Part II

Statistical Inference

Chapter 9

Introduction to Statistical Inference

9.1 Student Learning Objectives

The next section of this chapter introduces the basic issues and tools of statistical inference. These tools are the subject matter of the second part of this book. In Chapters 9–15 we use data on the specifications of cars in order to demonstrate the application of the tools for making statistical inference. In the third section of this chapter we present the data frame that contains this data. The fourth section reviews probability topics that were discussed in the first part of the book and are relevant for the second part. By the end of this chapter, the student should be able to:

- Define key terms that are associated with inferential statistics.
- Recognize the variables of the "`cars.csv`" data frame.
- Revise concepts related to random variables, the sampling distribution and the Central Limit Theorem.

9.2 Key Terms

The first part of the book deals with descriptive statistics and with probability. In descriptive statistics one investigates the characteristics of the data by using graphical tools and numerical summaries. The frame of reference is the observed data. In probability, on the other hand, one extends the frame of reference to include all data sets that could have potentially emerged, with the observed data as one among many.

The second part of the book deals with inferential statistics. The aim of statistical inference is to gain insight regarding the population parameters from the observed data. The method for obtaining such insight involves the application of formal computations to the data. The interpretation of the outcome of these formal computations is carried out in the probabilistic context, in which one considers the application of these formal computations to all potential data sets. The justification for using the specific form of computation on the observed

data stems from the examination of the probabilistic properties of the formal computations.

Typically, the formal computations will involve statistics, which are functions of the data. The assessment of the probabilistic properties of the computations will result from the sampling distribution of these statistics.

An example of a problem that requires statistical inference is the estimation of a parameter of the population using the observed data. *Point estimation* attempts to obtain the best guess to the value of that parameter. An *estimator* is a statistic that produces such a guess. One may prefer an estimator whose sampling distribution is more concentrated about the population parameter value over another estimator whose sampling distribution is less so. Hence, the justification for selecting a specific statistic as an estimator is a consequence of the probabilistic characteristics of this statistic in the context of the sampling distribution.

For example, a car manufacture may be interested in the fuel consumption of a new type of car. In order to do so the manufacturer may apply a standard test cycle to a sample of 10 new cars of the given type and measure their fuel consumptions. The parameter of interest is the average fuel consumption among *all* cars of the given type. The average consumption of the 10 cars is a point estimate of the parameter of interest.

An alternative approach for the estimation of a parameter is to construct an interval that is most likely to contain the population parameter. Such an interval, which is computed on the basis of the data, is called the a *confidence interval*. The sampling probability that the confidence interval will indeed contain the parameter value is called the *confidence level*. Confidence intervals are constructed so as to have a prescribed confidence level.

A different problem in statistical inference is *hypothesis testing*. The scientific paradigm involves the proposal of new theories and hypothesis that presumably provide a better description for the laws of Nature. On the basis of these hypothesis one may propose predictions that can be examined empirically. If the empirical evidence is consistent with the predictions of the new hypothesis but not with those of the old theory then the old theory is rejected in favor of the new one. Otherwise, the established theory maintains its status. Statistical hypothesis testing is a formal method for determining which of the two hypothesis should prevail that uses this paradigm.

Each of the two hypothesis, the old and the new, predicts a different distribution for the empirical measurements. In order to decide which of the distributions is more in tune with the data a statistic is computed. This statistic is called the *test statistic*. A threshold is set and, depending on where the test statistic falls with respect to this threshold, the decision is made whether or not to reject the old theory in favor of the new one.

This decision rule is not error proof, since the test statistic may fall by chance on the wrong side of the threshold. Nonetheless, by the examination of the sampling distribution of the test statistic one is able to assess the probability of making an error. In particular, the probability of erroneously rejecting the currently accepted theory (the old one) is called the *significance level* of the test. Indeed, the threshold is selected in order to assure a small enough significance level.

Returning to the car manufacturer. Assume that the car in question is manufactured in two different factories. One may want to examine the hypothesis

that the car's fuel consumption is the same for both factories. If 5 of the tested cars were manufactured in one factory and the other 5 in the other factory then the test may be based on the absolute value of the difference between the average consumption of the first 5 and the average consumption of the other 5.

The method of testing hypothesis is also applied in other practical settings where it is required to make decisions. For example, before a new treatment to a medical condition is approved for marketing by the appropriate authorities it must undergo a process of objective testing through clinical trials. In these trials the new treatment is administered to some patients while other obtain the (currently) standard treatment. Statistical tests are applied in order to compare the two groups of patient. The new treatment is released to the market only if it is shown to be beneficial with statistical significance and it is shown to have no unacceptable side effects.

In subsequent chapters we will discuss in more details the computation of point estimation, the construction of confidence intervals, and the application of hypothesis testing. The discussion will be initiated in the context of a single measurement but will later be extended to settings that involve comparison of measurements.

An example of such analysis is the analysis of clinical trials where the response of the patients treated with the new procedure is compared to the response of patients that were treated with the conventional treatment. This comparison involves the same measurement taken for two sub-samples. The tools of statistical inference – hypothesis testing, point estimation and the construction of confidence intervals – may be used in order to carry out this comparison.

Other comparisons may involve two measurements taken for the entire sample. An important tool for the investigation of the relations between two measurements, or variables, is *regression*. Models of regression describe the change in the distribution in one variable as a function of the other variable. Again, point estimation, confidence intervals, and hypothesis testing can be carried out in order to examine regression models. The variable whose distribution is the target of investigation is called the response. The other variable that may affect that distribution is called the explanatory variable.

9.3 The Cars Data Set

Statistical inference is applied to data in order to address specific research question. We will demonstrate different inferential procedures using a specific data set with the aim of making the discussion of the different procedures more concrete. The same data set will be used for all procedures that are presented in Chapters 10–15[1]. This data set contains information on various models of cars and is stored in the CVS file "`cars.csv`"[2]. The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/cars.csv`. You are advised to download this file to your computer and store it in the working directory of `R`.

[1] Other data sets will be used in Chapter 16 and in the quizzes and assignments.

[2] The original "Automobiles" data set is accessible at the UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml`). This data was assembled by Jeffrey C. Schlimmer, using as source the 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook. The current file "`cars.csv`" is based on all 205 observations of the original data set. We selected 17 of the 26 variables available in the original source.

Let us read the content of the CSV file into an R data frame and produce a brief summary:

```
> cars <- read.csv("cars.csv")
> summary(cars)
         make       fuel.type   num.of.doors       body.style
 toyota    : 32   diesel: 20   four:114     convertible: 6
 nissan    : 18   gas   :185   two : 89     hardtop    : 8
 mazda     : 17                NA's:  2     hatchback  :70
 honda     : 13                             sedan      :96
 mitsubishi: 13                             wagon      :25
 subaru    : 12
 (Other)   :100
 drive.wheels engine.location   wheel.base         length
 4wd:  9      front:202       Min.   : 86.60   Min.   :141.1
 fwd:120      rear :  3       1st Qu.: 94.50   1st Qu.:166.3
 rwd: 76                      Median : 97.00   Median :173.2
                              Mean   : 98.76   Mean   :174.0
                              3rd Qu.:102.40   3rd Qu.:183.1
                              Max.   :120.90   Max.   :208.1

     width           height       curb.weight    engine.size
 Min.   :60.30   Min.   :47.80   Min.   :1488   Min.   : 61.0
 1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145   1st Qu.: 97.0
 Median :65.50   Median :54.10   Median :2414   Median :120.0
 Mean   :65.91   Mean   :53.72   Mean   :2556   Mean   :126.9
 3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935   3rd Qu.:141.0
 Max.   :72.30   Max.   :59.80   Max.   :4066   Max.   :326.0

   horsepower       peak.rpm       city.mpg      highway.mpg
 Min.   : 48.0   Min.   :4150   Min.   :13.00   Min.   :16.00
 1st Qu.: 70.0   1st Qu.:4800   1st Qu.:19.00   1st Qu.:25.00
 Median : 95.0   Median :5200   Median :24.00   Median :30.00
 Mean   :104.3   Mean   :5125   Mean   :25.22   Mean   :30.75
 3rd Qu.:116.0   3rd Qu.:5500   3rd Qu.:30.00   3rd Qu.:34.00
 Max.   :288.0   Max.   :6600   Max.   :49.00   Max.   :54.00
 NA's   :  2.0   NA's   :   2
     price
 Min.   : 5118
 1st Qu.: 7775
 Median :10295
 Mean   :13207
 3rd Qu.:16500
 Max.   :45400
 NA's   :    4
```

Observe that the first 6 variables are factors, i.e. they contain qualitative data that is associated with categorization or the description of an attribute. The last 11 variable are numeric and contain quantitative data.

Factors are summarized in R by listing the attributes and the frequency of each attribute value. If the number of attributes is large then only the most

frequent attributes are listed. Numerical variables are summarized in R with the aid of the smallest and largest values, the three quartiles (Q1, the median, and Q3) and the average (mean).

The third factor variable, "num.of.doors", as well as several of the numerical variables have a special category titled "NA's". This category describes the number of missing values among the observations. For a given variable, the observations for which a value for the variable is not recorded, are marked as missing. R uses the symbol "NA" to identify a missing value[3].

Missing observations are a concern in the analysis of statistical data. If the relative frequency of missing values is substantial and the reason for not obtaining the data for specific observations is related to the phenomena under investigation than naïve statistical inference may produce biased conclusions. In the "cars" data frame missing values are less of a concern since their relative frequency is low.

One should be on the lookout for missing values when applying R to data since the different functions may have different ways for dealing with missing values. One should make sure that the appropriate way is applied for the specific analysis.

Consider the variables of the data frame "cars":

make: The name of the car producer (a factor).

fuel.type: The type of fuel used by the car, either diesel or gas (a factor).

num.of.doors: The number of passenger doors, either two or four (a factor).

body.style: The type of the car (a factor).

drive.wheels: The wheels powered by the engine (a factor).

engine.location: The location in the car of the engine (a factor).

wheel.base: The distance between the centers of the front and rear wheels in inches (numeric).

length: The length of the body of the car in inches (numeric).

width: The width of the body of the car in inches (numeric).

height: The height of the car in inches (numeric).

curb.weight: The total weight in pounds of a vehicle with standard equipment and a full tank of fuel, but with no passengers or cargo (numeric).

engine.size: The volume swept by all the pistons inside the cylinders in cubic inches (numeric).

horsepower: The power of the engine in horsepowers (numeric).

peak.rpm: The top speed of the engine in rounds-per-minute (numeric).

city.mpg: The fuel consumption of the car in city driving conditions, measured as miles per gallon of fuel (numeric).

[3]Indeed, if you scan the CSV file directly by opening it with a spreadsheet then every now and again you will encounter this symbol.

`highway.mpg`: The fuel consumption of the car in highway driving conditions, measured as miles per gallon of fuel (numeric).

`price`: The retail price of the car in US Dollars (numeric).

9.4 The Sampling Distribution

9.4.1 Statistics

Statistical inferences, be it point estimation, confidence intervals, or testing hypothesis, are based on statistics computed from the data. Examples of statistics are the sample average and the sample standard deviation. These are important examples, but clearly not the only ones. Given numerical data, one may compute the smallest value, the largest value, the quartiles, and the median. All are examples of statistics. Statistics may also be associated with factors. The frequency of a given attribute among the observations is a statistic. (An example of such statistic is the frequency of diesel cars in the data frame.) As part of the discussion in the subsequent chapters we will consider these and other types of statistics.

Any statistic, when computed in the context of the data frame being analyzed, obtains a single numerical value. However, once a sampling distribution is being considered then one may view the same statistic as a random variable. A statistic is a function or a formula which is applied to the data frame. Consequently, when a random collection of data frames is the frame of reference then the application of the formula to each of the data frames produces a random collection of values, which is the sampling distribution of the statistic.

We distinguish in the text between the case where the statistic is computed in the context of the given data frame and the case where the computation is conducted in the context of the random sample. This distinguishing is emphasized by the use of small letters for the former and capital letters for the later. Consider, for example, the sample average. In the context of the observed data we denote the data values for a specific variable by $x_1, x_2, \ldots, x_n$. The sample average computed for these values is denoted by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \, .$$

On the other hand, if the discussion of the sample average is conducted in the context of a random sample then the sample is a sequence $X_1, X_2, \ldots, X_n$ of random variables. The sample average is denoted in this context as

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \, .$$

The same formula that was applied to the data values is applied now to the random components of the random sample. In the first context $\bar{x}$ is an observed non-random quantity. In the second context $\bar{X}$ is a random variable, an abstract mathematical concept.

A second example is the sample variance. When we compute the sample variance for the observed data we use the formula:

$$s^2 = \frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample} - 1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \, .$$

However, when we discuss the sampling distribution of the sample variance we apply the same formula to the random sample:

$$S^2 = \frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample} - 1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \ .$$

Again, S^2 is a random variable whereas s^2 is a non-random quantity: The evaluation of the random variable at the specific sample that is being observed.

9.4.2 The Sampling Distribution

The sampling distribution may emerge as random selection of samples from a particular population. In such a case, the sampling distribution of the sample, and hence of the statistic, is linked to the distribution of values of the variable in the population.

Alternatively, one may assign theoretical distribution to the measurement associated with the variable. In this other case the sampling distribution of the statistic is linked to the theoretical model.

Consider, for example, the variable "`price`" that describes the prices of the 205 car types (with 4 prices missing) in the data frame "`cars`". In order to define a sampling distribution one may imagine a larger population of car types, perhaps all the car types that were sold during the 80's in the United States, or some other frame of reference, with the car types that are included in the data frame considered as a random sample from that larger population. The observed sample corresponds to car types that where sold in 1985. Had one chosen to consider car types from a different year then one may expect to obtain other evaluations of the price variable. The reference population, in this case, is the distribution of the prices of the car types that were sold during the 80's and the sampling distribution is associated with a random selection of a particular year within this period and the consideration of prices of car types sold in that year. The data for 1985 is what we have at hand. But in the sampling distribution we take into account the possibility that we could have obtained data for 1987, for example, rather than the data we did get.

An alternative approach for addressing sampling distribution is to consider a theoretical model. Referring again to the variable "`price`" one may propose an Exponential model for the distribution of the prices of cars. This model implies that car types in the lower spectrum of the price range are more frequent than cars with a higher price tag. With this model in mind, one may propose the sampling distribution to be composed of 205 unrelated copies from the Exponential distribution (or 201 if we do not want to include the missing values). The rate λ of the associated Exponential distribution is treated as an unknown parameter. One of the roles of statistical inference is to estimate the value of this parameter with the aid of the data at hand.

Sampling distribution is relevant also for factor variables. Consider the variable "`fuel.type`" as an example. In the given data frame the frequency of diesel cars is 20. However, had one considered another year during the 80's one may have obtained a different frequency, resulting in a sampling distribution. This type of sampling distribution refers to all cars types that were sold in the United States during the 80's as the frame of reference.

Alternatively, one may propose a theoretical model for the sampling distribution. Imagine there is a probability p that a car runs on diesel (and probability

$1 - p$ that it runs on gas). Hence, when one selects 205 car types at random then one obtains that the distribution of the frequency of car types that run on diesel has the Binomial$(205, p)$ distribution. This is the sampling distribution of the frequency statistic. Again, the value of p is unknown and one of our tasks is to estimate it from the data we observe.

In the context of statistical inference the use of theoretical models for the sampling distribution is the standard approach. There are situation, such as the application surveys to a specific target population, where the consideration of the entire population as the frame of reference is more natural. But, in most other applications the consideration of theoretical models is the method of choice. In this part of the book, where we consider statistical inference, we will always use the theoretical approach for modeling the sampling distribution.

9.4.3 Theoretical Distributions of Observations

In the first part of the book we introduced several theoretical models that may describe the distribution of an observation. Let us take the opportunity and review the list of models:

Binomial: The Binomial distribution is used in settings that involve counting the number of occurrences of a particular outcome. The parameters that determine the distribution are n, the number of observations, and p, the probability of obtaining the particular outcome in each observation. The expression "Binomial(n, p)" is used to mark the Binomial distribution. The sample space for this distribution is formed by the integer values $\{0, 1, 2, \ldots, n\}$. The expectation of the distribution is np and the variance is $np(1 - p)$. The functions "`dbinom`", "`pbinom`", and "`qbinom`" may be used in order to compute the probability, the cumulative probability, and the percentiles, respectively, for the Binomial distribution. The function "`rbinom`" can be used in order to simulate a random sample from this distribution.

Poisson: The Poisson distribution is also used in settings that involve counting. This distribution approximates the Binomial distribution when the number of examinations n is large but the probability p of the particular outcome is small. The parameter that determines the distribution is the expectation λ. The expression "Poisson(λ)" is used to mark the Poisson distribution. The sample space for this distribution is the entire collection of natural numbers $\{0, 1, 2, \ldots\}$. The expectation of the distribution is λ and the variance is also λ. The functions "`dpois`", "`ppois`", and "`qpois`" may be used in order to compute the probability, the cumulative probability, and the percentiles, respectively, for the Poisson distribution. The function "`rpois`" can be used in order to simulate a random sample from this distribution.

Uniform: The Uniform distribution is used in order to model measurements that may have values in a given interval, with all values in this interval equally likely to occur. The parameters that determine the distribution are a and b, the two end points of the interval. The expression "Uniform(a, b)" is used to identify the Uniform distribution. The sample space for this distribution is the interval $[a, b]$. The expectation of the distribution is

$(a+b)/2$ and the variance is $(b-a)^2/12$. The functions "`dunif`", "`punif`", and "`qunif`" may be used in order to compute the density, the cumulative probability, and the percentiles for the Uniform distribution. The function "`runif`" can be used in order to simulate a random sample from this distribution.

Exponential: The Exponential distribution is frequently used to model times between events. It can also be used in other cases where the outcome of the measurement is a positive number and where a smaller value is more likely than a larger value. The parameter that determines the distribution is the rate λ. The expression "Exponential(λ)" is used to identify the Exponential distribution. The sample space for this distribution is the collection of positive numbers. The expectation of the distribution is $1/\lambda$ and the variance is $1/\lambda^2$. The functions "`dexp`", "`pexp`", and "`qexp`" may be used in order to compute the density, the cumulative probability, and the percentiles, respectively, for the Exponential distribution. The function "`rexp`" can be used in order to simulate a random sample from this distribution.

Normal: The Normal distribution frequently serves as a generic model for the distribution of a measurement. Typically, it also emerges as an approximation of the sampling distribution of statistics. The parameters that determine the distribution are the expectation μ and the variance σ^2. The expression "Normal(μ, σ^2)" is used to mark the Normal distribution. The sample space for this distribution is the collection of all numbers, negative or positive. The expectation of the distribution is μ and the variance is σ^2. The functions "`dnorm`", "`pnorm`", and "`qnorm`" may be used in order to compute the density, the cumulative probability, and the percentiles for the Normal distribution. The function "`rnorm`" can be used in order to simulate a random sample from this distribution.

9.4.4 Sampling Distribution of Statistics

Theoretical models describe the distribution of a measurement as a function of a parameter, or a small number of parameters. For example, in the Binomial case the distribution is determined by the number of trials n and by the probability of success in each trial p. In the Poisson case the distribution is a function of the expectation λ. For the Uniform distribution we may use the end-points of the interval, a and b, as the parameters. In the Exponential case, the rate λ is a natural parameter for specifying the distribution and in Normal case the expectation μ and the variance σ^2 my be used for that role.

The general formulation of statistical inference problems involves the identification of a theoretical model for the distribution of the measurements. This theoretical model is a function of a parameter whose value is unknown. The goal is to produce statements that refer to this unknown parameter. These statements are based on a sample of observations from the given distribution.

For example, one may try to guess the value of the parameter (point estimation), one may propose an interval which contains the value of the parameter with some subscribed probability (confidence interval) or one may test the hypothesis that the parameter obtains a specific value (hypothesis testing).

The vehicles for conducting the statistical inferences are statistics that are computed as a function of the measurements. In the case of point estimation these statistics are called *estimators*. In the case where the construction of an interval that contains the value of the parameter is the goal then the statistics are called *confidence interval*. In the case of testing hypothesis these statistics are called *test statistics*.

In all cases of inference, The relevant statistic possesses a distribution that it inherits from the sampling distribution of the observations. This distribution is the sampling distribution of the statistic. The properties of the statistic as a tool for inference are assessed in terms of its sampling distribution. The sampling distribution of a statistic is a function of the sample size and of the parameters that determine the distribution of the measurements, but otherwise may be of complex structure.

In order to assess the performance of the statistics as agents of inference one should be able to determine their sampling distribution. We will apply two approaches for this determination. One approach is to use a Normal approximation. This approach relies on the Central Limit Theorem. The other approach is to simulate the distribution. This other approach relies on the functions available in `R` for the simulation of a random sample from a given distribution.

9.4.5 The Normal Approximation

In general, the sampling distribution of a statistic is not the same as the sampling distribution of the measurements from which it is computed. For example, if the measurements are from the Uniform distributed then the distribution of a function of the measurements will, in most cases, not possess the Uniform distribution. Nonetheless, in many cases one may still identify, at least approximately, what the sampling distribution of the statistic is.

The most important scenario where the limit distribution of the statistic has a known shape is when the statistic is the sample average or a function of the sample average. In such a case the Central Limit Theorem may be applied in order to show that, at least for a sample size not too small, the distribution of the statistic is approximately Normal.

In the case where the Normal approximation may be applied then a probabilistic statement associated with the sampling distribution of the statistic can be substituted by the same statement formulated for the Normal distribution. For example, the probability that the statistic falls inside a given interval may be approximated by the probability that a Normal random variable with the same expectation and the same variance (or standard deviation) as the statistic falls inside the given interval.

For the special case of the sample average one may use the fact that the expectation of the average of a sample of measurements is equal to the expectation of a single measurement and the fact that the variance of the average is the variance of a single measurement, divided by the sample size. Consequently, the probability that the sample average falls within a given interval may be approximate by the probability of the same interval according to the Normal distribution. The expectation that is used for the Normal distribution is the expectation of the measurement. The standard deviation is the standard deviation of the measurement, divided by the square root of the number of observations.

The Normal approximation of the distribution of a statistic is valid for cases other than the sample average or functions thereof. For example, it can be shown (under some conditions) that the Normal approximation applies to the sample median, even though the sample median is not a function of the sample average.

On the other hand, one need not always assume that the distribution of a statistic is necessarily Normal. In many cases it is not, even for a large sample size. For example, the minimal value of a sample that is generated from the Exponential distribution can be shown to follow the Exponential distribution with an appropriate rate[4], regardless of the sample size.

9.4.6 Simulations

In most problems of statistical inference that will be discussed in this book we will be using the Normal approximation for the sampling distribution of the statistic. However, every now and then we may want to check the validity of this approximation in order to reassure ourselves of its appropriateness. Computerized simulations can be carried out for that checking. The simulations are equivalent to those used in the first part of the book.

A model for the distribution of the observations is assumed each time a simulation is carried out. The simulation itself involves the generation of random samples from that model for the given sample size and for a given value of the parameter. The statistic is evaluated and stored for each generated sample. Thereby, via the generation of many samples, an approximation of the sampling distribution of the statistic is produced. A probabilistic statement inferred from the Normal approximation can be compared to the results of the simulation. Substantial disagreement between the Normal approximation and the outcome of the simulations is an evidence that the Normal approximation may not be valid in the specific setting.

As an illustration, assume the statistic is the average price of a car. It is assumed that the price of a car follows an Exponential distribution with some unknown rate parameter λ. We consider the sampling distribution of the average of 201 Exponential random variables. (Recall that in our sample there are 4 missing values among the 205 observations.) The expectation of the average is $1/\lambda$, which is the expectation of a single Exponential random variable. The variance of a single observation is $1/\lambda^2$. Consequently, the standard deviation of the average is $\sqrt{(1/\lambda^2)/201} = (1/\lambda)/\sqrt{201} = (1/\lambda)/14.17745 = 0.0705/\lambda$.

In the first part of the book we found out that for Normal(μ, σ^2), the Normal distribution with expectation μ and variance σ^2, the central region that contains 95% of the distribution takes the form $\mu \pm 1.96\,\sigma$ (namely, the interval $[\mu - 1.96\,\sigma, \mu + 1.96\,\sigma]$). Thereby, according to the Normal approximation for the sampling distribution of the average price we state that the region $1/\lambda \pm 1.96 \cdot 0.0705/\lambda$ should contain 95% of the distribution.

We may use simulations in order to validate this approximation for selected values of the rate parameter λ. Hence, for example, we may choose $\lambda = 1/12,000$ (which corresponds to an expected price of \$12,000 for a car) and validate the approximation for that parameter value.

[4] If the rate of an Exponential measurement is λ then the rate of the minimum of n such measurements is $n\lambda$.

The simulation itself is carried out by the generation of a sample of size $n = 201$ from the Exponential(1/1200) distribution using the function "`rexp`" for generating Exponential samples[5]. The function for computing the average (`mean`) is applied to each sample and the result stored. We repeat this process a large number of times (100,000 is the typical number we use) in order to produce an approximation of the sampling distribution of the sample average. Finally, we check the relative frequency of cases where the simulated average is within the given range[6]. This relative frequency is an approximation of the required probability and may be compared to the target value of 0.95.

Let us run the proposed simulation for the sample size of $n = 201$ and for a rate parameter equal to $\lambda = 1/12000$. Observe that the expectation of the sample average is equal to $12,000$ and the standard deviation is 0.0705×12000. Hence:

```
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rexp(201,1/12000)
+    X.bar[i] <- mean(X)
+ }
> mean(abs(X.bar-12000) <= 1.96*0.0705*12000)
[1] 0.9496
```

Observe that the simulation produces 0.9496 as the probability of the interval. This result is close enough to the target probability of 0.95, proposing that the Normal approximation is adequate in this example.

The simulation demonstrates the appropriateness of the Normal approximation for the specific value of the parameter that was used. In order to gain more confidence in the approximation we may want to consider other values as well. However, simulations in this book are used only for demonstration. Hence, in most cases where we conduct a simulation experiment, we conduct it only for a single evaluation of the parameters. We leave it to the curiosity of the reader to expand the simulations and try other evaluations of the parameters.

Simulations may also be used in order to compute probabilities in cases where the Normal approximation does not hold. As an illustration, consider the mid-range statistic. This statistic is computed as the average between the largest and the smallest values in the sample. This statistic is discussed in the next chapter.

Consider the case where we obtain 100 observations. Let the distribution of each observation be Uniform. Suppose we are interested as before in the central range that contains 95% of the distribution of the mid-range statistic. The Normal approximation does not apply in this case. Yet, if we specify the parameters of the Uniform distribution then we may use simulations in order to compute the range.

As a specific example let the distribution of an observation be Uniform$(3, 7)$. In the simulation we generate a sample of size $n = 100$ from this distribution[7]

[5] The expression for generating a sample is "`rexp(201,1/12000)`"

[6] In the case where the simulated averages are stored in the sequence "`X.bar`" then we may use the expression "`mean(abs(X.bar - 12000) <= 1.96*0.0705*12000)`" in order to compute the relative frequency.

[7] With the expression "`runif(100,3,7)`".

and compute the mid-range for the sample.

For the computation of the statistic we need to obtain the minimal and the maximal values of the sample. The minimal value of a sequence is compute with the function "`min`". The input to this function is a sequence and the output is the minimal value of the sequence. Similarly, the maximal value is computed with the function "`max`". Again, the input to the function is a sequence and the output is the maximal value in the sequence. The statistic itself is obtained by adding the two extreme values to each other and dividing the sum by two[8].

We produce, just as before, a large number of samples and compute the value of the statistic to each sample. The distribution of the simulated values of the statistic serves as an approximation of the sampling distribution of the statistic. The central range that contains 95% of the sampling distribution may be approximated with the aid of this simulated distribution.

Specifically, we approximate the central range by the identification of the 0.025-percentile and the 0.975-percentile of the simulated distribution. Between these two values are 95% of the simulated values of the statistic. The percentiles of a sequence of simulated values of the statistic can be identified with the aid of the function "`quantile`" that was presented in the first part of the book. The first argument to the function is a sequence of values and the second argument is a number p between 0 and 1. The output of the function is the p-percentile of the sequence[9]. The p-percentile of the simulated sequence serves as an approximation of the p-percentile of the sampling distribution of the statistic.

The second argument to the function "`quantile`" may be a sequence of values between 0 and 1. If so, the percentile for each value in the second argument is computed[10].

Let us carry out the simulation that produces an approximation of the central region that contains 95% of the sampling distribution of the mid-range statistic for the Uniform distribution:

```
> mid.range <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- runif(100,3,7)
+    mid.range[i] <- (max(X)+min(X))/2
+ }
> quantile(mid.range,c(0.025,0.975))
    2.5%    97.5%
4.941680 5.059004
```

Observe that (approximately) 95% of the sampling distribution of the statistic are in the range $[4.941680, 5.059004]$.

Simulations can be used in order to compute the expectation, the standard deviation or any other numerical summary of the sampling distribution of a

[8] If the sample is stored in an object by the name "`X`" then one may compute the mid-range statistic with the expression "`(max(X)+min(X))/2`".

[9] The p-percentile of a sequence is a number with the property that the proportion of entries with values smaller than that number is p and the proportion of entries with values larger than the number is $1 - p$.

[10] If the simulated values of the statistic are stored in a sequence by the name "`mid.range`" then the 0.025-percentile and the 0.975-percentile of the sequence can be computed with the expression "`quantile(mid.range,c(0.025,0.975))`".

statistic. All one needs to do is compute the required summary for the simulated sequence of statistic values and hence obtain an approximation of the required summary. For example, we my use the sequence "`mid.range`" in order to obtain the expectation and the standard deviation of the mid-range statistic of a sample of 100 observations from the Uniform$(3, 7)$ distribution:

```
> mean(mid.range)
[1] 5.000168
> sd(mid.range)
[1] 0.02767719
```

The expectation of the statistic is obtained by the application of the function "`mean`" to the sequence. Observe that it is practically equal to 5. The standard deviation is obtained by the application of the function "`sd`". Its value is approximately equal to 0.028.

9.5 Solved Exercises

Magnetic fields have been shown to have an effect on living tissue and were proposed as a method for treating pain. In the case study presented here, Carlos Vallbona and his colleagues[11] sought to answer the question "Can the chronic pain experienced by postpolio patients be relieved by magnetic fields applied directly over an identified pain trigger point?"

A total of 50 patients experiencing post-polio pain syndrome were recruited. Some of the patients were treated with an active magnetic device and the others were treated with an inactive placebo device. All patients rated their pain before (`score1`) and after application of the device (`score2`). The variable "`change`" is the difference between "`score1`" and "`score2`. The treatment condition is indicated by the variable "`active`." The value "1" indicates subjects receiving treatment with the active magnet and the value "2" indicates subjects treated with the inactive placebo.

This case study is taken from the Rice Virtual Lab in Statistics. More details on this case study can be found in the case study Magnets and Pain Relief that is presented in that site.

Question 9.1. The data for the 50 patients is stored in file "`magnets.csv`". The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/magnets.csv`. Download this file to your computer and store it in the working directory of `R`. Read the content of the file into an `R` data frame. Produce a summary of the content of the data frame and answer the following questions:

1. What is the sample average of the change in score between the patient's rating before the application of the device and the rating after the application?

2. Is the variable "`active`" a factor or a numeric variable?

[11] Vallbona, Carlos, Carlton F. Hazlewood, and Gabor Jurida. (1997). Response of pain to static magnetic fields in postpolio patients: A double-blind pilot study. Archives of Physical and Rehabilitation Medicine 78(11): 1200-1203.

3. Compute the average value of the variable "`change`" for the patients that received and active magnet and average value for those that received an inactive placebo. (Hint: Notice that the first 29 patients received an active magnet and the last 21 patients received an inactive placebo. The subsequence of the first 29 values of the given variables can be obtained via the expression "`change[1:29]`" and the last 21 vales are obtained via the expression "`change[30:50]`".)

4. Compute the sample standard deviation of the variable "`change`" for the patients that received and active magnet and the sample standard deviation for those that received an inactive placebo.

5. Produce a boxplot of the variable "`change`" for the patients that received and active magnet and for patients that received an inactive placebo. What is the number of outliers in each subsequence?

Solution (to Question 9.1.1): Let us read the data into a data frame by the name "`magnets`" and apply the function "`summary`" to the data frame:

```
> magnets <- read.csv("magnets.csv")
> summary(magnets)
     score1          score2          change     active
 Min.   : 7.00   Min.   : 0.00   Min.   : 0.0   "1":29
 1st Qu.: 9.25   1st Qu.: 4.00   1st Qu.: 0.0   "2":21
 Median :10.00   Median : 6.00   Median : 3.5
 Mean   : 9.58   Mean   : 6.08   Mean   : 3.5
 3rd Qu.:10.00   3rd Qu.: 9.75   3rd Qu.: 6.0
 Max.   :10.00   Max.   :10.00   Max.   :10.0
```

The variable "`change`" contains the difference between the patient's rating before the application of the device and the rating after the application. The sample average of this variable is reported as the "`Mean`" for this variable and is equal to 3.5.

Solution (to Question 9.1.2): The variable "`active`" is a factor. Observe that the summary of this variable lists the two levels of the variable and the frequency of each level. Indeed, the levels are coded with numbers but, nonetheless, the variable is a factor[12].

Solution (to Question 9.1.3): Based on the hint we know that the expressions "`change[1:29]`" and "`change[30:50]`" produce the values of the variable "`change`" for the patients that were treated with active magnets and by inactive placebo, respectively. We apply the function "`mean`" to these sub-sequences:

```
> mean(magnets$change[1:29])
[1] 5.241379
> mean(magnets$change[30:50])
[1] 1.095238
```

[12]The number codes are read as character strings into `R`. Notice that the codes are given in the data file "`magnets.csv`" between double quotes.

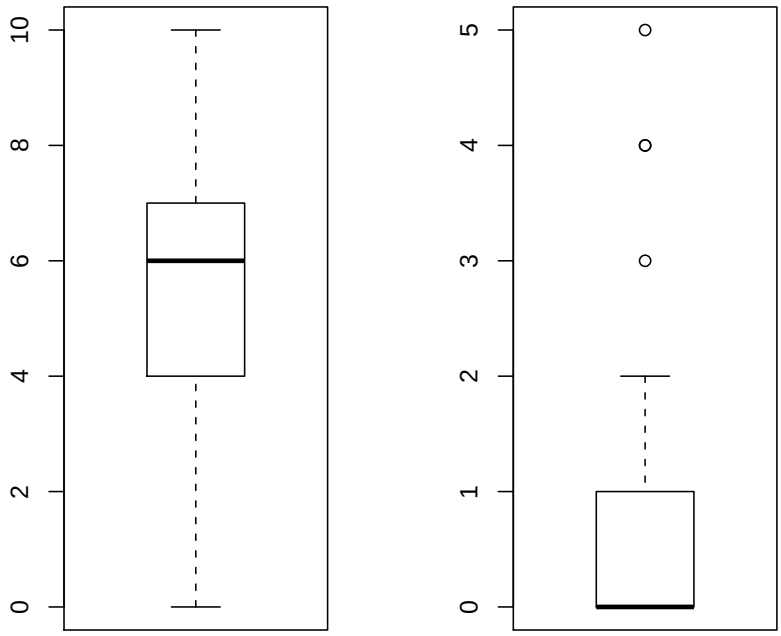

Figure 9.1: Two Box-plots

The sample average for the patients that were treated with active magnets is 5.241379 and sample average for the patients that were treated with inactive placebo is 1.095238.

Solution (to Question 9.1.4): We apply the function "`sd`" to these sub-sequences:

```
> sd(magnets$change[1:29])
[1] 3.236568
> sd(magnets$change[30:50])
[1] 1.578124
```

The sample standard deviation for the patients that were treated with active magnets is 3.236568 and sample standard deviation for the patients that were treated with inactive placebo is 1.578124.

Solution (to Question 9.1.5): We apply the function "`boxplot`" to each sub-sequences:

```
> boxplot(magnets$change[1:29])
```

```
> boxplot(magnets$change[30:50])
```

The box-plots are presented in Figure 9.1. The box-plot on the left correspond to the sub-sequence of the patients that received an active magnet. There are no outliers in this plot. The box-plot on the right correspond to the sub-sequence of the patients that received aa inactive placebo. Three values, the values "3", "4", and "5" are associated with outliers. Let us see what is the total number of observations that receive these values:

```
> table(magnets$change[30:50])

 0  1  2  3  4  5
11  5  1  1  2  1
```

One may see that a single observation obtained the value "3", another one obtained the value "5" and 2 observations obtained the value "4", a total of 4 outliers[13]. Notice that the single point that is associated with the value "4" actually represents 2 observations and not one.

Question 9.2. In Chapter 13 we will present a statistical test for testing if there is a difference between the patients that received the active magnets and the patients that received the inactive placebo in terms of the *expected* value of the variable that measures the change. The test statist for this problem is taken to be

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/29 + S_2^2/21}},$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample averages for the 29 patients that receive active magnets and for the 21 patients that receive inactive placebo, respectively. The quantities S_1^2 and S_2^2 are the sample variances for each of the two samples. Our goal is to investigate the sampling distribution of this statistic in a case where both expectations are equal to each other and to compare this distribution to the observed value of the statistic.

1. Assume that the expectation of the measurement is equal to 3.5, regardless of what the type of treatment that the patient received. We take the standard deviation of the measurement for patients the receives an active magnet to be equal to 3 and for those that received the inactive placebo we take it to be equal to 1.5. Assume that the distribution of the measurements is Normal and there are 29 patients in the first group and 21 in the second. Find the interval that contains 95% of the sampling distribution of the statistic.

2. Does the observed value of the statistic, computed for the data frame "`magnets`", falls inside or outside of the interval that is computed in 1?

Solution (to Question 9.2.1): Let us run the following simulation:

```
> mu1 <- 3.5
> sig1 <- 3
```

[13] An alternative method for obtaining the total count of the observations with values larger or equal to "3" is to run the expression "`sum(magnets$change[30:50] >= 3)`".

```
> mu2 <- 3.5
> sig2 <- 1.5
> test.stat <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X1 <- rnorm(29,mu1,sig1)
+    X2 <- rnorm(21,mu2,sig2)
+    X1.bar <- mean(X1)
+    X2.bar <- mean(X2)
+    X1.var <- var(X1)
+    X2.var <- var(X2)
+    test.stat[i] <- (X1.bar-X2.bar)/sqrt(X1.var/29 + X2.var/21)
+ }
> quantile(test.stat,c(0.025,0.975))
     2.5%     97.5%
-2.014838  2.018435
```

Observe that each iteration of the simulation involves the generation of two samples. One sample is of size 29 and it is generated from the Normal$(3.5, 3^2)$ distribution and the other sample is of size 21 and it is generated from the Normal$(3.5, 1.5^2)$ distribution. The sample average and the sample variance are computed for each sample. The test statistic is computed based on these averages and variances and it is stored in the appropriate position of the sequence "`test.stat`".

The values of the sequence "`test.stat`" at the end of all the iterations represent the sampling distribution of the static. The application of the function "`quantile`" to the sequence gives the 0.025-percentiles and the 0.975-percentiles of the sampling distribution, which are -2.014838 and 2.018435. It follows that the interval $[-2.014838, 2.018435]$ contains about 95% of the sampling distribution of the statistic.

Solution (to Question 9.2.2): In order to evaluate the statistic for the given data set we apply the same steps that were used in the simulation for the computation of the statistic:

```
> x1.bar <- mean(magnets$change[1:29])
> x2.bar <- mean(magnets$change[30:50])
> x1.var <- var(magnets$change[1:29])
> x2.var <- var(magnets$change[30:50])
> (x1.bar-x2.bar)/sqrt(x1.var/29 + x2.var/21)
[1] 5.985601
```

In the first line we compute the sample average for the first 29 patients and in the second line we compute it for the last 21 patients. In the third and fourth lines we do the same for the sample variances of the two types of patients. Finally, in the fifth line we evaluate the statistic. The computed value of the statistic turns out to be 5.985601, a value that does not belong to the interval $[-2.014838, 2.018435]$.

9.6 Summary

Glossary

Statistical Inferential: Methods for gaining insight regarding the population parameters from the observed data.

Point Estimation: An attempt to obtain the best guess of the value of a population parameter. An estimator is a statistic that produces such a guess. The estimate is the observed value of the estimator.

Confidence Interval: An interval that is most likely to contain the population parameter. The confidence level of the interval is the sampling probability that the confidence interval contains the parameter value.

Hypothesis Testing: A method for determining between two hypothesis, with one of the two being the currently accepted hypothesis. A determination is based on the value of the test statistic. The probability of falsely rejecting the currently accepted hypothesis is the significance level of the test.

Comparing Samples: Samples emerge from different populations or under different experimental conditions. Statistical inference may be used to compare the distributions of the samples to each other.

Regression: Relates different variables that are measured on the same sample. Regression models are used to describe the effect of one of the variables on the distribution of the other one. The former is called the explanatory variable and the later is called the response.

Missing Value: An observation for which the value of the measurement is not recorded. `R` uses the symbol "`NA`" to identify a missing value.

Discuss in the forum

A data set may contain missing values. Missing value is an observation of a variable for which the value is not recorded. Most statistical procedures delete observations with missing values and conduct the inference on the remaining observations.

Some people say that the method of deleting observations with missing values is dangerous and may lead to biased analysis. The reason is that missing values may contain information. What is your opinion?

When you formulate your answer to this question it may be useful to come up with an example from you own field of interest. Think of an example in which a missing value contains information relevant for inference or an example in which it does not. In the former case try to assess the possible effects on the analysis that may emerge due to the deletion of observations with missing values.

For example, the goal in some clinical trials is to assess the effect of a new treatment on the survival of patients with a life-threatening illness. The trial is conducted for a given duration, say two years, and the time of death of the patients is recorded. The time of death is missing for patients that survived the entire duration of the trial. Yet, one is advised not to ignore these patients in the analysis of the outcome of the trial.

Chapter 10

Point Estimation

10.1 Student Learning Objectives

The subject of this chapter is the estimation of the value of a parameter on the basis of data. An estimator is a statistic that is used for estimation. Criteria for selecting among estimators are discussed, with the goal of seeking an estimator that tends to obtain values that are as close as possible to the value of the parameter. Different examples of estimation problems are presented and analyzed. By the end of this chapter, the student should be able to:

- Recognize issues associated with the estimation of parameters.
- Define the notions of bias, variance and mean squared error (MSE) of an estimator.
- Estimate parameters from data and assess the performance of the estimation procedure.

10.2 Estimating Parameters

Statistic is the science of data analysis. The primary goal in statistic is to draw meaningful and solid conclusions on a given phenomena on the basis of observed data. Typically, the data emerges as a sample of observations. An observation is the outcome of a measurement (or several measurements) that is taken for a subject that belongs to the sample. These observations may be used in order to investigate the phenomena of interest. The conclusions are drawn from the analysis of the observations.

A key aspect in statistical inference is the association of a probabilistic model to the observations. The basic assumption is that the observed data emerges from some distribution. Usually, the assumption is that the distribution is linked to a theoretical model, such as the Normal, Exponential, Poisson, or any other model that fits the specifications of the measurement taken.

A standard setting in statistical inference is the presence of a sequence of observations. It is presumed that all the observations emerged from a common distribution. The parameters one seeks to estimate are summaries or characteristics of that distribution.

For example, one may be interested in the distribution of price of cars. A reasonable assumption is that the distribution of the prices is the Exponential(λ) distribution. Given an observed sample of prices one may be able to estimate the rate λ that specifies the distribution.

The target in statistical point estimation of a parameter is to produce the best possible guess of the value of a parameter on the basis of the available data. The statistic that tries to guess the value of the parameter is called an *estimator*. The estimator is a formula applied to the data that produces a number. This number is the *estimate* of the value of the parameter.

An important characteristic of a distribution, which is always of interest, is the expectation of the measurement, namely the central location of the distribution. A natural estimator of the expectation is the sample average. However, one may propose other estimators that make sense, such as the sample mid-range that was presented in the previous chapter. The main topic of this chapter is the identification of criteria that may help us choose which estimator to use for the estimation of which parameter.

In the next section we discuss issues associated with the estimation of the expectation of a measurement. The following section deals with the estimation of the variance and standard deviation – summaries that characterize the spread of the distribution. The last section deals with the theoretical models of distribution that were introduced in the first part of the book. It discusses ways by which one may estimate the parameters that characterize these distributions.

10.3 Estimation of the Expectation

A natural candidate for the estimation of the expectation of a random variable on the basis of a sample of observations is the sample average. Consider, as an example, the estimation of the expected price of a car using the information in the data file "`cars.csv`". Let us read the data into a data frame named "`cars`" and compute the average of the variable "`price`":

```
> cars <- read.csv("cars.csv")
> mean(cars$price)
[1] NA
```

The application of the function "`mean`" for the computation of the sample average produced a missing value. The reason is that the variable "`price`" contains 4 missing values. As default, when applied to a sequence that contains missing values, the function "`mean`" produce as output a missing value.

The behavior of the function "`mean`" at the presence of missing values is determined by the argument "`na.rm`"[1]. If we want to compute the average of the non-missing values in the sequence we should specify the argument "`na.rm`" as "`TRUE`". This can be achieved by the inclusion of the expression "`na.rm=TRUE`" in the arguments of the function:

[1] The name of the argument stands for "NA remove". If the value of the argument is set to "`TRUE`" then the missing values are removed in the computation of the average. Consequently, the average is computed for the sub-sequence of non-missing values. The default specification of the argument in the definition of the function is "`na.rm=FALSE`", which implies a missing value for the mean when computed on a sequence that contains missing values.

```
> mean(cars$price,na.rm=TRUE)
[1] 13207.13
```

The resulting average price is, approximately, \$13,000.

10.3.1 The Accuracy of the Sample Average

How close is the estimated value of the expectation – the average price – to the expected price?

There is no way of answering this question on the basis of the data we observed. Indeed, we think of the price of a random car as a random variable. The expectation we seek to estimate is the expectation of that random variable. However, the actual value of that expectation is unknown. Hence, not knowing what is the target value, how can we determine the distance between the computed average 13207.13 and that unknown value?

As a remedy for not being able to answer the question we would like to address we, instead, change the question. In the new formulation of the question we ignore the data at hand altogether. The new formulation considers the sample average as a statistic and the question is formulated in terms of the sampling distribution of that statistic. The question, in its new formulation is: How close is the sample average of the price, taken as a random variable, to the expected price?

Notice that in the new formulation of the question the observed average price $\bar{x} = 13207.13$ has no special role. The question is formulated in terms of the sampling distribution of the sample average ($\bar{X}$). The observed average value is only one among many in the sampling distribution of the average.

The advantage of the new formulation of the question is that it can be addressed. We do have means for investigating the closeness of the estimator to the parameter and thereby producing meaningful answers. Specifically, consider the current case where the estimator is the sample average $\bar{X}$. This estimator attempts to estimate the expectation $\mathrm{E}(X)$ of the measurement, which is the parameter. Assessing the closeness of the estimator to the parameter corresponds to the comparison between the distribution of the random variable, i.e. the estimator, and the value of the parameter.

For this comparison we may note that the expectation $\mathrm{E}(X)$ is also the expectation of the sample average $\bar{X}$. Consequently, in this problem the assessment of the closeness of the estimator to the parameter is equivalent to the investigation of the spread of the distribution of the sample average about its expectation.

Consider an example of such investigation. Imagine that the expected price of a car is equal to \$13,000. A question one may ask is how likely it is that the estimator's guess at the value is within \$1,000 of the actual value? In other words, what is the probability that sample average falls in the range $[12,000, 14,000]$?

Let us investigate this question using simulations. Recall our assumption that the distribution of the price is Exponential. An expectation of 13,000 corresponds to a rate parameter of $\lambda = 1/13,000$. We simulate the sampling distribution of the estimator by the generation of a sample of 201 Exponential random variables with this rate. The sample average is computed for each sample and stored. The sampling distribution of the sample average is approximated

via the production of a large number of sample averages:

```
> lam <- 1/13000
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rexp(201,lam)
+    X.bar[i] <- mean(X)
+ }
> mean(abs(X.bar - 1/lam) <= 1000)
[1] 0.7247
```

In the last line of the code we compute the probability of being within \$1,000 of the expected price. Recall that the expected price in the Exponential case is the reciprocal of the rate λ. In this simulation we obtained 0.7247 as an approximation of the probability.

In the case of the sample average we may also apply the Normal approximation in order to assess the probability under consideration. In particular, if $\lambda = 1/13,000$ then the expectation of an Exponential observation is $\mathrm{E}(X) = 1/\lambda = 13,000$ and the variance is $\mathrm{Var}(X) = 1/\lambda^2 = (13,000)^2$. The expectation of the sample average is equal to the expectation of the measurement, 13,000 in this example. The variance of the sample average is equal to the variance of the observation, divided by the sample size. In the current setting it is equal to $(13,000)^2/201$. The standard deviation is equal to the square root of the variance.

The Normal approximation uses the Normal distribution in order to compute probabilities associated with the sample average. The Normal distribution that is used has the same expectation and standard deviation as the sample average:

```
> mu <- 13000
> sig <- 13000/sqrt(201)
> pnorm(14000,mu,sig) - pnorm(12000,mu,sig)
[1] 0.7245391
```

The probability of falling within the interval $[12000, 14000]$ is computed as the difference between the cumulative Normal probability at 14,000 and the cumulative Normal probability at 12,000.

These cumulative probabilities are computed with the function "`pnorm`". Recall that this function computes the cumulative probability for the Normal distribution. If the first argument is 14,000 then the function produces the probability that a Normal random variable is less than or equal to 14,000. Likewise, if the first argument is 12,000 then the computed probability is the probability of being less than or equal to 12,000. The expectation of the Normal distribution enters in the second argument of the function and the standard deviation enters in the third argument.

The Normal approximation of falling in the interval $[12000, 14000]$, computed as the difference between the two cumulative probabilities, produces 0.7245391 as the probability[2]. Notice that the probability 0.7247 computed in the simulations is in agreement with the Normal approximation.

[2] As a matter of fact, the difference is the probability of falling in the half-open interval $(12000, 14000]$. However, for continuous distributions the probability of the end-points is zero and they do not contribute to the probability of the interval.

If we wish to assess the accuracy of the estimator at other values of the parameter, say $\mathrm{E}(X) = 12,000$ (which corresponds to $\lambda = 1/12,000$) or $\mathrm{E}(X) = 14,033$, (which corresponds to $\lambda = 1/14,033$) all we need to do is change the expression "`lam <- 1/13000`" to the new value and rerun the simulation.

Alternatively, we may use a Normal approximation with modified interval, expectation, and standard deviation. For example, consider the case where the expected price is equal to \$12,000. In order to asses the probability that the sample average falls within \$1,000 of the parameter we should compute the probability of the interval $[11,000, 13,000]$ and change the entries to the first argument of the function "`pnorm`" accordingly. The new expectation is 12,000 and the new standard deviation is $12,000/\sqrt{201}$:

```
> mu <- 12000
> sig <- 12000/sqrt(201)
> pnorm(13000,mu,sig) - pnorm(11000,mu,sig)
[1] 0.7625775
```

This time we get that the probability is, approximately, 0.763.

The fact that the computed value of the average 13,207.13 belongs to the interval $[12,000, 14,000]$ that was considered in the first analysis but does not belong to the interval $[11,000, 13,000]$ that was considered in the second analysis is irrelevant to the conclusions drawn from the analysis. In both cases the theoretical properties of the sample average as an estimator were considered and not its value at specific data.

In the simulation and in the Normal approximation we applied one method for assessing the closeness of the sample average to the expectation it estimates. This method involved the computation of the probability of being within \$1,000 of the expected price. The higher this probability, the more accurate is the estimator.

An alternative method for assessing the accuracy of an estimator of the expectation may involve the use of an overall summary of the spread of the distribution. A standard method for quantifying the spread of a distribution about the expectation is the variance (or its square root, the standard deviation). Given an estimator of the expectation of a measurement, the sample average for example, we may evaluate the accuracy of the estimator by considering its variance. The smaller the variance the more accurate is the estimator.

Consider again the case where the sample average is used in order to estimate the expectation of a measurement. In such a situation the variance of the estimator, i.e. the variance of the sample average, is obtained as the ratio between the variance of the measurement $\mathrm{Var}(X)$, divided by the sample size n:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n \ .$$

Notice that for larger sample sizes the estimator is more accurate. The lager the sample size n the smaller is the variance of the estimator, in which case the values of the estimator tend to be more concentrated about the expectation. Hence, one may make the estimator more accurate by increasing the sample size.

Another method for improving the accuracy of the average of measurements in estimating the expectation is the application of a more accurate measurement

device. If the variance $\mathrm{Var}(X)$ of the measurement device decreases so does the variance of the sample average of such measurements.

In the sequel, when we investigate the accuracy of estimators, we will generally use overall summaries of the spread of their distribution around the target value of the parameter.

10.3.2 Comparing Estimators

Notice that the formulation of the accuracy of estimation that we use replaces the question: "How close is the given value of the estimator to the unknown value of the parameter?" by the question: "How close are the unknown (and random) values of the estimator to a given value of the parameter?" In the second formulation the question is completely academic and unrelated to actual measurement values. In this academic context we can consider different potential values of the parameter. Once the value of the parameter has been selected it can be treated as known in the context of the academic discussion. Clearly, this does not imply that we actually know what is the true value of the parameter.

The sample average is a natural estimator of the expectation of the measurement. However, one may propose other estimators. For example, when the distribution of the measurement is symmetric about the expectation then the median of the distribution is equal to the expectation. The sample median, which is a natural estimator of the measurement median, is an alternative estimator of the expectation in such case. Which of the two alternatives, the sample average or the sample median, should we prefer as an estimator of the expectation in the case of a symmetric distribution?

The straightforward answer to this question is to prefer the better one, the one which is more accurate. As part of the solved exercises you are asked to compare the sample average to the sample median as estimators of the expectation. Here we compare the sample average to yet another alternative estimator – the mid-range estimator – which is the average between the smallest and the largest observations.

In the comparison between estimators we do not evaluate them in the context of the observed data. Rather, we compare them as random variables. The comparison deals with the properties of the estimators in a given theoretical context. This theoretical context is motivated by the realities of the situation as we know them. But, still, the frame of reference is the theoretical model and not the collected data.

Hence, depending on the context, we may assume in the comparison that the observations emerge from some distribution. We may specify parameter values for this distribution and select the appropriate sample size. After setting the stage we can compare the accuracy of one estimator against that of the other. Assessment at other parameter values in the context of the given theoretical model, or of other theoretical models, may provide insight and enhance our understanding regarding the relative merits and weaknesses of each estimator.

Let us compare the sample average to the sample mid-range as estimators of the expectation in a situation that we design. Consider a Normal measurement X with expectation $\mathrm{E}(X) = 3$ and variance that is equal to 2. Assume that the sample size is $n = 100$. Both estimators, due to the symmetry of the Normal distribution, are centered at the expectation. Hence, we may evaluate

the accuracy of the two estimators using their variances. These variances are the measure of the spread of the distributions of each estimator about the target parameter value.

We produce the sampling distribution and compute the variances using a simulation. Recall that the distribution of the mid-range statistic was simulated in the previous chapter. In the computation of the mid-range statistic we used the function "`max`" that computes the maximum value of its input and the function "`min`" that computes the minimum value:

```
> mu <- 3
> sig <- sqrt(2)
> X.bar <- rep(0,10^5)
> mid.range <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rnorm(100,mu,sig)
+    X.bar[i] <- mean(X)
+    mid.range[i] <- (max(X)+min(X))/2
+ }
> var(X.bar)
[1] 0.02020161
> var(mid.range)
[1] 0.1850595
```

We get that the variance of the sample average[3] is approximately equal to 0.02. The variance of the mid-range statistic is approximately equal to 0.185, more than 9 times as large. We see that the accuracy of the sample average is better in this case than the accuracy of the mid-range estimator. Evaluating the two estimators at other values of the parameter will produce the same relation. Hence, in the current example it seems as if the sample average is the better of the two.

Is the sample average necessarily the best estimator for the expectation? The next example will demonstrate that this need not always be the case.

Consider again a situation of observing a sample of size $n = 100$. However, this time the measurement X is Uniform and not Normal. Say $X \sim \text{Uniform}(0.5, 5.5)$ has the Uniform distribution over the interval $[0.5, 5.5]$. The expectation of the measurement is equal to 3 like before, since $\text{E}(X) = (0.5 + 5.5)/2 = 3$. The variance on an observation is $\text{Var}(X) = (5.5 - 0.5)^2/12 = 2.083333$, not much different from the variance that was used in the Normal case. The Uniform distribution, like the Normal distribution, is a symmetric distribution about the center of the distribution. Hence, using the mid-range statistic as an estimator of the expectation makes sense[4].

We re-run the simulations, using the function "`runif`" for the simulation of a sample from the Uniform distribution and the parameters of the Uniform distribution instead of the function "`rnorm`" that was used before:

[3] As a matter of fact, the variance of the sample average is exactly $\text{Var}(X)/100 = 0.02$. Due to the inaccuracy of the simulation we got a slightly different variance.

[4] Observe that the middle range of the Uniform(a, b) distribution, the middle point between the maximum value of the distribution b and the minimal value a, is $(a + b)/2$, which is equal to the expectation of the distribution

```
> a <- 0.5
> b <- 5.5
> X.bar <- rep(0,10^5)
> mid.range <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- runif(100,a,b)
+    X.bar[i] <- mean(X)
+    mid.range[i] <- (max(X)+min(X))/2
+ }
> var(X.bar)
[1] 0.02074304
> var(mid.range)
[1] 0.001209732
```

Again, we get that the variance of the sample average is approximately equal to 0.02, which is close to the theoretical value[5]. The variance of mid-range statistic is approximately equal to 0.0012.

Observe that in the current comparison between the sample average and the mid-range estimator we get that the latter is a clear winner. Examination of other values of a and b for the Uniform distribution will produce the same relation between the two competitors. Hence, we may conclude that for the case of the Uniform distribution the sample average is an inferior estimator.

The last example may serve as yet another reminder that life is never simple. A method that is good in one situation may not be as good in a different situation.

Still, the estimator of choice of the expectation is the sample average. Indeed, in some cases we may find that other methods may produce more accurate estimates. However, in most settings the sample average beats its competitors. The sample average also possesses other useful benefits. Its sampling distribution is always centered at the expectation it is trying to estimate. Its variance has a simple form, i.e. it is equal to the variance of the measurement divided by the sample size. Moreover, its sampling distribution can be approximated by the Normal distribution. Henceforth, due to these properties, we will use the sample average whenever estimation of the expectation is required.

10.4 Estimation of the Variance and Standard Deviation

The spread of the measurement about its expected value may be measured by the variance or by the standard deviation, which is the square root of the variance. The standard estimator for the variance of the measurement is the sample variance and the square root of the sample variance is the default estimator of the standard deviation.

The computation of the sample variance from the data is discussed in Chap-

[5]Actually, the exact value of the variance of the sample average is $\mathrm{Var}(X)/100 = 0.02083333$. The results of the simulation are consistent with this theoretical computation.

ter 3. Recall that the sample variance is computed via the formula:

$$s^2 = \frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample} - 1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} ,$$

where $\bar{x}$ is the sample average and n is the sample size. The term $x_i - \bar{x}$ is the deviation from the sample average of the ith observation and $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sum of the squares of deviations. It is pointed out in Chapter 3 that the reason for dividing the sum of squares by $(n-1)$, rather than n, stems from considerations of statistical inference. A promise was made that these reasonings will be discussed in due course. Now we want to deliver on this promise.

Let us compare between two competing estimators for the variance, both considered as random variables. One is the estimator S^2, which is equal to the formula for the sample variance applied to a random sample:

$$S^2 = \frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample} - 1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} ,$$

The computation of this statistic can be carried out with the function "`var`".

The second estimator is the one obtained when the sum of squares is divided by the sample size (instead of the sample size minus 1):

$$\frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} .$$

Observe that the second estimator can be represented in the form:

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \frac{n-1}{n} \cdot \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = [(n-1)/n]S^2 .$$

Hence, the second estimator may be obtained by the multiplication of the first estimator S^2 by the ratio $(n-1)/n$. We seek to compare between S^2 and $[(n-1)/n]S^2$ as estimators of the variance.

In order to make the comparison concrete, let us consider it in the context of a Normal measurement with expectation $\mu = 5$ and variance $\sigma^2 = 3$. Let us assume that the sample is of size 20 ($n = 20$).

Under these conditions we carry out a simulation. Each iteration of the simulation involves the generation of a sample of size $n = 20$ from the given Normal distribution. The sample variance S^2 is computed from the sample with the application of the function "`var`". The resulting estimate of the variance is stored in an object that is called "`X.var`":

```
> mu <- 5
> std <- sqrt(3)
> X.var <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rnorm(20,mu,std)
+    X.var[i] <- var(X)
+ }
```

The content of the object "`X.var`", at the end of the simulation, approximates the sampling distribution of the estimator S^2.

Our goal is to compare between the performance of the estimator of the variance S^2 and that of the alternative estimator. In this alternative estimator the sum of squared deviations is divided by the sample size ($n = 20$) and not by the sample size minus 1 ($n-1 = 19$). Consequently, the alternative estimator is obtained by multiplying S^2 by the ratio 19/20. The sampling distribution of the values of S^2 is approximated by the content of the object "`X.var`". It follows that the sampling distribution of the alternative estimator is approximated by the object "`(19/20)*X.var`", in which each value of S^2 is multiplied by the appropriate ratio. The comparison between the sampling distribution of S^2 and the sampling distribution of the alternative estimator is obtained by comparing between "`X.var`" and "`(19/20)*X.var`".

Let us start by the investigation of the expectation of the estimators. Recall that when we analyzed the sample average as an estimator of the expectation of a measurement we obtained that the expectation of the sampling distribution of the estimator is equal to the value of the parameter it is trying to estimate. One may wonder: What is the situation for the estimators of the variance? Is it or is it not the case that the expectation of their sampling distribution equals the value of the variance? In other words, is the distribution of either estimators of the variance centered at the value of the parameter they are trying to estimate?

Compute the expectations of the two estimators:

```
> mean(X.var)
[1] 2.995400
> mean((19/20)*X.var)
[1] 2.845630
```

Note that 3 is the value of the variance of the measurement that was used in the simulation. Observe that the expectation of S^2 is essentially equal to 3, whereas the expectation of the alternative estimator is less than 3. Hence, at least in the example that we consider, the center of the distribution of S^2 is located on the target value. On the other hand, the center of the sampling distribution of the alternative estimator is located off that target value.

As a matter of fact it can be shown mathematically that the expectation of the estimator S^2 is always equal to the variance of the measurement. This holds true regardless of what is the actual value of the variance. On the other hand the expectation of the alternative estimator is always off the target value[6].

An estimator is called *unbiased* if its expectation is equal to the value of the parameter that it tries to estimate. We get that S^2 is an unbiased estimator of the variance. Similarly, the sample average is an unbiased estimator of the expectation. Unlike these two estimators, the alternative estimator of the variance is a *biased* estimator.

The default is to use S^2 as the estimator of the variance of the measurement and to use its square root as the estimator of the standard deviation of the measurement. A justification, which is frequently quoted to justify this selection, is the fact that S^2 is an unbiased estimator of the variance[7].

[6] For the estimator S^2 we get that $\mathrm{E}(S^2) = \mathrm{Var}(X)$. On the other hand, for the alternative estimator we get that $\mathrm{E}([(n-1)/n] \cdot S^2) = [(n-1)/n]\mathrm{Var}(X) \neq \mathrm{Var}(X)$. This statement holds true also in the cases where the distribution of the measurement is not Normal.

[7] As part of your homework assignment you are required to investigate the properties of S, the square root of S^2, as an estimator of the standard deviation of the measurement. A conclusion of this investigation is that S is a biased estimator of the standard deviation.

In the previous section, when comparing two competing estimators of the expectation, or main concern was the quantification of the spread of the sampling distribution of either estimator about the target value of the parameter. We used that spread as a measure of the distance between the estimator and the value it tries to estimate. In the setting of the previous section both estimators were unbiased. Consequently, the variance of the estimators, which measures the spread of the distribution about its expectation, could be used in order to quantify the distance between the estimator and the parameter. (Since, for unbiased estimators, the parameter is equal to the expectation of the sampling distribution.)

In the current section one of the estimators (S^2) is unbiased, but the other (the alternative estimator) is not. In order to compare their accuracy in estimation we need to figure out a way to quantify the distance between a biased estimator and the value it tries to estimate.

Towards that end let us recall the definition of the variance. Given a random variable X with an expectation $\mathrm{E}(X)$, we consider the square of the deviations $(X - \mathrm{E}(X))^2$, which measure the (squared) distance between each value of the random variable and the expectation. The variance is defined as the expectation of the squared distance: $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2]$. One may think of the variance as an overall measure of the distance between the random variable and the expectation.

Assume now that the goal is to assess the distance between an estimator and the parameter it tries to estimate. In order to keep the discussion on an abstract level let us use the Greek letter θ (read: theta) to denote this parameter[8]. The estimator is denoted by $\hat{\theta}$ (read: theta hat). It is a statistic, a formula applied to the data. Hence, with respect to the sampling distribution, $\hat{\theta}$ is a random variable[9]. The issue is to measure the distance between the random variable $\hat{\theta}$ and the parameter θ.

Motivated by the method that led to the definition of the variance we consider the deviations between the estimator and the parameter. The square deviations $(\hat{\theta} - \theta)^2$ may be considered in the current context as a measure of the (squared) distance between the estimator and the parameter. When we take the expectation of these square deviations we get an overall measure of the distance between the estimator and the parameter. This overall distance is called the *mean square error* of the estimator and is denoted by MSE:

$$\mathrm{MSE} = \mathrm{E}\big[(\hat{\theta} - \theta)^2\big] \ .$$

The mean square error of an estimator is tightly linked to the bias and the variance of the estimator. The bias of an estimator $\hat{\theta}$ is the difference between

[8]The letter θ is frequently used in the statistical literature to denote a parameter of the distribution. In the previous section we considered $\theta = \mathrm{E}(X)$ and in this section we consider $\theta = \mathrm{Var}(X)$.

[9]Observe that we diverge here slightly from our promise to use capital letters to denote random variables. However, denoting the parameter by θ and denoting the estimator of the parameter by $\hat{\theta}$ is standard in the statistical literature. As a matter of fact, we will use the "hat" notation, where a hat is placed over a Greek letter that represents the parameter, in other places in this book. The letter with the hat on top will represent the estimator and will always be considered as a random variable. For Latin letters we will still use capital letters, with or without a hat, to represent a random variable and small letter to represent evaluation of the random variable for given data.

the expectation of the estimator and the parameter it seeks to estimate:

$$\text{Bias} = \text{E}(\hat{\theta}) - \theta \ .$$

In an unbiased estimator the expectation of the estimator and the estimated parameter coincide, i.e. the bias is equal to zero. For a biased estimator the bias is either negative, as is the case for the alternative estimator of the variance, or else it is positive.

The variance of the estimator, Variance $= \text{Var}(\hat{\theta})$, is a measure of the spread of the sampling distribution of the estimator about its expectation.

The link between the mean square error, the bias, and the variance is described by the formula:

$$\text{MSE} = \text{Variance} + (\text{Bias})^2 \ .$$

Hence, the mean square error of an estimator is the sum of its variance, the (squared) distance between the estimator and its expectation, and the square of the bias, the square of the distance between the expectation and the parameter. The mean square error is influenced both by the spread of the distribution about the expected value (the variance) and by the distance between the expected value and the parameter (the bias). The larger either of them become the larger is the mean square error, namely the distance between the estimator and the parameter.

Let us compare between the mean square error of the estimator S^2 and the mean square error of the alternative estimator $[19/20]S^2$. Recall that we have computed their expectations and found out that the expectation of S^2 is essentially equal to 3, the target value of the variance. The expectation of the alternative estimator turned out to be equal to 2.845630, which is less than the target value[10]. It turns out that the bias of S^2 is zero (or essentially zero in the simulations) and the bias of the alternative estimator is $2.845630 - 3 = -0.15437 \approx -0.15$.

In order to compute the mean square errors of both estimators, let us compute their variances:

```
> var(X.var)
[1] 0.9361832
> var((19/20)*X.var)
[1] 0.8449054
```

Observe that the variance of S^2 is essentially equal to 0.936 and the variance of the alternative estimator is essentially equal to 0.845.

The estimator S^2 is unbiased. Consequently, the mean square error of S^2 is equal to its variance. The bias of the alternative is -0.15. As a result we get that the mean square error of this estimator, which is the sum of the variance and the square of the bias, is essentially equal to

$$0.845 + (-0.15)^2 = 0.845 + 0.0225 = 0.8675 \ .$$

Observe that the mean square error of the estimator S^2, which is equal to 0.936, is larger than the mean square error of the alternative estimator.

[10] It can be shown mathematically that $\text{E}([(n-1)/n]S^2) = [(n-1)/n]\text{E}(S^2)$. Consequently, the actual value of the expectation of the alternative estimator in the current setting is $[19/20] \cdot 3 = 2.85$ and the bias is -0.15. The results of the simulation are consistent with this fact.

Notice that even though the alternative estimator is biased it still has a smaller mean square error than the default estimator S^2. Indeed, it can be prove mathematically that when the measurement has a Normal distribution then the mean square error of the alternative estimator is always smaller than the mean square error of the sample variance S^2.

Still, although the alternative estimator is slightly more accurate than S^2 in the estimation of the variance, the tradition is to use the latter. Obeying this tradition we will henceforth use S^2 whenever estimation of the variance is required. Likewise, we will use S, the square root of the sample variance, to estimate the standard deviation.

In order to understand how is it that the biased estimator produced a smaller mean square error than the unbiased estimator let us consider the two components of the mean square error. The alternative estimator is biased but, on the other hand, it has a smaller variance. Both the bias and the variance contribute to the mean square error of an estimator. The price for reducing the bias in estimation is usually an increase in the variance and vice versa. The consequence of producing an unbiased estimator such as S^2 is an inflated variance. A better estimator is an estimator that balances between the error that results from the bias and the error that results from the variance. Such is the alternative estimator.

We will use S^2 in order to estimate the variance of a measurement. A context in which an estimate of the variance of a measurement is relevant is in the assessment of the variance of the sample mean. Recall that the variance of the sample mean is equal to $\mathrm{Var}(X)/n$, where $\mathrm{Var}(X)$ is the variance of the measurement and n is the size of the sample. In the case where the variance of the measurement is not known one may estimate it from the sample using S^2. It follows that the estimator of the variance of the sample average is S^2/n. Similarly, $S/\sqrt{n}$ can be used as an estimator of the standard deviation of the sample average.

10.5 Estimation of Other Parameters

In the previous two section we considered the estimation of the expectation and the variance of a measurement. The proposed estimators, the sample average for the expectation and the sample variance for the variance, are not tied to any specific model for the distribution of the measurement. They may be applied to data whether or not a theoretical model for the distribution of the measurement is assumed.

In the cases where a theoretical model for the measurement is assumed one may be interested in the estimation of the specific parameters associated with this model. In the first part of the book we introduced the Binomial, the Poisson, the Uniform, the Exponential, and the Normal models for the distribution of measurements. In this section we consider the estimation of the parameters that determine each of these theoretical distributions based on a sample generated from the same distribution. In some cases the estimators coincide with the estimators considered in the previous sections. In other cases the estimators are different.

Start with the Binomial distribution. We will be interested in the special case $X \sim \mathrm{Binomial}(1, p)$. This case involves the outcome of a single trial. The

trial has two possible outcomes, one of them is designated as "success" and the other as "failure". The parameter p is the probability of the success. The Binomial$(1, p)$ distribution is also called *the Bernoulli distribution*. Our concern is the estimation of the parameter p based on a sample of observations from this Bernoulli distribution.

This estimation problem emerges in many settings that involve the assessment of the probability of an event based on a sample of n observations. In each observation the event either occurs or not. A natural estimator of the probability of the event is its relative frequency in the sample. Let us show that this estimator can be represented as an average of a Bernoulli sample and the sample average is used for the estimation of a Bernoulli expectation.

Consider an event, one may code a measurement X, associated with an observation, by 1 if the event occurs and by 0 if it does not. Given a sample of size n, one thereby produces n observations with values 0 or 1. An observation has the value 1 if the event occurs for that observation or, else, the value is 0.

Notice that $\mathrm{E}(X) = 1 \cdot p = p$. Consequently, the probability of the event is equal to the expectation of the Bernoulli measurement[11]. It turns out that the parameter one seeks to estimate is the expectation of a Bernoulli measurement. The estimation is based on a sample of size n of Bernoulli observations.

In Section 10.3 it was proposed to use the sample average as an estimate of the expectation. The sample average is the sum of the observations, divided by the number of observation. In the specific case of a sample of Bernoulli observations, the sum of observation is the sum of zeros and one. The zeros do not contribute to the sum. Hence, the sum is equal to the number of times that 1 occurs, namely the frequency of the occurrences of the event. When we divide by the sample size we get the relative frequency of the occurrences. The conclusion is that the sample average of the Bernoulli observations and the relative frequency of occurrences of the event in the sample are the same. Consequently, the sample relative frequency of the event is also a sample average that estimates the expectation of the Bernoulli measurement.

We seek to estimate p, the probability of the event. The estimator is the relative frequency of the event in the sample. We denote this estimator by $\hat{P}$. This estimator is a sample average of Bernoulli observations that is used in order to estimate the expectation of the Bernoulli distribution. From the discussion in Section 10.3 one may conclude that this estimator is an unbiased estimator of p (namely, $\mathrm{E}(\hat{P}) = p$) and that its variance is equal to:

$$\mathrm{Var}(\hat{P}) = \mathrm{Var}(X)/n = p(1-p)/n \ ,$$

where the variance of the measurement is obtained from the formula for the variance of a Binomial$(1, p)$ distribution[12].

The second example of an integer valued random variable that was considered in the first part of the book is the Poisson(λ) distribution. Recall that λ is the expectation of a Poisson measurement. Hence, one may use the sample average of Poisson observations in order to estimate this parameter.

The first example of a continuous distribution that was discussed in the first part of the book is the Uniform(a, b) distribution. This distribution is param-

[11] The expectation of $X \sim$ Binomial(n, p) is $\mathrm{E}(X) = np$. In the Bernoulli case $n = 1$. Therefore, $\mathrm{E}(X) = 1 \cdot p = p$.

[12] The variance of $X \sim$ Binomial(n, p) is $\mathrm{Var}(X) = np(1-p)$. In the Bernoulli case $n = 1$. Therefore, $\mathrm{Var}(X) = 1 \cdot p(1-p) = p(1-p)$.

eterized by a and b, the end-points of the interval over which the distribution is defined. A natural estimator of a is the smallest value observed and a natural estimator of b is the largest value. One may use the function "`min`" for the computation of the former estimate from the sample and use the function "`max`" for the computation of the later. Both estimators are slightly biased but have a relatively small mean square error.

Next considered the $X \sim \text{Exponential}(\lambda)$ random variable. This distribution was applied in this chapter to model the distribution of the prices of cars. The distribution is characterized by the rate parameter λ. In order to estimate the rate one may notice the relation between it and the expectation of the measurement:

$$\text{E}(X) = 1/\lambda \quad \Longrightarrow \quad \lambda = 1/\text{E}(X) \ .$$

The rate is equal to the reciprocal of the expectation. The expectation can be estimated by the sample average. Hence a natural proposal is to use the reciprocal of the sample average as an estimator of the rate:

$$\hat{\lambda} = 1/\bar{X} \ .$$

The final example that we mention is the $\text{Normal}(\mu, \sigma^2)$ case. The parameter μ is the expectation of the measurement and may be estimated by the sample average $\bar{X}$. The parameter σ^2 is the variance of a measurement, and can be estimated using the sample variance S^2.

10.6 Solved Exercises

Question 10.1. In Subsection 10.3.2 we compare the average against the mid-range as estimators of the expectation of the measurement. The goal of this exercise is to repeat the analysis, but this time compare the average to the median as estimators of the expectation in symmetric distributions.

1. Simulate the sampling distribution of average and the median of a sample of size $n = 100$ from the $\text{Normal}(3, 2)$ distribution. Compute the expectation and the variance of the sample average and of the sample median. Which of the two estimators has a smaller mean square error?

2. Simulate the sampling distribution of average and the median of a sample of size $n = 100$ from the $\text{Uniform}(0.5, 5.5)$ distribution. Compute the expectation and the variance of the sample average and of the sample median. Which of the two estimators has a smaller mean square error?

Solution (to Question 10.1.1): We simulate the sampling distribution of the average and the median in a sample generated from the Normal distribution. In order to do so we copy the code that was used in Subsection 10.3.2, replacing the object "`mid.range`" by the object "`X.med`" and using the function "`median`" in order to compute the sample median instead of the computation of the mid-range statistic:

```
> mu <- 3
> sig <- sqrt(2)
> X.bar <- rep(0,10^5)
> X.med <- rep(0,10^5)
```

```
> for(i in 1:10^5)
+ {
+    X <- rnorm(100,mu,sig)
+    X.bar[i] <- mean(X)
+    X.med[i] <- median(X)
+ }
```

The sequence "`X.bar`" represents the sampling distribution of the sample average and the sequence "`X.med`" represents the sampling distribution of the sample median. We apply the function "`mean`" to these sequences in order to obtain the expectations of the estimators:

```
> mean(X.bar)
[1] 3.000010
> mean(X.med)
[1] 3.000086
```

The expectation of the measurement, the parameter of interest is equal to 3. Observe that expectations of the estimators are essentially equal to the expectation[13]. Consequently, both estimators are unbiased estimators of the expectation of the measurement.

In order to obtain the variances of the estimators we apply the function "`var`" to the sequences that represent their sampling distributions:

```
> var(X.bar)
[1] 0.02013529
> var(X.med)
[1] 0.03120206
```

Observe that the variance of the sample average is essentially equal to 0.020 and the variance of the sample median is essentially equal to 0.0312. The mean square error of an unbiased estimator is equal to its variance. Hence, these numbers represent the mean square errors of the estimators. It follows that the mean square error of the sample average is less than the mean square error of the sample median in the estimation of the expectation of a Normal measurement.

Solution (to Question 10.1.2): We repeat the same steps as before for the Uniform distribution. Notice that we use the parameters $a = 0.5$ and $b = 5.5$ the same way we did in Subsection 10.3.2. These parameters produce an expectation $\mathrm{E}(X) = 3$ and a variance $\mathrm{Var}(X) = 2.083333$:

```
> a <- 0.5
> b <- 5.5
> X.bar <- rep(0,10^5)
> X.med <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- runif(100,a,b)
+    X.bar[i] <- mean(X)
```

[13] It can be proved mathematically that for a symmetric distribution the expectation of the sample average and the expectation of the sample median are both equal to the expectation of the measurement. The Normal distribution is a symmetric distribution.

```
+    X.med[i] <- median(X)
+ }
```

Applying the function "mean" to the sequences that represent the sampling distribution of the estimators we obtain that both estimators are essentially unbiased[14]:

```
> mean(X.bar)
[1] 3.000941
> mean(X.med)
[1] 3.001162
```

Compute the variances:

```
> var(X.bar)
[1] 0.02088268
> var(X.med)
[1] 0.06069215
```

Observe 0.021 is, essentially, the value of the variance of the sample average[15]. The variance of the sample median is essentially equal to 0.061. The variance of each of the estimators is equal to it's mean square error. This is the case since the estimators are unbiased. Consequently, we again obtain that the mean square error of the sample average is less than that of the sample median.

Question 10.2. The goal in this exercise is to assess estimation of a proportion in a population on the basis of the proportion in the sample.

The file "pop2.csv" was introduced in Exercise 7.1 of Chapter 7. This file contains information associated to the blood pressure of an imaginary population of size 100,000. The file can be found on the internet (http://pluto.huji.ac.il/~msby/StatThink/Datasets/pop2.csv). One of the variables in the file is a factor by the name "group" that identifies levels of blood pressure. The levels of this variable are "HIGH", "LOW", and "NORMAL".

The file "ex2.csv" contains a sample of size $n = 150$ taken from the given population. This file can also be found on the internet (http://pluto.huji.ac.il/~msby/StatThink/Datasets/ex2.csv). It contains the same variables as in the file "pop2.csv". The file "ex2.csv" corresponds in this exercise to the observed sample and the file "pop2.csv" corresponds to the unobserved population.

Download both files to your computer and answer the following questions:

1. Compute the proportion in the sample of those with a high level of blood pressure[16].

2. Compute the proportion in the population of those with a high level of blood pressure.

[14]The Uniform distribution is symmetric. Consequently, both estimators are unbiased.

[15]As a matter of fact, the variance is equal to 0.02. The discrepancy results from the fact that simulations serves only as an approximation to the sampling distribution.

[16]Hint: You may use the function summary or you may note that the expression "*variable*==*level*" produces a sequence with logical "TRUE" or "FALSE" entries that identify entries in the sequence "*variable*" that obtain the value "*level*".

3. Simulate the sampling distribution of the sample proportion and compute its expectation.

4. Compute the variance of the sample proportion.

5. It is proposed in Section 10.5 that the variance of the sample proportion is $\mathrm{Var}(\hat{P}) = p(1-p)/n$, where p is the probability of the event (having a high blood pressure in our case) and n is the sample size ($n = 150$ in our case). Examine this proposal in the current setting.

Solution (to Question 10.2.1): Assuming that the file "`ex2.csv`" is saved in the working directory, one may read the content of the file into a data frame and produce a summary of the content of the data frame using the code:

```
> ex2 <- read.csv("ex2.csv")
> summary(ex2)
       id              sex          age              bmi
 Min.   :1024982   FEMALE:74   Min.   :26.00   Min.   :15.12
 1st Qu.:3172783   MALE  :76   1st Qu.:32.00   1st Qu.:22.02
 Median :5200484               Median :35.00   Median :25.16
 Mean   :5463304               Mean   :35.09   Mean   :24.76
 3rd Qu.:7982902               3rd Qu.:37.00   3rd Qu.:27.49
 Max.   :9934175               Max.   :45.00   Max.   :35.24
    systolic       diastolic         group
 Min.   :100.8   Min.   : 51.98   HIGH  : 37
 1st Qu.:118.1   1st Qu.: 75.02   LOW   :  3
 Median :124.3   Median : 83.19   NORMAL:110
 Mean   :125.3   Mean   : 82.44
 3rd Qu.:132.6   3rd Qu.: 88.83
 Max.   :154.5   Max.   :112.71
```

Examine the variable "`group`". Observe that the sample contains 37 subjects with high levels of blood pressure. Dividing 37 by the sample size we get:

```
> 37/150
[1] 0.2466667
```

Consequently, the sample proportion is 0.2466667.

Alternatively, we compute the sample proportion using the code:

```
> mean(ex2$group == "HIGH")
[1] 0.2466667
```

Notice that the expression "`ex2$group == "HIGH"`" produces a sequence of length 150 with logical entries. The entry is equal to "`TRUE`" if the equality holds and it is equal to "`FALSE`" if it dos not[17]. When the function "`mean`" is applied to a sequence with logical entries it produces the relative frequeny of the `TRUE`s in the sequence. This corresponds, in the corrent context, to the sample proportion of the level "`HIGH`" in the variable "`ex2$group`".

[17] Pay attention to the fact that we use "`==`" in order to express equivalence and not "`=`". The later may be used as an assignment operator similar to "`<-`" and in the determination of an argument of a function.

Solution (to Question 10.2.2): Make sure that the file "`pop2.csv`" is saved in the working directory. In order to compute the proportion in the population we read the content of the file into a data frame and compute the relative frequency of the level "`HIGH`" in the variable "`group`":

```
> pop2 <- read.csv("pop2.csv")
> mean(pop2$group == "HIGH")
[1] 0.28126
```

We get that the proporion in the population is $p = 0.28126$.

Solution (to Question 10.2.3): The simulation of the sampling distribution involves a selection of a random sample of size 150 from the population and the computation of the proportion of the level "`HIGH`" in that sample. This procedure is iterated 100,000 times in order to produce an approximation of the distribution:

```
> P.hat <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- sample(pop2$group,150)
+    P.hat[i] <- mean(X == "HIGH")
+ }
> mean(P.hat)
[1] 0.2812307
```

Observe that the sampling distribution is stored in the object "`P.hat`". The function "`sample`" is used in order to sample 150 observation from the sequence "`pop2$group`". The sample is stored in the object "`X`". The expression "`mean(X == "HIGH")`" computes the relative frequency of the level "`HIGH`" in the sequence "`X`".

At the last line, after the production of the sequence "`P.hat`" is completed, the function "`mean`" is applied to the sequence. The result is the expected value of estimator $\hat{P}$, which is equal to 0.2812307. This expectation is essentially equal to the probability of the event $p = 0.28126$.[18]

Solution (to Question 10.2.4): The application of the function "`var`" to the sequence "`P.hat`" produces:

```
> var(P.hat)
[1] 0.001350041
```

Hence, the variance of the estimator is (approximately) equal to 0.00135.

Solution (to Question 10.2.5): Compute the variance according to the formula that is proposed in Section:

```
> p <- mean(pop2$group == "HIGH")
> p*(1-p)/150
[1] 0.001347685
```

[18] It can be shown mathematically that for random sampling from a population we have $\mathrm{E}(\hat{P}) = p$. The discrepancy from the mathematical theory results from the fact that simulations serves only as an approximation to the sampling distribution.

We get that the proposed variance in Section 10.5 is 0.0013476850, which is in good agreement with the value 0.00135 that was obtained in the simulation[19].

10.7 Summary

Glossary

Point Estimation: An attempt to obtain the best guess of the value of a population parameter. An estimator is a statistic that produces such a guess. The estimate is the observed value of the estimator.

Bias: The difference between the expectation of the estimator and the value of the parameter. An estimator is unbiased if the bias is equal to zero. Otherwise, it is biased.

Mean Square Error (MSE): A measure of the concentration of the distribution of the estimator about the value of the parameter. The mean square error of an estimator is equal to the sum of the variance and the square of the bias. If the estimator is unbiased then the mean square error is equal to the variance.

Bernoulli Random Variable: A random variable that obtains the value "1" with probability p and the value "0" with probability $1 - p$. It coincides with the Binomial$(1, p)$ distribution. Frequently, the Bernoulli random variable emerges as the indicator of the occurrence of an event.

Discuss in the forum

Performance of estimators is assessed in the context of a theoretical model for the sampling distribution of the observations. Given a criteria for optimality, an optimal estimator is an estimator that performs better than any other estimator with respect to that criteria. A robust estimator, on the other hand, is an estimator that is not sensitive to misspecification of the theoretical model. Hence, a robust estimator may be somewhat inferior to an optimal estimator in the context of an assumed model. However, if in actuality the assumed model is not a good description of reality then robust estimator will tend to perform better than the estimator denoted optimal.

Some say that optimal estimators should be preferred while other advocate the use of more robust estimators. What is your opinion?

When you formulate your answer to this question it may be useful to come up with an example from you own field of interest. Think of an estimation problem and possible estimators that can be used in the context of this problem. Try to identify a model that is natural to this problem an ask yourself in what ways may this model err in its attempt to describe the real situation in the estimation problem.

[19] It can be shown theoretically that the variance of the sample proportion, in the case of sampling from a population, is equal to $[(N-n)/(N-1)] \cdot p(1-p)/n$, where n is the sample size, and N is the population size. The factor $[(N-n)/(N-1)]$ is called *the finite population correction*. In the current setting the finite population correction is equal to 0.99851, which is practically equal to one.

As an example consider estimation of the expectation of a Uniform measurement. We demonstrated that the mid-range estimator is better than the sample average if indeed the measurements emerge from the Uniform distribution. However, if the modeling assumption is wrong then this may no longer be the case. If the distribution of the measurement in actuality is not symmetric or if the distribution is more concentrated in the center than in the tails then the performance of the mid-range estimator may deteriorate. The sample average, on the other hand is not sensitive to the distribution not being symmetric.

Formulas:

- Bias: $\text{Bias} = \text{E}(\hat{\theta}) - \theta$.
- Variance: $\text{Var}(\hat{\theta}) = \text{E}\big[(\hat{\theta} - \text{E}(\hat{\theta}))^2\big]$.
- Mean Square Error: $\text{MSE} = \text{E}\big[(\hat{\theta} - \theta)^2\big]$.