Chapter 4

Point estimation

Contents

4.1 Introduction

In this chapter, we consider how to estimate some characteristic of a population such as its mean μ or variance σ^2, given a sample from that population. By "point" estimate, we mean a single number as the estimate. (In the next chapter, we will consider *interval* estimates: estimates in the form of a *range* of values).

4.2 Estimating a population mean

We wish to estimate the population mean μ using a random sample $X_1, \ldots, X_n$, where we have established that

$$\mathbb{E}(X_j) = \mu, \tag{4.1}$$
$$Var(X_j) = \sigma^2. \tag{4.2}$$

An obvious thing to do would be to estimate μ using

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{4.3}$$

Should we expect to obtain a 'good' estimate of the population mean using only a sample mean? After all, the sample may only constitute a very small proportion of the population.

4.2.1 The problem with estimating a population mean with a sample mean: an example

Consider the PISA example, and the population of test scores in the UK. Suppose we want to estimate the population mean score μ without testing everyone: we take a sample of n 15-year-olds and test them, and we estimate μ using the sample mean. Obviously, different samples of n will produce different sample means and hence different estimates of μ. The concern is therefore whether any single sample is likely to produce a 'good' estimate or not.

We illustrate this in Figure 4.1, with a simple simulation experiment in R. We suppose that the unknown population distribution is $N(\mu = 500, \sigma^2 = 100^2)$, and we generate three samples of size 10 from this distribution. We get three different sample means, the first of which is quite close to μ, but the other two are somewhat further away. Of course, we could try larger sample sizes, but the basic problem remains: a sample mean will (almost certainly) not equal the population mean: how different might it be?

```
# Arrange the plots in a 3x1 array
par(mfrow=c(3,1))
# Use a 'for loop' to repeat the process of drawing a random sample 3 times
for(i in 1:3){
  # Generate a sample of 10 observations from the population distribution
  x <- rnorm(10, 500, 100)
  # Plot the population distribution and the population mean
  curve(dnorm(x, mean=500, sd=100), from = 200, to = 800,
        xlab = "test score", ylab = "density",
```

```
        main = paste("Sample ",i,". Sample mean = ",
                     signif(mean(x), 4), sep=""))
  abline(v=500, lty =2)
  # plot the sample and the sample mean
  points(x, rep(0.0001, 10), pch = 4, col = "red")
  abline(v=mean(x), col = "red")
}
```

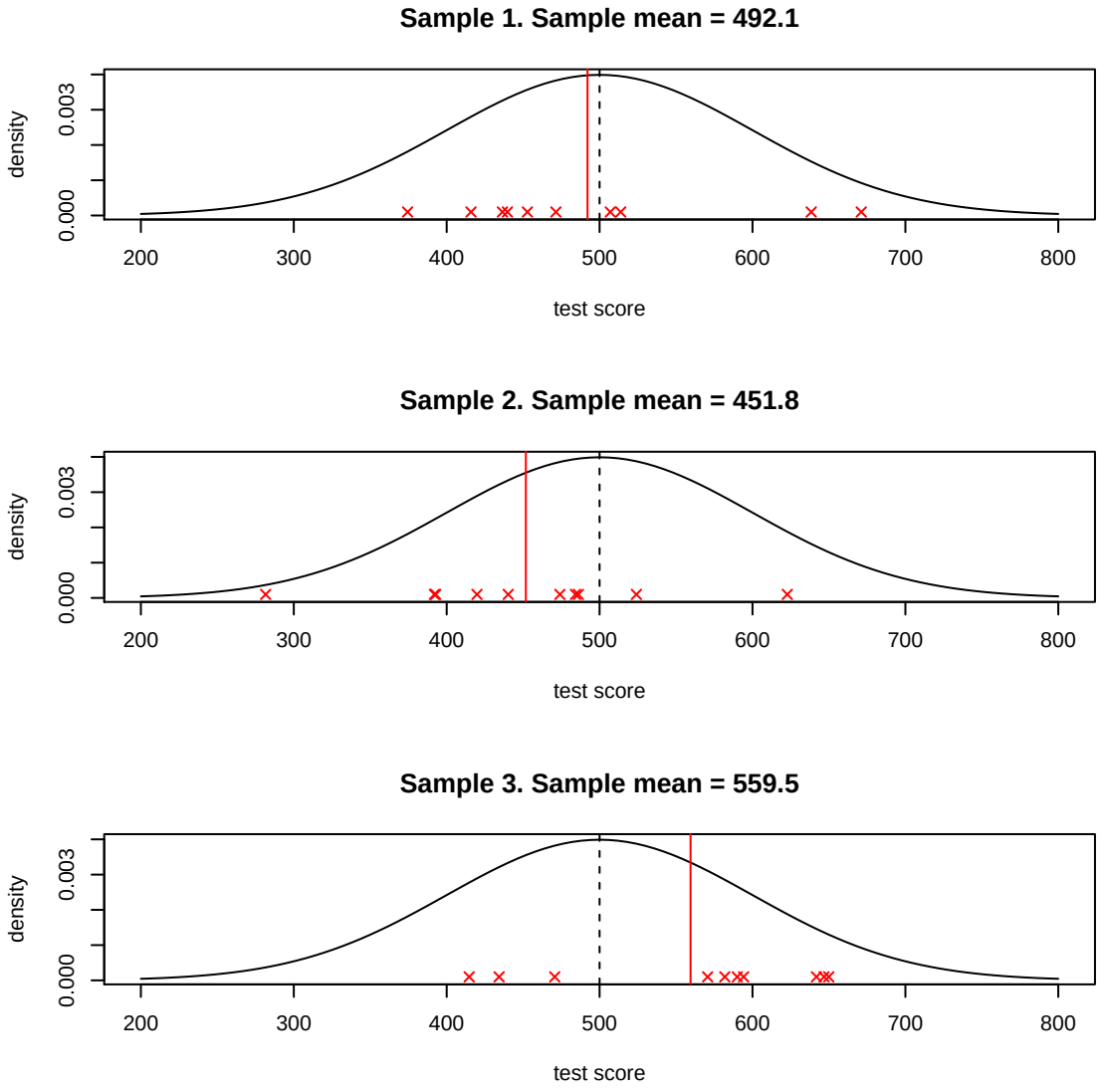

Figure 4.1: The black curve shows the population distribution, with the dashed line indicating the population mean. The red crosses show the individual sampled values, with the red line indication the sample mean. Note the variability in the sample means between the three samples.

4.2.2 Properties of the sample mean

We want to understand how far a sample mean could be from a population mean. If the individual sampled observations $X_1, \ldots, X_n$ are modelled as random variables, the sample mean $\bar{X}$ is a random variable. We already have useful results from Section 16 in the Probability notes which we can use to understand the properties of $\bar{X}$. (You may wish to revise that section). But before we review them...

Confusion alert number 2!
X_i and x_i

A second source of confusion when studying statistics is the 'big X_i, little x_i' notation. We use X_i to represent a **random variable**, and x_i to represent the **observed value** of a random variable. To see why this distinction is important, consider the following.

- Suppose we are *going to* roll a 6-sided die n times. Let X_i be the outcome of the i-th roll of the die. We don't yet know what the outcomes are, so $X_1, \ldots, X_n$ are thought of as random variables, and we can, for example, make statements such as $P(X_i = 2) = \frac{1}{6}$.
- After we roll the die n times, denote the numbers we *actually observe* by $x_1, \ldots, x_n$. For example, if we see a 4 on the ith roll, we would write $x_i = 4$. But we soon get in a mess if we write "$X_i = 4$": if $X_i = 4$ and $P(X_i = 2) = \frac{1}{6}$, then surely $P(4 = 2) = \frac{1}{6}$?! Proper use of notation is important!

Note also:

- any expression involving $X_1, \ldots, X_n$ corresponds to the time **before** we have observed the data: we are using probability to consider what values the data *could* take.
- Any expression involving $x_1, \ldots, x_n$ corresponds to the time **after** we have observed the data: we are presenting calculations to be performed with the values we have observed.
- With this notation, we **never** write expressions such as $\mathbb{E}(x_i) = \mu$, $Var(x_i) = \sigma^2$ etc. The term x_i is a constant, not a random variable, so we would simply have $\mathbb{E}(x_i) = x_i$, $Var(x_i) = 0$.
- $\bar{X}$ is a random variable: it is a function of the random variables $X_1, \ldots, X_n$, but

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  is a constant: $\bar{x}$ is a function of the constants $x_1, \ldots, x_n$.

Hence by 'properties of the sample mean', we mean the properties of the random variable $\bar{X}$, not the properties of a particular number $\bar{x}$.

The expectation of the sample mean

We have

$$\mathbb{E}\left(\bar{X}\right) = \mu, \tag{4.4}$$

We can interpret equation (4.4) to mean that the process of drawing a random sample and estimating μ with the sample mean will give the right answer 'on average': we shouldn't 'expect' an overestimate and we shouldn't 'expect' an underestimate. Note that, assuming we have $\mathbb{E}(X_i) = \mu$ for $i = 1, \ldots, n$, this result always holds; it doesn't matter what the population distribution is.

The variance of the sample mean

We have

$$Var\left(\bar{X}\right) = \frac{\sigma^2}{n}. \tag{4.5}$$

This tells us that as we increase the sample size n, the variance of the sample mean decreases, and so we should expect $\bar{X}$ to be closer to μ as we get more data. This result also always holds, but there is one extra condition: $X_1, \ldots, X_n$ must be independent.

Exercise 4.1. *(Revision of Semester 1 material) Why do we need $X_1, \ldots, X_n$ independent for (4.5) to hold?*

In Figure 4.2, we repeat the simulation experiment from before, but now with random five samples of size 10 with five samples of size 100. Equation (4.5) tells us we should expect to see the sample means closer to the population mean in the case of the larger sample size.

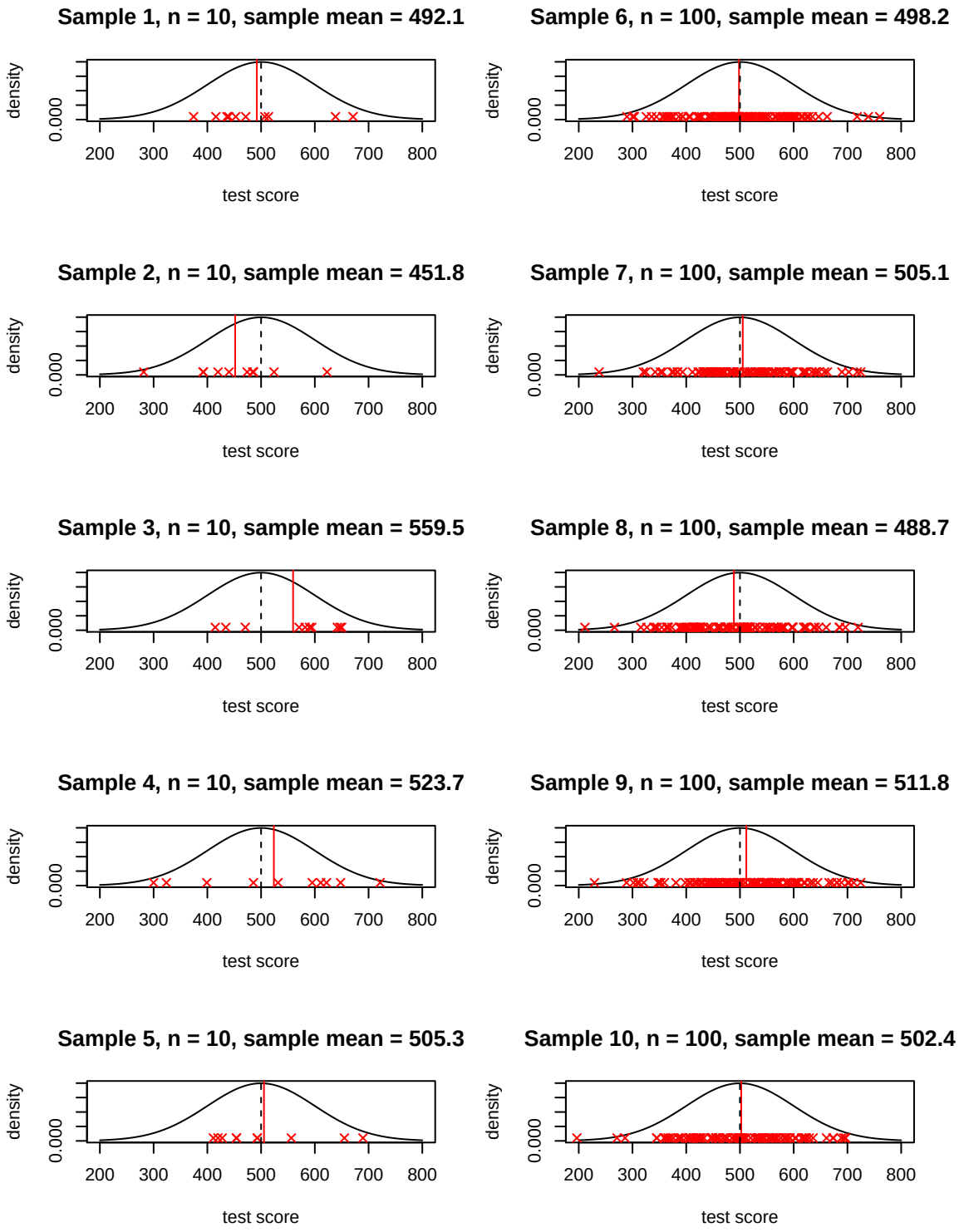

Figure 4.2: The black curve shows the population distribution, with the dashed line indicating the population mean. The red crosses show the individual sampled values, with the red line indication the sample mean. Note the smaller variability in sample means in the second column, where a larger sample size has been used.

The distribution of the sample mean

If the sample size is 'large enough', we have the additional result that, by the Central Limit Theorem, $\bar{X}$ is approximately normally distributed:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \tag{4.6}$$

even if the population distribution is **not** normal. We'll see an illustration for this in a non-normal population distribution shortly.

4.2.3 How large does the sample need to be to get a 'good' estimate?

This is somewhat like asking how long is a piece of string! There is no single answer to how large a sample should be; it will depend on how accurately we want to estimate the population mean μ, and that will depend on the context. We will revisit the topic of sample sizes in later chapters, but there is one special case we can consider now, for exponentially distributed populations.

Estimation error

We will use the term "estimation error" to mean the difference

$$\bar{X} - \mu$$

We won't be able to calculate this, but we can consider how large an estimation error would be acceptable. Sometimes, it will make more sense to consider estimation errors relative to μ: we will use the term "relative estimation error" to mean

$$\frac{\bar{X} - \mu}{\mu},$$

e.g. we would want to estimate μ within 10% of its true value.

Exercise 4.2. *A new drug has been developed as a "second-line" treatment for bowel cancer (a drug only to be used when the initial treatment stops working). Once a patient starts treatment on this new drug, it is supposed that their survival time (in months) will have an exponential distribution*[1]*, with unknown rate parameter λ.*

We wish to estimate the mean survival time on this drug, and would like the absolute relative estimation error $|\bar{X} - \mu|/\mu$ to be no more than 10%. How many patients would we need to observe, such that there is a 95% chance this error will be less than 10%? Use the following R output to help you, and assume that Equation (4.6) holds.

```
qnorm(0.975)

## [1] 1.959964
```

[1] the analysis of survival time data is a major topic in medical statistics, and more complex models/methods are taught in MAS361

We will test our answer (385 patients) with an experiment in R:

1. We suppose that the population distribution is $Exp(rate = 1/12)$ (this is so we can generate random samples of data - otherwise we'll pretend $\lambda = 1/12$ is unknown to us.) This gives a true value for the population mean as 12 months.

2. We generate a large number (1000) of samples of size 385, and for each sample, we calculate the sample mean.

3. We will count how many times the estimation error is more than 0.1×12; we'd expect this to be happen about 50 times out of 1000.

4. We will also show a histogram of the 1000 sample means, together with the population distribution (the density function of the $Exp(rate = 1/12)$). Based on the Central Limit Theorem, the distribution of the sample mean is approximately normal, so we would expect this histogram to resemble a normal distribution, even though the population distribution is exponential.

```
population.mean <- 12
n <- 385
# Generate sample data
X <- matrix(rexp(n * 1000, 1/population.mean), nrow = n, ncol = 1000)
# Each column of x represents one sample of size 385
# Calculate the 1000 sample means by calculating the mean of each column
sample.means <- colMeans(X)
# How many sample means were outside the range [10.8, 13.2]?
sum(sample.means < 10.8) + sum(sample.means > 13.2)

## [1] 57

# Compare the distribution of the sample means with the population distribution
hist(sample.means, xlim=c(0, 40), prob = T, xlab = "survival time (months)",
     main = "histogram of sample means")
curve(dexp(x, rate= 1/population.mean),
      from = 0, to =1000, add = T, col = "red", n =301)
abline(v = population.mean, col = "blue", lwd = 2)
```

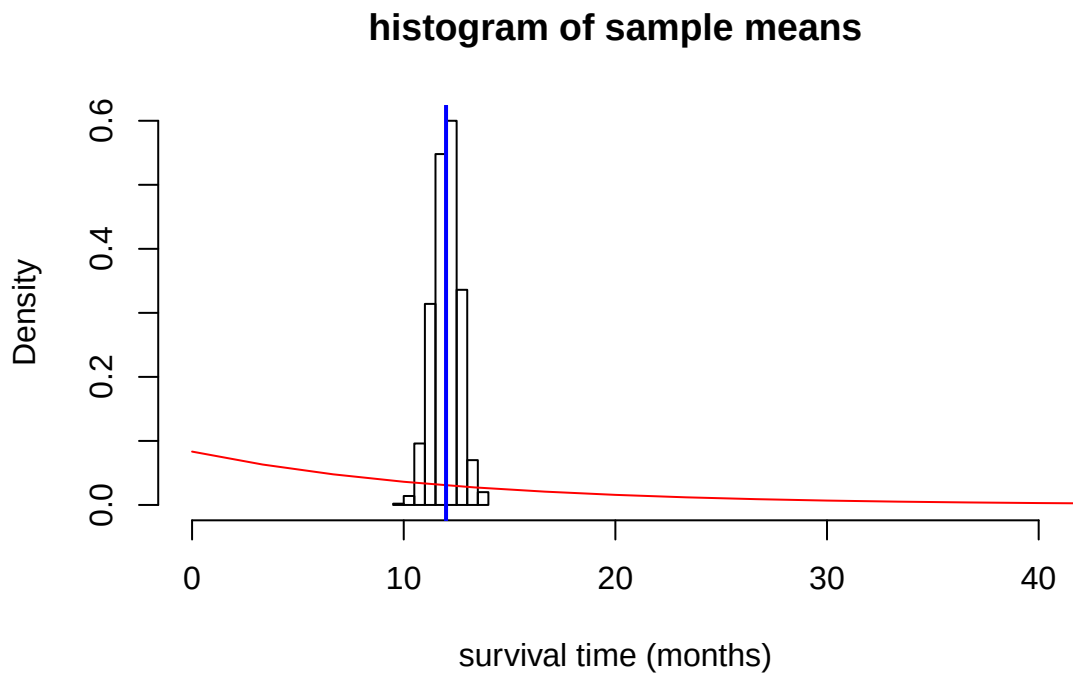

Figure 4.3: The red line shows the population distribution (an exponential distribution), with the blue line indicating the population mean. The histogram shows the distribution of 1000 separate sample means, with each sample of size 385. Note how the histogram is centred around the population mean, and suggests a normal distribution, even though the population distribution is exponential. In this example, 57 samples out of 1000 gave absolute estimation errors more than 10% of the population mean, roughly as we were expecting.

4.3 Point estimation: general theory

Having looked at the case of estimating a population mean, we now introduce some general concepts and theory of point estimation.

Let θ be some *parameter* of the population that we seek to estimate. So far, we have considered $\theta = \mu$, the population mean, but we will consider other parameters later. We wish to estimate θ using a random sample $X_1, X_2, \ldots, X_n$ from the population. An *estimator* of θ is defined to be a function $\hat{\theta}$ of the random variables $X_1, X_2, \ldots, X_n$. So we could write:

$$\hat{\theta} = f(X_1, X_2, \ldots, X_n). \tag{4.7}$$

In the case $\theta = \mu$, we have considered

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{4.8}$$

but there are circumstances in which we might use a different function to estimate μ, and in general there may be different estimators we might consider for a population parameter θ.

We now consider some conditions/properties for which we might judge $\hat{\theta}$ to be a good estimator.

4.3.1 Unbiased Estimators

We define the *bias* $B(\hat{\theta}, \theta)$ of an estimator to be the difference between its expected value and the true value of the parameter we are trying to estimate:

$$B(\hat{\theta}, \theta) = \mathbb{E}(\hat{\theta}) - \theta, \tag{4.9}$$

and we say that an estimator is *unbiased* if it has zero bias, i.e. if

$$\mathbb{E}(\hat{\theta}) = \theta. \tag{4.10}$$

We have already seen that the sample mean is an unbiased estimator of the population mean.

4.3.2 The standard error of an estimator

We have a special term for the standard deviation of an estimator: we refer to it as the *standard error* and denote it by $\text{SE}(\hat{\theta})$. Hence we have the distinction:

- if we have some random variables, $X_1, X_2, \ldots, X_n$, we may refer to the *standard deviation of a single variable* X_i, defined as $\sqrt{Var(X_i)}$.
- if we have a function $\hat{\theta} = f(X_1, \ldots, X_n)$, used to estimate some parameter of the population distribution, we may refer to the *standard error of the estimator*, defined as $\sqrt{Var(\hat{\theta})}$.

Again, we have already seen that the standard error of the sample mean is σ^2/n. If $\hat{\theta}$ is an unbiased estimator, then the smaller the standard error, the more likely $\hat{\theta}$ will be close to what we want to estimate: θ.

4.3.3 The sampling distribution of an estimator

Thinking of $\hat{\theta}$ as a random variable, we refer to its probability distribution as the *sampling distribution*. The Central Limit Theorem tells that for large n, the sampling distribution of the sample mean is approximately $N(\mu, \sigma^2/n)$

4.3.4 Consistent estimators

The final property of estimators that we'll briefly consider is consistency, which is concerned with whether and estimator is likely to be arbitrarily close (however close we choose to specify) to the true value as the sample size increases

Suppose that we have a sequence $\hat{\theta}_n$ of estimators of θ, where the label n corresponds to increasing sample size. For example, if estimating a population mean μ with a sample mean, the sequence of estimators $\hat{\theta}_n$ would represent a sequence of sample means based on 1,2,3, ... observations. Informally, we say that these are *consistent* if the probability that $\hat{\theta}_n$ is close to θ is close to 1, for large n.

More precisely, for the sequence of estimators to be consistent, we require that for any $\epsilon > 0$ (no matter how small it is)

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1. \tag{4.11}$$

The next result gives us a convenient method for finding consistent estimators.

Theorem 4.1. *Suppose that $(\hat{\theta}_n)$ are unbiased estimators of θ. If*

$$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0,$$

then $\hat{\theta}_n$ are consistent.

This theorem enables us to check quickly whether an unbiased estimator is consistent, without needing to directly evaluate the limit of $P(|\hat{\theta}_n - \theta| < \epsilon)$.

Exercise 4.3. *Show that $\bar{X}$ is a consistent estimator for μ.*

Be aware that, in general, it is possible to find consistent estimators that are biased, and unbiased estimators that fail to be consistent.

4.4 Estimating a population variance

In this section we will aim to find an unbiased estimator for the population variance σ^2. If μ was known, a candidate for an estimator of σ^2 is:

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{4.12}$$

Exercise 4.4. *Show that S_1^2 is unbiased for σ^2.*

However, in most applications we will not know μ. One idea is to replace μ by its estimator $\bar{X}$, and so to consider as an estimator for σ^2

$$S_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

It turns out that S_2^2 is not an unbiased estimator for σ^2. The reason is that we have lost some information by estimating μ. We slightly modify S_2^2 to

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{4.13}$$

Note the difference between S^2 and S_2^2.

Theorem 4.2. *S^2 is an unbiased estimator for σ^2, i.e.*

$$\mathbb{E}(S^2) = \sigma^2. \tag{4.14}$$

It can also be shown to be a consistent estimator, but we will not consider the proof here.

4.4.1 Computing sample variances

We use the notation s^2 to denote the observed sample variance, computed once we have the data:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{4.15}$$

It can be helpful (if computing s^2 by hand) to note the following.

Theorem 4.3.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \tag{4.16}$$

Note that Equation (4.15) is the formula R uses in the `var` command:

```
x <- c(10, 17, 8, 22, 15)
var(x)

## [1] 31.3

sum((x - mean(x))^2) / 4

## [1] 31.3

(sum(x^2) - 5 * mean(x)^2) / 4

## [1] 31.3
```

Exercise 4.5. *Explain the difference between σ^2, S^2 and s^2.*