\}}_{\text{components failed}} \underbrace{\prod_{j=m+1}^{n} \exp\left\{-\left(\frac{100}{b}\right)^a\right\}}_{\text{components survived}}.$$

Then the log-likelihood function is

$$\ell(a, b) = m\ln(a) - ma\ln(b) + (a-1)\sum_{i=1}^{m}\ln(t_i) - \sum_{i=1}^{n}(t_i/b)^a\,,$$

where for convenience we have written $t_{m+1} = \cdots = t_n = 100$. This yields score functions

$$S_a(a, b) = \frac{m}{a} - m\ln(b) + \sum_{i=1}^{m}\ln(t_i) - \sum_{i=1}^{n}(t_i/b)^a\ln(t_i/b)\,,$$

and

$$S_b(a, b) = -\frac{ma}{b} + \frac{a}{b}\sum_{i=1}^{n}(t_i/b)^a\,. \tag{4.10.2}$$

It is not obvious how to solve $S_a(a, b) = S_b(a, b) = 0$ for a and b.

Setting $S_b(a, b) = 0$ from equation (4.10.2) we can solve for b in terms of a as

$$b = \left[\frac{1}{m}\sum_{i=1}^{n} t_i^a\right]^{1/a}. \tag{4.10.3}$$

This reduces our two parameter problem to a search over the "new" one-parameter profile log-likelihood

$$\begin{aligned} \ell_a(a) &= \ell(a, b(a)), \\ &= m\ln(a) - m\ln\left(\frac{1}{m}\sum_{i=1}^{n} t_i^a\right) + (a-1)\sum_{i=1}^{m}\ln(t_i) - m. \end{aligned} \tag{4.10.4}$$

Given an initial guess a_0 for a, an improved estimate a_1 can be obtained using a single parameter Newton-Raphson updating step $a_1 = a_0 + S(a_0)/I(a_0)$, where $S(a)$ and $I(a)$ are now obtained from $\ell_a(a)$. The estimates $\hat{a} = 1.924941$ and $\hat{b} = 78.12213$ were obtained by applying this method to the Weibull data using starting values $a_0 = 0.001$ and $a_0 = 5.8$ in 16 and 13 iterations respectively. However, the starting value $a_0 = 5.9$ produced the sequence of estimates $a_1 = -5.544163, a_2 = 8.013465, a_3 = -16.02908, a_4 = 230.0001$ and subsequently crashed.

4.10.6 Reparameterization

Negative parameter estimates can be avoided by reparameterizing the profile log-likelihood in (4.10.4) using $\alpha = \ln(a)$. Since $a = e^\alpha$ we are guaranteed to obtain $a > 0$. The reparameterized profile log-likelihood becomes

$$\ell_\alpha(\alpha) = m\alpha - m\ln\left(\frac{1}{m}\sum_{i=1}^{n} t_i^{e^\alpha}\right) + (e^\alpha - 1)\sum_{i=1}^{m}\ln(t_i) - m,$$

implying score function

$$S(\alpha) = m - \frac{me^\alpha \sum_{i=1}^n t_i^{e^\alpha} \ln(t_i)}{\sum_{i=1}^n t_i^{e^\alpha}} + e^\alpha \sum_{i=1}^m \ln(t_i)$$

and information function

$$\begin{aligned} I(\alpha) &= \frac{me^\alpha}{\sum_{i=1}^n t_i^{e^\alpha}} \left[\sum_{i=1}^n t_i^{e^\alpha} \ln(t_i) - e^\alpha \frac{[\sum_{i=1}^n t_i^{e^\alpha} \ln(t_i)]^2}{\sum_{i=1}^n t_i^{e^\alpha}} \right. \\ & \qquad \left. + e^\alpha \sum_{i=1}^n t_i^{e^\alpha} [\ln(t_i)]^2 \right] - e^\alpha \sum_{i=1}^m \ln(t_i). \end{aligned}$$

The estimates $\hat{a} = 1.924941$ and $\hat{b} = 78.12213$ were obtained by applying this method to the Weibull data using starting values $a_0 = 0.07$ and $a_0 = 76$ in 103 and 105 iterations respectively. However, the starting values $a_0 = 0.06$ and $a_0 = 77$ failed due to division by computationally tiny (1.0e-300) values.

4.10.7 The step-halving scheme

The Newton-Raphson method uses the (first and second) derivatives of $\ell(\boldsymbol{\theta})$ to maximize the function $\ell(\boldsymbol{\theta})$, but the function itself is not used in the algorithm. The log-likelihood can be incorporated into the Newton-Raphson method by modifying the updating step to

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \lambda_i \mathbf{I}(\boldsymbol{\theta}_i)^{-1} \mathbf{S}(\boldsymbol{\theta}_i), \qquad (4.10.5)$$

where the search direction has been multiplied by some $\lambda_i \in (0, 1]$ chosen so that the inequality

$$\ell\left(\boldsymbol{\theta}_i + \lambda_i \mathbf{I}(\boldsymbol{\theta}_i)^{-1} \mathbf{S}(\boldsymbol{\theta}_i)\right) > \ell\left(\boldsymbol{\theta}_i\right) \qquad (4.10.6)$$

holds. This requirement protects the algorithm from converging towards minima or saddle points. At each iteration the algorithm sets $\lambda_i = 1$, and if (4.10.6) does not hold λ_i is replaced with $\lambda_i/2$. The process is repeated until the inequality in (4.10.6) is satisfied. At this point the parameter estimates are updated using (4.10.5) with the value of λ_i for which (4.10.6) holds. If the function $\ell(\boldsymbol{\theta})$ is concave and unimodal convergence is guaranteed. Finally, when $\bar{\mathbf{I}}(\boldsymbol{\theta})$ is used in place of $\mathbf{I}(\boldsymbol{\theta})$ convergence to a (local) maxima is guaranteed, even if $\ell(\boldsymbol{\theta})$ is not concave.

4.11 Bayesian estimation

4.11.1 Bayes' theorem for random variables

Some of the following results are presented without elaboration, intended as a revision to provide the framework for Bayesian data analyses.

$$\begin{gathered} P(E|FH)P(F|H) = P(EF|H) = P(EFH|H) \\ P(E) = \sum_n P(E|H_n)P(H_n) \\ P(H_n|E)P(E) = P(EH_n) = P(H_n)P(E|H_n) \\ P(H_n|E) \propto P(H_n)P(E|H_n). \end{gathered}$$

The last result is Baye's theorem and in this form shows how we can "invert" probabilities, getting $P(H_n|E)$ from $P(E|H_n)$.

When H_n consists of exclusive and exhaustive events,

$$P(H_n|E) = \frac{P(H_n)P(E|H_n)}{\sum_m P(H_m)P(E|H_m)} \tag{4.11.1}$$

$$p(y|x) \propto p(y)p(x|y) \tag{4.11.2}$$

The constant of proportionality is

$$\begin{aligned} \text{continuous :} \quad & \frac{1}{p(x)} = \frac{1}{\int p(x|y)p(y)dy} \\ \text{discrete :} \quad & \frac{1}{p(x)} = \frac{1}{\sum_y p(x|y)p(y)dy} \end{aligned}$$

4.11.2 Post 'is' prior × likelihood

Suppose we are interested in the values of k unknown quantities,

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$$

and *apriori* beliefs about their values can be expressed in terms of the pdf $p(\boldsymbol{\theta})$.

Then we collect data,

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

which have a probability distribution that depends on $\boldsymbol{\theta}$, expressed as

$$p(\mathbf{X}|\boldsymbol{\theta})$$

From (4.11.2),

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \tag{4.11.3}$$

The term, $p(\mathbf{X}|\boldsymbol{\theta})$ may be considered as a function of $\mathbf{X}$ for fixed $\boldsymbol{\theta}$, i.e. a density of $\mathbf{X}$ which is parameterized by $\boldsymbol{\theta}$.

We can also consider the same term as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{X}$ and then it is termed the *likelihood function*,

$$\ell(\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta})$$

The terms in expression 4.11.3 represent the 3 components in Baeyesian inference:

- $p(\boldsymbol{\theta})$ is the prior
- $\ell(\boldsymbol{\theta}|\mathbf{X})$ is the likelihood
- $p(\boldsymbol{\theta}|\mathbf{X})$ is the posterior

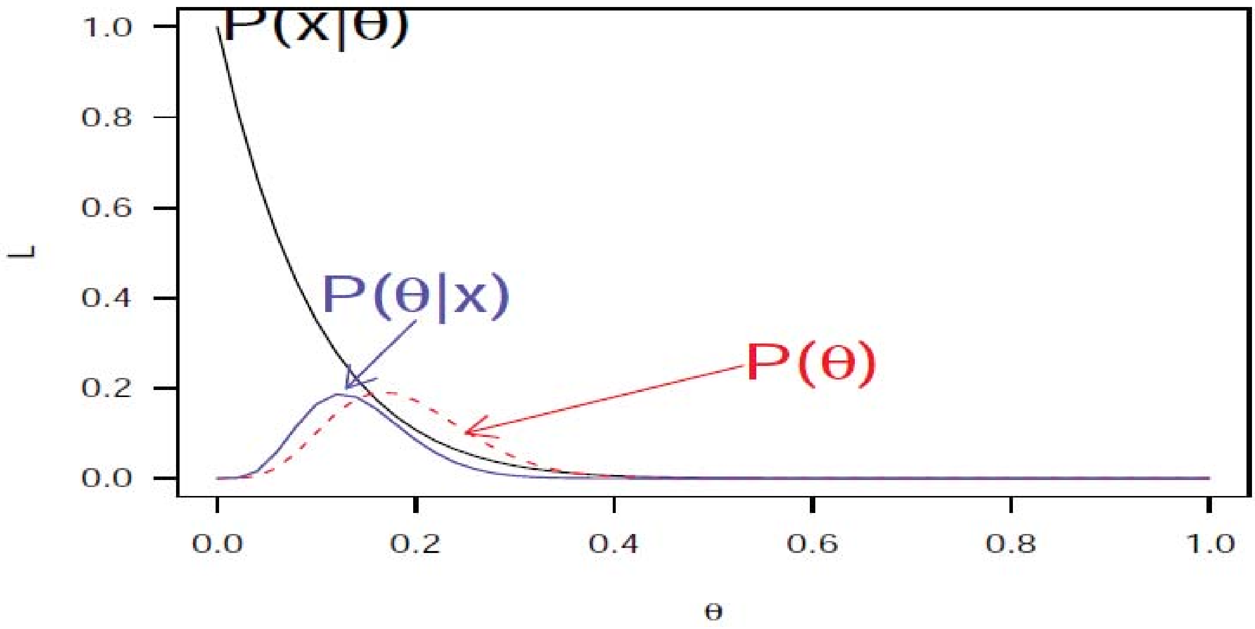


Figure 4.11.1: Posterior distribution

The Bayesian mantra therefore is:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

or **a posterior density is proportional to the prior times the likelihood**.

Note that the function $p(\cdot)$ is not the same in each instance but is a generic symbol to represent the density appropriate for prior, density of the data given the parameters, and the posterior. The form of p is understood by considering its arguments, i.e. $p(\theta)$, $p(x|\theta)$ or $p(\theta|x)$.

A diagram depicting the relationships amongst the different densities is shown in Fig. 4.11.1.

The posterior is a combination of the likelihood, where information about $\boldsymbol{\theta}$ comes from the data $\mathbf{X}$ and the prior $p(\boldsymbol{\theta})$ where the information is the knowledge of $\boldsymbol{\theta}$ independent of $\mathbf{X}$. This knowledge may come from previous sampling, (say). The *posterior* represents an update on $P(\theta)$ with the new information at hand, i.e. $\mathbf{x}$.

If the likelihood is weak due to insufficient sampling or wrong choice of likelihood function, the prior can dominate so that the posterior is just an adaptation of the prior. Alternatively, if the sample size is large so that the likelihood function is strong, the prior will not have much impact and the Bayesian analysis is the same as the maximum likelihood.

The output is a distribution, $p(\theta|\mathbf{X})$ and we may interpret it using summaries such as the median. A tabular view on the difference between Bayesian inference and frequentist inference is given in Table ??. In a frequentist world, θ is considered a fixed but unknown feature of the population from which data is being (randomly) sampled. In a Bayesian world it is the estimator $\widehat{\theta}$ that is fixed (a function of the data available for analysis) and θ is a random variable, subject to (subjective) uncertainty.

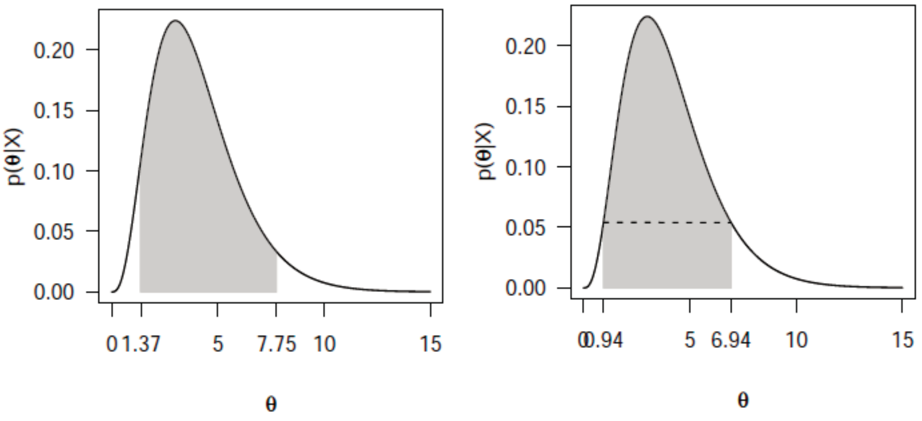
From the output (i.e., posterior distribution of θ) we can derive intervals where the true value of θ would lie with a certain probability. We differentiate between credible and highest probability density regions.

	Bayesian	Frequentist
θ	random	fixed but unknown
$\widehat{\theta}$	fixed	random
‘randomness’	subjective	sampling
distribution of interest	posterior	sampling distribution

Definition 4.19 (Credible Region)**.** A region $C \subseteq \mathbf{\Omega}$ such that $\int_C p(\theta)d\theta = -\alpha, 0 \leq \alpha \leq 1$ is a $100(1 - alpha)\%$ credible region for θ. For single-parameter problems, if C is not a set of disjoint intervals, then C is a credible interval. If $p(\theta)$ is a (prior/posterior) density, then C is a (prior/posterior) credible region.

Definition 4.20 (Highest Probability Density Region)**.** A region $C \subseteq \mathbf{\Omega}$ is a $100(1 - \alpha)\%$ highest probability density region for θ under $p(\theta)$ if $P(\theta \in C) = 1 - \alpha$ and $P(\theta_1) \geq P(\theta_2), \forall \theta_1 \in C, \theta_2 \notin C$.

Figure 4.11.2 depicts a density with shaded areas of 0.9 in 2 cases. In frame 4.2(a), observe that there are quantiles outside the interval $(1.37, 7.75)$ for which the density is greater than quantiles within the interval. Frame 4.2(b) depicts the HDR as (0.94, 6.96).



(a) Confidence interval (b) A HDR

Figure 4.11.2: Comparison of 2 types of regions, a Confidence Interval and a Highest Density Region.

We may now have a notion about uncertainty, but how to derive a Bayesian point estimate? Bayes estimates are single number summaries of a posterior density. As there are modes, medians, means, etc, it is the question which summary to choose best. Different loss functions rationalize different point estimates. The quadratic loss $l(\theta, \tilde{\theta} = (\theta - \tilde{\theta})^2$ indicates the quadratic loss from the use of the estimate $\tilde{\theta}$ instead of θ. The posterior mean as Bayes estimate under quadratic loss is given by

$$E(\theta|\mathbf{y}) = \tilde{\theta} = \int_\theta \theta p(\theta|\mathbf{y})d\theta. \tag{4.11.4}$$

CHAPTER 5

CONFIDENCE INTERVALS

5.1 Introduction

We have precedently been considering point estimators of a parameter, and in the Bayesian context, we have also briefly touched upon ‘credibility’, credible intervals. Also in the frequentist context we can construct an interval to which we can attach a probability that the true value lies within the limits of the interval. What these intervals are and how we can construct them, will be the subject of this chapter.

Recall that by *point estimator* we are referring to the fact that, after the sampling has been done and the observed value of the estimator computed, our end-product is the single number which is hopefully a good approximation for the unknown true value of the parameter. If the estimator is good according to some criteria (and we have seen several of these criteria in Chapter 4, then the estimate should be reasonably close to the unknown true value. But the single number itself does not include any indication of how high the probability might be that the estimator has taken on a value close to the true unknown value. The method of confidence intervals gives both an idea of the actual numerical value of the parameter, by giving it a range of possible values, and a measure of how confident we are that the true value of the parameter is in that range. To pursue this idea further consider the following example.

Example 5.1.
Consider a random sample of size n for a normal distribution with mean μ (unknown) and known variance σ^2. Find a 95% CI for the unknown mean, μ.

Solution of Example 5.1. *We know that the best estimator of μ is $\overline{X}$ and the sampling distribution of $\overline{X}$ is $N\left(\mu, \frac{\sigma^2}{n}\right)$. Then from the standard normal,*

$$P\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} < 1.96\right) = .95.$$

The event $\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} < 1.96$ is equivalent to the event

$$\mu - \frac{1.96\sigma}{\sqrt{n}} < \overline{X} < \mu + \frac{1.96\sigma}{\sqrt{n}},$$

which is equivalent to the event

$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Hence

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95 \tag{5.1.1}$$

The two statistics $\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}$, $\overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$ are the endpoints of a 95% CI for μ. This is reported as

The 95% CI for μ is $\left(\widehat{X} - 1.96\frac{\sigma}{\sqrt{n}}, \widehat{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$.

Computer Exercise 5.1. *Generate 100 samples of size 9 from a $N(0,1)$ distribution. Find the 95% CI for μ for each of these samples and count the number that do (don't) contain zero. (You could repeat this say 10 times to build up the total number of CI's generated to 1000.) You should observe that about 5% of the intervals don't contain the true value of μ $(= 0)$.*

Solution of Computer Exercise 5.1. *Use the commands:*

```
#___________ ConfInt.R __________
sampsz <- 9
nsimulations <- 100
non.covered <- 0
for (i in 1:nsimulations){
rn <- rnorm(mean=0,sd=1,n=sampsz)

Xbar <- mean(rn)
s <- sd(rn)
CI <- qnorm(mean=Xbar,sd=s/sqrt(sampsz),p=c(0.025,0.975) )

non.covered <- non.covered + (CI[1] > 0) + (CI[2] < 0)

}
cat("Rate of non covering CI's",100*non.covered/nsimulations," % \n")


> source("ConfInt.R")
Rate of non covering CI's 8 %
```

This implies that 8 of the CI's don't contain 0. With a larger sample size we would expect that about 5% of the CI's would not contain zero.

We make the following definition:

Definition 5.1 (Random interval). An interval, at least one of whose endpoints is a random variable is called a **random interval**.

In (5.1.1), we are saying that the probability is 0.95 that the random interval

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

contains μ.

A CI has to be interpreted carefully. For a particular sample, where $\overline{x}$ is the observed value of $\overline{X}$, a 95% CI for μ is

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right), \tag{5.1.2}$$

but the statement

$$\overline{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < 1.96\frac{\sigma}{\sqrt{n}}$$

is either true or false. The parameter μ is a constant and either the interval contains it in which case the statement is true, or it does not contain it, in which case the statement is false. A probability 0.95 needs to be interpreted in terms of the relative frequency with which the indicated event will occur "in the long run" of similar sampling experiments. Each time we take a sample of size n, a different $\overline{x}$, and hence a different interval (5.1.2) would be obtained. Some of these intervals will contain μ as claimed, and some will not. In fact, if we did this many times, we'd expect that 95 times out of 100 the interval obtained would contain μ. The measure of our confidence is then 0.95 because before a sample is drawn there is a probability of 0.95 that the confidence interval to be constructed will cover the true mean.

A statement such as $P(3.5 < \mu < 4.9) = 0.95$ is incorrect and should be replaced by

A 95% CI for μ is $(3.5, 4.9)$.

We can generalize the above as follows: Let $z_{\frac{\alpha}{2}}$ be defined by

$$\Phi(z_{\frac{\alpha}{2}}) = 1 - (\alpha/2). \tag{5.1.3}$$

That is, the area under the normal curve **above** $z_{\alpha/2}$ is $\alpha/2$. Then

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha. \tag{5.1.4}$$

So a $100(1-\alpha)\%$ CI for μ is

$$\left(\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right). \tag{5.1.5}$$

Commonly used values of α are $0.1, 0.05, 0.01$.

Obviously, confidence intervals for a given parameter are not unique. For example, we have considered a **symmetric**, **two-sided** interval, but

$$\left(\overline{x} - z_{2\alpha/3}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/3}\frac{\sigma}{\sqrt{n}}\right)$$

is also a $100(1-\alpha)\%$ CI for μ. Likewise, we could have one-sided CI's for μ. For example,

$$\left(-\infty, \overline{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right) \quad \text{or} \quad \left(\overline{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right).$$

The second of these arises from considering $P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1-\alpha$. Also, alternatively, we could have a CI based on say, the sample median instead of the sample mean.

As for methods to derive point estimates, methods of obtaining CIs must be judged by their various statistical properties. For example, one desirable property is to have the length (or expected length) of a $100(1-\alpha)\%$ CI as short as possible. Note that for the CI in (5.1.5), the length is constant for given n.

5.2 Exact confidence intervals

Suppose that we are going to observe the value of a random vector $\mathbf{X}$. Let $\mathcal{X}$ denote the set of possible values that $\mathbf{X}$ can take and, for $\mathbf{x} \in \mathcal{X}$, let $g(\mathbf{x}|\theta)$ denote the probability that $\mathbf{X}$ takes the value $\mathbf{x}$ where the parameter θ is some unknown element of the set Θ. Consider the problem of quoting a subset of θ values which are in some sense plausible in the light of the data $\mathbf{x}$. We need a procedure which for each possible value $\mathbf{x} \in \mathcal{X}$ specifies a subset $C(\mathbf{x})$ of Θ which we should quote as a set of plausible values for θ.

Example 5.2.
Suppose we are going to observe data $\mathbf{x}$ where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and $x_1, x_2, \ldots, x_n$ are the observed values of random variables $X_1, X_2, \ldots, X_n$ which are thought to be iid $N(\theta, 1)$ for some unknown parameter $\theta \in (-\infty, \infty) = \Theta$. Consider the subset $C(\mathbf{x}) = [\bar{x} - 1.96/\sqrt{n}, \bar{x} + 1.96/\sqrt{n}]$. If we carry out an infinite sequence of independent repetitions of the experiment then we will get an infinite sequence of $\mathbf{x}$ values and thereby an infinite sequence of subsets $C(\mathbf{x})$. We might ask what proportion of this infinite sequence of subsets actually contain the fixed but unknown value of θ?

Solution of Example 5.2. *Since $C(\mathbf{x})$ depends on $\mathbf{x}$ only through the value of $\bar{x}$ we need to know how $\bar{x}$ behaves in the infinite sequence of repetitions. This follows from the fact that $\bar{X}$ has a $N(\theta, \frac{1}{n})$ density and so $Z = \frac{\bar{X}-\theta}{\frac{1}{\sqrt{n}}} = \sqrt{n}(\bar{X}-\theta)$ has a $N(0,1)$ density. Thus eventhough θ is unknown we can calculate the probability that the value of Z will exceed 2.78, for example, using the standard normal tables. Remember that (from a frequentist viewpoint) the probability is the proportion of experiments in the infinite sequence of repetitions which produce a value of Z greater than 2.78.*

In particular we have that $P[|Z| \leq 1.96] = 0.95$. *Thus* 95% *of the time* Z *will lie between* -1.96 *and* $+1.96$. *But*

$$\begin{aligned} -1.96 \leq Z \leq +1.96 &\Rightarrow -1.96 \leq \sqrt{n}(\bar{X} - \theta) \leq +1.96 \\ &\Rightarrow -1.96/\sqrt{n} \leq \bar{X} - \theta \leq +1.96/\sqrt{n} \\ &\Rightarrow \bar{X} - 1.96/\sqrt{n} \leq \theta \leq \bar{X} + 1.96/\sqrt{n} \\ &\Rightarrow \theta \in C(\mathbf{X}) \end{aligned}$$

Thus we have answered the question we started with. The proportion of the infinite sequence of subsets given by the formula $C(\mathbf{X})$ *which will actually include the fixed but unknown value of* θ *is* 0.95. *For this reason the set* $C(\mathbf{X})$ *is called a* 95% *confidence set or* confidence interval *for the parameter* θ.

It is well to bear in mind that once we have actually carried out the experiment and observed *our* value of $\mathbf{x}$, the resulting interval $C(\mathbf{x})$ either does or does not contain the unknown parameter θ. We do not know which is the case. All we know is that the procedure we used in constructing $C(\mathbf{x})$ is one which 95% of the time produces an interval which contains the unknown parameter.

The crucial step in the last example was finding the quantity $Z = \sqrt{n}(\bar{X} - \theta)$ whose value depended on the parameter of interest θ but whose distribution was *known* to be that of a standard normal variable. This leads to the following definition.

Definition 5.2 (Pivotal Quantity)**.** A pivotal quantity for a parameter θ is a random variable $Q(\mathbf{X}|\theta)$ whose value depends both on (the data) $\mathbf{X}$ and on the value of the unknown parameter θ but whose distribution is known.

The quantity Z in the example above is a pivotal quantity for θ. The following lemma provides a method of finding pivotal quantities in general.

Lemma 5.1. Let X be a random variable and define $F(a) = P[X \leq a]$. Consider the random variable $U = -2\log[F(X)]$. Then U has a χ^2_2 density. Consider the random variable $V = -2\log[1 - F(X)]$. Then V has a χ^2_2 density.

Proof. Observe that, for $a \geq 0$,

$$\begin{aligned} P[U \leq a] &= P[F(X) \geq \exp(-a/2)] \\ &= 1 - P[F(X) \leq \exp(-a/2)] \\ &= 1 - P[X \leq F^{-1}(\exp(-a/2))] \\ &= 1 - F[F^{-1}(\exp(-a/2))] \\ &= 1 - \exp(-a/2). \end{aligned}$$

Hence, U has distribution $\frac{1}{2}\exp(-a/2)$ which corresponds to the distribution of a χ^2_2 variable as required. The corresponding proof for V is left as an exercise. □

This lemma has an immediate, and very important, application.

Suppose that we have data $X_1, X_2, \ldots, X_n$ which are iid with density $f(x|\theta)$. Define $F(a|\theta) = \int_{-\infty}^{a} f(x|\theta)dx$ and, for $i = 1, 2, \ldots, n$, define $U_i = -2\log[F(X_i|\theta)]$. Then $U_1, U_2, \ldots, U_n$ are iid each having a χ_2^2 density. Hence $Q_1(\mathbf{X}, \theta) = \sum_{i=1}^n U_i$ has a χ_{2n}^2 density and so is a pivotal quantity for θ. Another pivotal quantity (also having a χ_{2n}^2 density) is given by $Q_2(\mathbf{X}, \theta) = \sum_{i=1}^n V_i$ where $V_i = -2\log[1 - F(X_i|\theta)]$.

According to [5], the pivotal method thus depends on finding a pivotal quantitity that has 2 characteristics:

1. It is a function of the sample observations and the unknown parameter θ, say $H(X_1, X_2, \dots, X_n; \theta)$ where θ is the only unknown quantity,

2. It has a probability distribution that does not depend on θ.

Then any probability statement of the form

$$P(a < H(X_1, X_2, \dots, X_n; \theta) < b) = 1 - \alpha$$

will give rise to a probability statement about θ.

Example 5.3 ([5]).
Given $X_1, X_2, \dots, X_{n_1}$ from $N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \dots, Y_{n_2}$ from $N(\mu_2, \sigma_2^2)$ where σ_1^2, σ_2^2 are known, find a symmetric 95% CI for $\mu_1 - \mu_2$.

Solution of Example 5.3 ([5]). *Consider $\mu_1 - \mu_2$ ($= \theta$, say) as a single parameter. Then $\overline{X}$ is distributed $N(\mu_1, \sigma_1^2/n_1)$ and $\overline{Y}$ is distributed $N(\mu_2, \sigma_2^2/n_2)$ and further, $\overline{X}$ and $\overline{Y}$ are independent. It follows that $\overline{X} - \overline{Y}$ is normally distributed, and writing it in standardized form,*

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \quad \textit{is distributed as } N(0,1).$$

So we have found the pivotal quantity which is a function of $\mu_1 - \mu_2$ but whose distribution does not depend on $\mu_1 - \mu_2$. A 95% CI for $\theta = \mu_1 - \mu_2$ is found by considering

$$P\left(-1.96 < \frac{\overline{X} - \overline{Y} - \theta}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} < 1.96\right) = .95,$$

which, on rearrangement, gives the appropriate CI for $\mu_1 - \mu_2$. That is,

$$\left(\overline{x} - \overline{y} - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{x} - \overline{y} + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right). \tag{5.2.1}$$

Example 5.4 ([5]).
In many problems where we need to estimate proportions, it is reasonable to assume that sampling is from a binomial population, and hence that the problem is to estimate p in the bin(n, p) *distribution, where p is unknown. Find a $100(1-\alpha)$% CI for p, making use of the fact that for large sample sizes, the binomial distribution can be approximated by the normal.*

Solution of Example 5.4 ([5]). *Given X is distributed as* bin(n, p), *an unbiased estimate of p is $\widehat{p} = X/n$. For n large, X/n is approximately normally distributed. Then,*

$$E(\widehat{p}) = E(X)/n = p,$$

and

$$Var(\widehat{p}) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

so that

$$\frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \quad \textit{is distributed approximately } N(0,1).$$

[Note that we have found the required pivotal quantity whose distribution does not depend on p.]

An approximate $100(1-\alpha)\%$ CI for p is obtained by considering

$$P\left(-z_{\alpha/2} < \frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) = 1-\alpha \tag{5.2.2}$$

where $z_{\alpha/2}$ is defined in (5.1.3). Rearranging (5.2.2), the confidence limits for p are obtained as

$$\frac{2n\widehat{p} + z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{4n\widehat{p}(1-\widehat{p}) + z^2_{\alpha/2}}}{2(n+z^2_{\alpha/2})}. \tag{5.2.3}$$

A simpler expression can be found by dividing both numerator and denominator of (5.2.3) by $2n$ and neglecting terms of order $1/n$. That is, a 95% CI for p is

$$\left(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\right). \tag{5.2.4}$$

Note that this is just the expression we would have used if we replaced $Var(\widehat{p}) = p(1-p)/n$ in (5.2.2) by $\widehat{Var(\widehat{p})} = \widehat{p}(1-\widehat{p})/n$. In practice, confidence limits for p when n is small will need to be constructed in another way. We will see more about this in future chapters.

Example 5.5 ([5]).
Construct an appropriate 90% confidence interval for λ in the Poisson distribution. Evaluate this if a sample of size 30 yields $\sum x_i = 240$.

Solution of Example 5.5 ([5])**.** *Now $\overline{X}$ is an unbiased estimator of λ for this problem, so λ can be estimated by $\widehat{\lambda} = \overline{x}$ with $E(\widehat{\lambda}) = \lambda$ and $Var(\widehat{\lambda}) = Var(\overline{X}) = \sigma^2/n = \lambda/n$. By the Central Limit theorem, for large n, the distribution of $\overline{X}$ is approximately normal, so*

$$\frac{\overline{X}-\lambda}{\sqrt{\lambda/n}} \quad \textit{is distributed approximately } N(0,1).$$

An approximate 90% CI for λ can be obtained from considering

$$P\left(-1.645 < \frac{\overline{X}-\lambda}{\sqrt{\lambda/n}} < 1.645\right) = .90. \tag{5.2.5}$$

Rearrangement of the inequality in (5.2.5) to give an inequality for λ, is similar to that in Example 5.4 where it was necessary to solve a quadratic. But, noting the comment following (5.2.4), replace the variance of $\overline{X}$ by its estimate $\frac{\widehat{\lambda}}{n} = \frac{\overline{X}}{n}$, giving for the 90% CI for λ,

$$\left(\overline{x} - 1.645\sqrt{\overline{x}/n}\overline{x} + 1.645\sqrt{\overline{x}/n}\right)$$

which on substitution of the observed value $240/30 = 8$ for $\overline{x}$ gives $(7.15, 8.85)$.

Example 5.6.
Suppose that we have data $X_1, X_2, \dots, X_n$ which are iid with density

$$f(x|\theta) = \theta \exp(-\theta x)$$

for $x \geq 0$ and suppose that we want to construct a 95% confidence interval for θ. We need to find a pivotal quantity for θ. Observe that

$$\begin{aligned} F(a|\theta) &= \int_{-\infty}^{a} f(x|\theta)\mathrm{d}x \\ &= \int_{0}^{a} \theta \exp(-\theta x)\mathrm{d}x \\ &= 1 - \exp(-\theta a). \end{aligned}$$

Hence

$$Q_1(\mathbf{X}, \theta) = -2\sum_{i=1}^{n} \log\left[1 - \exp(-\theta X_i)\right]$$

is a pivotal quantity for θ having a χ^2_{2n} density. Also

$$Q_2(\mathbf{X}, \theta) = -2\sum_{i=1}^{n} \log\left[\exp(-\theta X_i)\right] = 2\theta\sum_{i=1}^{n} X_i$$

is another pivotal quantity for θ having a χ^2_{2n} density.

Using the tables, find $A < B$ such that $P[\chi^2_{2n} < A] = P[\chi^2_{2n} > B] = 0.025$. Then

$$\begin{aligned} 0.95 &= P[A \leq Q_2(\mathbf{X}, \theta) \leq B] \\ &= P\left[A \leq 2\theta\sum_{i=1}^{n} X_i \leq B\right] \\ &= P\left[\frac{A}{2\sum_{i=1}^{n} X_i} \leq \theta \leq \frac{B}{2\sum_{i=1}^{n} X_i}\right] \end{aligned}$$

and so the interval

$$\left[\frac{A}{2\sum_{i=1}^{n} X_i}, \frac{B}{2\sum_{i=1}^{n} X_i}\right]$$

is a 95% confidence interval for θ.

The other pivotal quantity $Q_1(\mathbf{X}, \theta)$ is more awkward in this example since it is not straightforward to determine the set of θ values which satisfy $A \leq Q_1(\mathbf{X}, \theta) \leq B$.

5.3 Pivotal quantities for use with normal data

Many exact pivotal quantities have been developed for use with Gaussian data.

Example 5.7.
Suppose that we have data $X_1, X_2, \dots, X_n$ which are iid observations from a $\mathcal{N}(\theta, \sigma^2)$ density where σ is known. Define

$$Q = \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma}.$$

Then Q has a $\mathcal{N}(0, 1)$ density and so is a pivotal quantity for θ. In particular we can be 95% sure that

$$-1.96 \leq \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma} \leq +1.96$$

which is equivalent to

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \theta \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

The `R` *command* `qnorm(p=0.975,mean=0,sd=1)` *returns the value 1.959964 as the* $97\frac{1}{2}\%$ *quantile from the standard normal distribution.*

Example 5.8.
Suppose that we have data $X_1, X_2, \ldots, X_n$ *which are iid observations from a* $N(\theta, \sigma^2)$ *density where* θ *is known. Define*

$$Q = \frac{\sum_{i=1}^{n}(X_i - \theta)^2}{\sigma^2}$$

We can write $Q = \sum_{i=1}^{n} Z_i^2$ *where* $Z_i = (X_i - \theta)/\sigma$. *If* Z_i *has a* $\mathcal{N}(0,1)$ *density then* Z_i^2 *has a* χ_1^2 *density. Hence,* Q *has a* χ_n^2 *density and so is a pivotal quantity for* σ. *If* $n = 20$ *then we can be* 95% *sure that*

$$9.591 \le \frac{\sum_{i=1}^{n}(X_i - \theta)^2}{\sigma^2} \le 34.170$$

which is equivalent to

$$\sqrt{\frac{1}{34.170}\sum_{i=1}^{n}(X_i - \theta)^2} \le \sigma \le \sqrt{\frac{1}{9.591}\sum_{i=1}^{n}(X_i - \theta)^2}.$$

The `R` *command* `qchisq(p=c(.025,.975),df=20)` *returns the values 9.590777 and 34.169607 as the* $2\frac{1}{2}\%$ *and* $97\frac{1}{2}\%$ *quantiles from a Chi-squared distribution on 20 degrees of freedom.*

Lemma 5.2 (The Student t-distribution)**.** Suppose the random variables X and Y are independent, and $X \sim N(0,1)$ and $Y \sim \chi_n^2$. Then the ratio

$$T = \frac{X}{\sqrt{Y/n}}$$

is Student-t distributed with n degrees of freedom (df).

Example 5.9.
Suppose that we have data $X_1, X_2, \ldots, X_n$ *which are iid observations from a* $\mathcal{N}(\theta, \sigma^2)$ *density where both* θ *and* σ *are unknown. Define*

$$Q = \frac{\sqrt{n}(\bar{X} - \theta)}{s}$$

where

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

We can write

$$Q = \frac{Z}{\sqrt{W/(n-1)}}$$

where

$$Z = \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma}$$

has a $\mathcal{N}(0,1)$ *density and*

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

has a χ^2_{n-1} *density (see lemma 4.15).*

In lemma 5.2 we show that Q *has a* t_{n-1} *density, and so is a pivotal quantity for* θ*. If* $n = 31$ *then we can be* 95% *sure that*

$$-2.042 \leq \frac{\sqrt{n}(\bar{X} - \theta)}{s} \leq +2.042$$

which is equivalent to

$$\bar{X} - 2.042\frac{s}{\sqrt{n}} \leq \theta \leq \bar{X} + 2.042\frac{s}{\sqrt{n}}. \tag{5.3.1}$$

The `R` *command* `qt(p=.975,df=30)` *returns the value 2.042272 as the* $97\frac{1}{2}$*% quantile from a Student t-distribution on 30 degrees of freedom.*

It also follows immediately that W *is a pivotal quantity for* σ*. If* $n = 31$ *then we can be* 95% *sure that*

$$16.79077 \leq \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \leq 46.97924$$

which is equivalent to

$$\sqrt{\frac{1}{46.97924}\sum_{i=1}^{n}(X_i - \bar{X})^2} \leq \sigma \leq \sqrt{\frac{1}{16.79077}\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{5.3.2}$$

The `R` *command* `qchisq(p=c(.025,.975),df=30)` *returns the values 16.79077 and 46.97924 as the* $2\frac{1}{2}$*% and* $97\frac{1}{2}$*% quantiles from a Chi-squared distribution on 30 degrees of freedom.*

It is important to point out that although a probability statement involving 95% confidence has been attached the two intervals (5.3.1) and (5.3.2) separately, this does not imply that both intervals simultaneously hold with 95% confidence.)

Example 5.10.
Suppose that we have data $X_1, X_2, \ldots, X_n$ *which are iid observations from a* $\mathcal{N}(\theta_1, \sigma^2)$ *density and data* $Y_1, Y_2, \ldots, Y_m$ *which are iid observations from a* $\mathcal{N}(\theta_2, \sigma^2)$ *density where* $\theta_1, \theta_2,$ *and* σ *are unknown. Let* $\delta = \theta_1 - \theta_2$ *and define*

$$Q = \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}}$$

where

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{n + m - 2}.$$

We know that $\bar{X}$ *has a* $\mathcal{N}(\theta_1, \frac{\sigma^2}{n})$ *density and that* $\bar{Y}$ *has a* $\mathcal{N}(\theta_2, \frac{\sigma^2}{m})$ *density. Then the difference* $\bar{X} - \bar{Y}$ *has a* $\mathcal{N}(\delta, \sigma^2[\frac{1}{n} + \frac{1}{m}])$ *density. Hence*

$$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma^2[\frac{1}{n} + \frac{1}{m}]}}$$

has a $\mathcal{N}(0, 1)$ *density. Let* $W_1 = \sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2$ *and let* $W_2 = \sum_{j=1}^{m}(Y_j - \bar{Y})^2/\sigma^2$*. Then,* W_1 *has a* χ^2_{n-1} *density and* W_2 *has a* χ^2_{m-1} *density, and* $W = W_1 + W_2$ *has a* χ^2_{n+m-2} *density. We can write*

$$Q_1 = Z/\sqrt{W/(n + m - 2)}$$

and so, Q_1 *has a* t_{n+m-2} *density and so is a pivotal quantity for* δ*. Define*

$$Q_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{\sigma^2}.$$

Then Q_2 *has a* χ^2_{n+m-2} *density and so is a pivotal quantity for* σ*.*

Lemma 5.3 (The Fisher F-distribution)**.** Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be iid $\mathcal{N}(0,1)$ random variables. The ratio

$$Z = \frac{\sum\limits_{i=1}^{n} X_i^2/n}{\sum\limits_{i=1}^{m} Y_i^2/m}$$

has the distribution called Fisher, or F distribution with parameters (degrees of freedom) n, m, or the $\mathrm{F}_{n,m}$ distribution for short. The corresponding pdf $f_{\mathrm{F}_{n,m}}$ is concentrated on the positive half axis:

$$f_{\mathrm{F}_{n,m}}(z) = \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} \left(\frac{n}{m}\right)^{n/2} z^{n/2-1} \left(1 + \frac{n}{m} z\right)^{-(n+m)/2} \qquad \text{for } z > 0.$$

Observe that if $T \sim \mathrm{t}_m$, then $T^2 = Z \sim \mathrm{F}_{1,m}$, and if $Z \sim \mathrm{F}_{n,m}$, then $Z^{-1} \sim \mathrm{F}_{m,n}$.
If $W_1 \sim \chi^2_n$ and $W_2 \sim \chi^2_m$, then $Z = (mW_1)/(nW_2) \sim \mathrm{F}_{n,m}$.

Example 5.11.
Suppose that we have data $X_1, X_2, \ldots, X_n$ which are iid observations from a $\mathcal{N}(\theta_X, \sigma_X^2)$ density and data $Y_1, Y_2, \ldots, Y_m$ which are iid observations from a $\mathcal{N}(\theta_Y, \sigma_Y^2)$ density where $\theta_X, \theta_Y, \sigma_X$, and σ_Y are all unknown. Let

$$\lambda = \sigma_X/\sigma_Y$$

and define

$$F^* = \frac{\hat{s}_X^2}{\hat{s}_Y^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \frac{(m-1)}{\sum_{j=1}^{m}(Y_j - \bar{Y})^2}.$$

Let

$$W_X = \sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma_X^2$$

and let

$$W_Y = \sum_{j=1}^{m}(Y_j - \bar{Y})^2/\sigma_Y^2.$$

Then, W_X has a χ^2_{n-1} density and W_Y has a χ^2_{m-1} density. Hence, by lemma 5.3,

$$Q = \frac{W_X/(n-1)}{W_Y/(m-1)} \equiv \frac{F^*}{\lambda^2}$$

has an F density with $n-1$ and $m-1$ degrees of freedom and so is a pivotal quantity for λ. Suppose that $n = 25$ and $m = 13$. Then we can be 95% sure that $0.39 \le Q \le 3.02$ which is equivalent to

$$\sqrt{\frac{F^*}{3.02}} \le \lambda \le \sqrt{\frac{F^*}{0.39}}.$$

To see how this might work in practice try the following `R` *commands one at a time:*

```
x = rnorm(25, mean = 0, sd = 2)
y = rnorm(13, mean = 1, sd = 1)
Fstar = Var(x)/Var(y); Fstar
CutOffs = qf(p=c(.025,.975), df1=24, df2=12)
CutOffs; rev(CutOffs)
Fstar / rev(CutOffs)
var.test(x, y)
```

The search for a nice pivotal quantity for $\delta = \theta_X - \theta_X$ (both variances σ_X^2 and σ_Y^2 unknown and not known to be equal) continues and is one of the great unsolved problems in statistics, referred to as the Behrens-Fisher Problem. One reason for its fame is that it can be proven that there is no exact solution satisfying the classical criteria for good tests. First-best solutions that are uniformly most powerful and invariant either do not exist or have strange properties. One needs to look for second-best solutions...

5.4 Approximate confidence intervals

Let $X_1, X_2, \ldots, X_n$ be iid with density $f(x|\theta)$. Let $\hat{\theta}$ be the MLE of θ. We saw before that the quantities $W_1 = \sqrt{\mathrm{E}(I(\theta))}(\hat{\theta} - \theta), W_2 = \sqrt{I(\theta)}(\hat{\theta} - \theta), W_3 = \sqrt{\mathrm{E}(I(\hat{\theta}))}(\hat{\theta} - \theta)$, and $W_4 = \sqrt{I(\hat{\theta})}(\hat{\theta} - \theta)$ all had densities which were approximately $\mathcal{N}(0,1)$. Hence they are all approximate pivotal quantities for θ. W_3 and W_4 are the simplest to use in general.

For W_3 the approximate 95% confidence interval is given by $[\hat{\theta} - 1.96/\sqrt{\mathrm{E}I(\hat{\theta})}, \hat{\theta} + 1.96/\sqrt{\mathrm{E}I(\hat{\theta})}]$. For W_4 the approximate 95% confidence interval is given by $[\hat{\theta} - 1.96/\sqrt{I(\hat{\theta})}, \hat{\theta} + 1.96/\sqrt{I(\hat{\theta})}]$. The quantity $1/\sqrt{\mathrm{E}(I(\hat{\theta}))}$ (or $1/\sqrt{I(\hat{\theta})}$) is often referred to as the approximate standard error of the MLE $\hat{\theta}$.

Let $X_1, X_2, \ldots, X_n$ be iid with density $f(x|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ consists of m unknown parameters. Let $\boldsymbol{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ be the MLE of $\boldsymbol{\theta}$. We saw before that for $r = 1, 2, \ldots, m$ the quantities $W_{1r} = (\hat{\theta}_r - \theta_r)/\sqrt{CRLB_r}$ where $CRLB_r$ is the lower bound for $\mathrm{Var}(\hat{\theta}_r)$ given in the generalisation of the Cramer-Rao theorem had a density which was approximately $\mathcal{N}(0,1)$. Recall that $CRLB_r$ is the rth diagonal element of the matrix $[\mathrm{E}(I(\boldsymbol{\theta}))]^{-1}$. In certain cases $CRLB_r$ may depend on the values of unknown parameters other than θ_r and in those cases W_{1r} will not be an approximate pivotal quantity for θ_r.

We also saw that if we define W_{2r} by replacing $CRLB_r$ by the rth diagonal element of the matrix $[I(\boldsymbol{\theta})]^{-1}, W_{3r}$ by replacing $CRLB_r$ by the rth diagonal element of the matrix $[\mathrm{E}I(\hat{\boldsymbol{\theta}})]^{-1}$ and W_{4r} by replacing $CRLB_r$ by the rth diagonal element of the matrix $[I(\hat{\boldsymbol{\theta}})]^{-1}$ we get three more quantities all of whom have a density which is approximately $\mathcal{N}(0,1)$. W_{3r} and W_{4r} only depend on the unknown parameter θ_r and so are approximate pivotal quantities for θ_r. However in certain cases the rth diagonal element of the matrix $[I(\boldsymbol{\theta})]^{-1}$ may depend on the values of unknown parameters other than θ_r and in those cases W_{2r} will not be an approximate pivotal quantity for θ_r. Generally W_{3r} and W_{4r} are most commonly used.

We now examine the use of approximate pivotal quantities based on the MLE in a series of examples :

Example 5.12 (Poisson sampling continued).
Recall that $\hat{\theta} = \bar{x}$ *and* $I(\theta) = \sum_{i=1}^n x_i/\theta^2 = n\hat{\theta}/\theta^2$ *with* $\mathrm{E}[I(\theta)] = n/\theta$. *Hence* $\mathrm{E}[I(\hat{\theta})] = I(\hat{\theta}) = n/\hat{\theta}$ *and the usual approximate* 95% *confidence interval is given by*

$$\left[\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}}{n}}, \quad \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}}{n}} \right].$$

Example 5.13 (Bernoulli trials continued).
Recall that $\hat{\theta} = \bar{x}$ *and*

$$I(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta^2} + \frac{n - \sum_{i=1}^{n} x_i}{(1-\theta)^2}$$

with

$$\mathrm{E}[I(\theta)] = \frac{n}{\theta(1-\theta)}.$$

Hence

$$\mathrm{E}[I(\hat{\theta})] = I(\hat{\theta}) = \frac{n}{\hat{\theta}(1-\hat{\theta})}$$

and the usual approximate 95% *confidence interval is given by*

$$[\ \hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \ \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\ ].$$

5.5 Bootstrap confidence intervals

The *Bootstrap* is a Monte-Carlo method which uses (computer) simulation in lieu of mathematical theory. It is not necessarily simpler. Exercises with the bootstrap are mostly numerical although the underlying theory follows much of the analytical methods.

5.5.1 The empirical cumulative distribution function

An important tool in non-parametric statistics is the *empirical cumulative distribution function* (acronym ecdf) which uses the ordered data as quantiles and probabilities are steps of $\frac{1}{(n+1)}$. We have used the word *empirical* for this plot because it uses only the information in the sample.

Recall that the values for cumulative area that are associated with each datum are determined by the following argument: The sample as collected is denoted by $x_1, x_2, \ldots, x_n$. The subscript represents the position in the list or row in the data file. Bracketed subscripts denote ordering of the data, where $x_{(1)}$ is the smallest, $x_{(2)}$ is the second smallest, $x_{(n)}$ is the largest. In general x_i is not the same datum $x_{(i)}$ but of course this correspondence *could* happen. The n sample points are considered to divide the sampling interval into $(n+1)$ subintervals,

$$(0, x_{(1)}), (x_{(1)}, x_{(2)}), (x_{(2)}, x_{(3)}), \ldots, (x_{(n-1)}, x_{(n)}), (x_{(n)}, \infty)$$

The total area under the density curve (area = 1) has been subdivided into $(n+1)$ subregions with individual areas approximated as $\frac{1}{(n+1)}$. The values of the cumulative area under the density curve is then approximated as

Interval	$(0, x_1)$	(x_1, x_2)	$\ldots$	$(x_{(n-1)}, x_n)$	(x_n, ∞)
Cumulative area	$\frac{0}{(n+1)}$	$\frac{1}{(n+1)}$	$\ldots$	$\frac{n}{(n+1)}$	1

A diagram of this is shown in Fig. 5.5.1.

If there are tied data, say k of them, the step size is $\frac{k}{(n+1)}$. Recall the notion of quantiles and quantile-quantile (Q-Q) plots discussed in Chapter 3.

The following data are the numbers of typhoons in the North Pacific Ocean over 88 years and assume that they are saved in a file called `TYPHOON.txt`

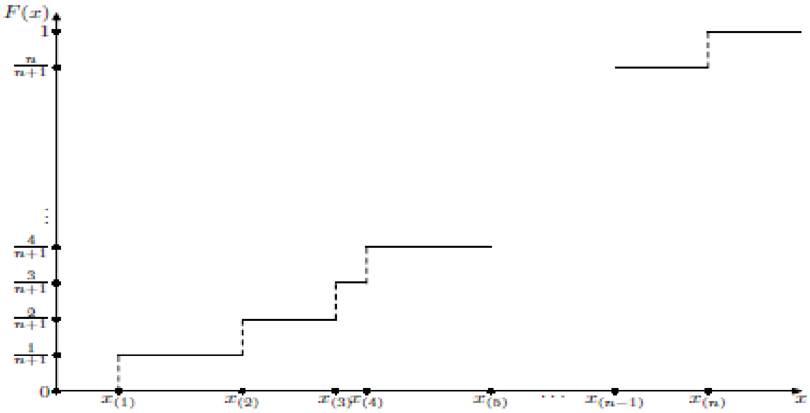


Figure 5.5.1: The ecdf is a step function with step size $\frac{1}{(n+1)}$ between data points.

```
13 7 14 20 13 12 12 15 20 17 11 14 16 12 17 17
16 7 14 15 16 20 17 20 15 22 26 25 27 18 23 26 18 15 20 24
19 25 23 20 24 20 16 21 20 18 20 18 24 27 27 21 21 22 28 38
39 27 26 32 19 33 23 38 30 30 27 25 33 34 16 17 22 17 26 21
30 30 31 27 43 40 28 31 24 15 22 31
```

A plot of the ecdf shown in Figure 5.5.2 and was generated with the following R code:

```
typhoons <- scan("TYPHOON.txt")
plot(ecdf(typhoons),las=1)
```

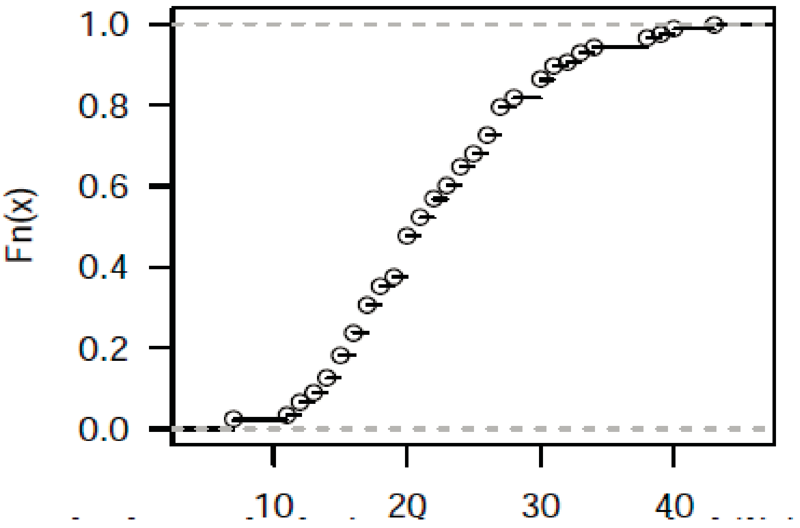


Figure 5.5.2: An empirical distribution function for typhoon data.

The reason why we introduced the concept of an empirical cumulative distribution function is that empirical cumulative distributions are used when calculating bootstrap probabilities. Here is an example. Suppose that in Example 5.5, the data were

```
8 6 5 10 8 12 9 9 8 11 7 3 6 7 5 8 10 7 8 8 10 8 5 10 8 6 10 6 8 14
```

Denote this sample by $x_1, x_2, \ldots, x_n$ where $n = 30$. The summary statistics are

$$\sum_{i=1}^{n} x_i = 240 \quad \overline{X} = 8$$

We shall use this example to illustrate (a) resampling, and (b) the bootstrap distribution.

The sample, $x_1, x_2, \ldots, x_n$, are independently and identically distributed (i.i.d.) as Poisson(λ) which means that each observation is as important as any other for providing information about the population from which this sample is drawn. That infers we can replace any number by one of the others and the "new" sample will still convey the same information about the population. This is also demonstrated in Figure 5.5.3. Three "new" samples have been generated by taking samples of size $n = 30$ with replacement from $\mathbf{x}$. The ecdf of $\mathbf{x}$ is shown in bold and the ecdf's of the "new" samples are shown with different line types. We observe that there is little change in the empirical distributions or estimates of quantiles. If a statistic (e.g. a quantile) were estimated from this process a large number of times, it would be a reliable estimate of the population parameter. The "new" samples are termed *bootstrap samples*.

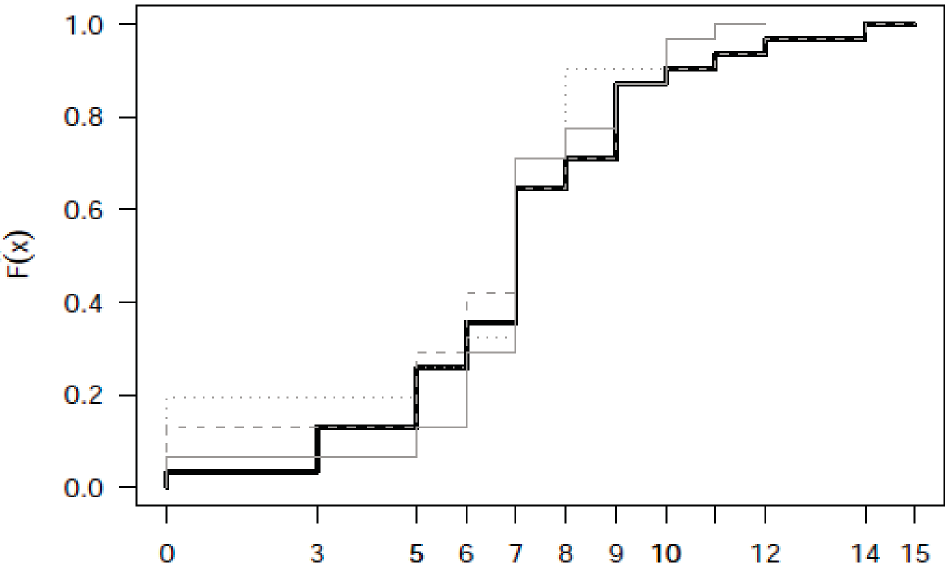


Figure 5.5.3: Resampling with replacement from original sample.

The bootstrap procedure for the CI for λ in the current example then goes as follows:

1. Nominate the number of bootstrap samples that will be drawn, e.g. `nBS=99`.
2. Sample with replacement from $\mathbf{x}$ a bootstrap sample of size n, $\mathbf{x}_1^\star$.
3. For each bootstrap sample, calculate the statistic of interest, $\widehat{\lambda_1^\star}$.
4. Repeat steps 2 and 3 `nBS` times.
5. Use the empirical cumulative distribution function of $\widehat{\lambda_1^\star}, \widehat{\lambda_2^\star}, \ldots, \widehat{\lambda_{nBS}^\star}$ to get the Confidence Interval.

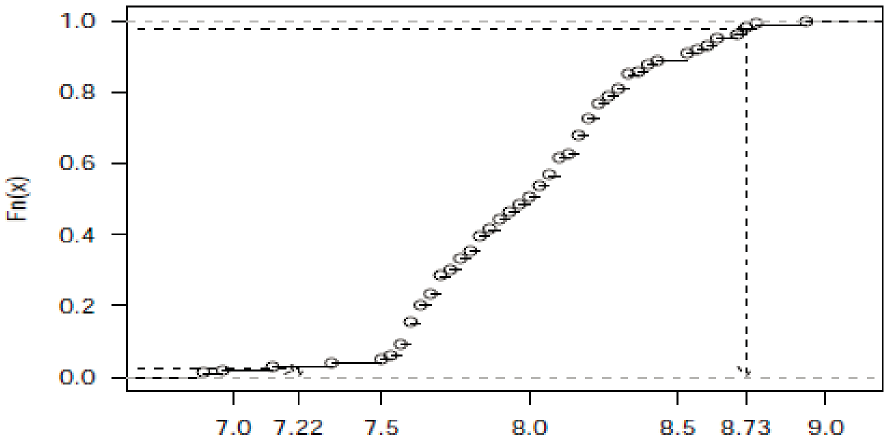


Figure 5.5.4: Deriving the 95% CI from the ecdf of bootstrap estimates of the mean

This is shown in Fig. 5.5.4.

The bootstrap estimate of the 95% CI for λ is $(7.22, 8.73)$. Note that although there is a great deal of statistical theory underpinning this (the ecdf, i.i.d., order statistics, etc.), there is no theoretical formula for the CI and it is determined numerically from the sample.

The R code that generated the graph in Figure 5.5.4, further illustrates the bootstrap approach to construct confidence intervals.

```
x <- c(8,6,5,10,8,12,9,9,8,11,7,3,6,7,5,8,10,7,8,8,10,8,5,10,8,6,10,6,8,14)
n <- length(x)
nBS <- 99 # number of bootstrap simulations
BS.mean <- numeric(nBS)
i <- 1
while (i < (nBS+1) ){
BS.mean[i] <- mean(sample(x,replace=T,size=n))
i <- i + 1
} # end of the while() loop

Quantiles <- quantile(BS.mean,p = c(0.025,0.975))
cat(" 95\% CI = ",Quantiles,"\n")
plot(ecdf(BS.mean),las=1)
```

R has several packages that incorporate bootstrap techniques, including the `boot` package with functions particularly focusing on bootstrapping. The following code uses the `boot` package to construct the same confidence interval as before, but with less effort involved..., although more effort may be required to fully understand how the 'boot' function works...

```
library(boot)
mnz <- function(z,id){mean(z[id])} # user must supply this
bs.samples <- boot(data=x,statistic=mnz,R=99)
boot.ci(bs.samples,conf=0.95,type=c("perc","bca"))
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 99 bootstrap replicates
CALL :
boot.ci(boot.out = bs.samples, conf = 0.95)
Intervals :
Level     Percentile         BCa
95%   ( 7.206, 8.882 ) ( 7.106, 8.751 )
```

It seems that the user must supply a function (e.g. `mnz` here) to generate the bootstrap samples. The variable `id` is recogised by R as a vector `1:length(z)` so that it can draw the samples.