Chapter 7

The Sampling Distribution

7.1 Student Learning Objective

In this section we integrate the concept of *data* that is extracted from a sample with the concept of a *random variable*. The new element that connects between these two concepts is the notion of *sampling distribution*. The data we observe results from the specific sample that was selected. The sampling distribution, in a similar way to random variables, corresponds to all samples that could have been selected. (Or, stated in a different tense, to the sample that will be selected prior to the selection itself.) Summaries of the distribution of the data, such as the sample mean and the sample standard deviation, become random variables when considered in the context of the sampling distribution. In this section we investigate the sampling distribution of such data summaries. In particular, it is demonstrated that (for large samples) the sampling distribution of the sample average may be approximated by the Normal distribution. The mathematical theorem that proves this approximation is called the *Central Limit Theory*. By the end of this chapter, the student should be able to:

- Comprehend the notion of sampling distribution and simulate the sampling distribution of the sample average.
- Relate the expectation and standard deviation of a measurement to the expectation and standard deviation of the sample average.
- Apply the Central Limit Theorem to the sample averages.

7.2 The Sampling Distribution

In Chapter 5 the concept of a random variable was introduced. As part of the introduction we used an example that involved the selection of a random person from the population and the measuring of his/her height. Prior to the action of selection, the height of that person is a *random variable*. It has the potential of obtaining any of the heights that are present in the population, which is the *sample space* of this example, with a distribution that reflects the relative frequencies of each of the heights in the population: the *probabilities* of the values. After the selection of the person and the measuring of the height

we get a particular value. This is the *observed value* and is no longer a random variable. In this section we extend the concept of a random variable and define the concept of *a random sample*.

7.2.1 A Random Sample

The relation between the random sample and the data is similar to the relation between a random variable and the observed value. The data is the observed values of a sample taken from a population. The content of the data is known. The random sample, similarly to a random variable, is the data that *will* be selected when taking a sample, prior to the selection itself. The content of the random sample is unknown, since the sample has not yet been taken. Still, just like for the case of the random variable, one is able to say what the possible evaluations of the sample may be and, depending on the mechanism of selecting the sample, what are the probabilities of the different potential evaluations. The collection of all possible evaluations of the sample is the *sample space of the random sample* and the probabilities of the different evaluations produce the *distribution* of the random sample.

(Alternatively, if one prefers to speak in past tense, one can define the sample space of a random sample to be the evaluations of the sample that could have taken place, with the distribution of the random sample being the probabilities of these evaluations.)

A *statistic* is a function of the data. Example of statistics are the average of the data, the sample variance and standard deviation, the median of the data, etc. In each case a given formula is applied to the data. In each type of statistic a different formula is applied.

The same formula that is applied to the observed data may, in principle, be applied to random samples. Hence, for example, one may talk of the sample average, which is the average of the elements in the data. The average, considered in the context of the observed data, is a number and its value is known. However, if we think of the average in the context of a random sample then it becomes a random variable. Prior to the selection of the actual sample we do not know what values it will include. Hence, we cannot tell what the outcome of the average of the values will be. However, due to the identification of all possible evaluations that the sample can possess we may say in advance what is the collection of values the sample average can have. This is the sample space of the sample average. Moreover, from the sampling distribution of the random sample one may identify the probability of each value of the sample average, thus obtaining the *sampling distribution* of the sample average.

The same line of argumentation applies to any statistic. Computed in the context of the observed data, the statistic is a known number that may, for example, be used to characterize the variation in the data. When thinking of a statistic in the context of a random sample it becomes a random variable. The distribution of the statistic is called the sampling distribution of the statistic. Consequently, we may talk of the sampling distribution of the median, the sample distribution of the sample variance, etc.

Random variables are also applied as models for uncertainty in future measurements in more abstract settings that need not involve a specific population. Specifically, we introduced the Binomial and Poisson random variables for settings that involve counting and the Uniform, Exponential, and Normal random

variables for settings where the measurement is continuous.

The notion of a sampling distribution may be extended to a situation where one is taking several measurements, each measurement taken independently of the others. As a result one obtains a *sequence* of measurements. We use the term "sample" to denote this sequence. The distribution of this sequence is also called the sampling distribution. If all the measurements in the sequence are Binomial then we call it a *Binomial sample*. If all the measurements are Exponential we call it an *Exponential sample* and so forth.

Again, one may apply a formula (such as the average) to the content of the random sequence and produce a random variable. The term *sampling distribution* describes again the distribution that the random variable produced by the formula inherits from the sample.

In the next subsection we examine an example of a sample taken from a population. Subsequently, we discuss examples that involves a sequence of measurements from a theoretical model.

7.2.2 Sampling From a Population

Consider taking a sample from a population. Let us use again for the illustration the file "`pop1.csv`" like we did in Chapter 4. The data frame produced from the file contains the sex and hight of the 100,000 members of some imaginary population. Recall that in Chapter 4 we applied the function "`sample`" to randomly sample the height of a single subject from the population. Let us apply the same function again, but this time in order to sample the heights of 100 subjects:

```
> pop.1 <- read.csv("pop1.csv")
> X.samp <- sample(pop.1$height,100)
> X.samp
  [1] 168 177 172 174 154 179 145 160 188 172 175 174 176 144 164
 [16] 171 167 158 181 165 166 173 184 174 169 176 168 154 167 175
 [31] 178 179 175 187 160 171 175 172 178 167 181 193 163 181 168
 [46] 153 200 168 169 194 177 182 167 183 177 155 167 172 176 168
 [61] 164 162 188 163 166 156 163 185 149 163 157 155 161 177 176
 [76] 153 162 180 177 156 162 197 183 166 185 178 188 198 175 167
 [91] 185 160 148 160 174 162 161 178 159 168
```

In the first line of code we produce a data frame that contains the information on the entire population. In the second line we select a sample of size 100 from the population, and in the third line we present the content of the sample.

The first argument to the function "`sample`" that selects the sample is the sequence of length 100,000 with the list of heights of all the members of the population. The second argument indicates the sample size, 100 in this case. The outcome of the random selection is stored in the object "`X.samp`", which is a sequence that contains 100 heights.

Typically, a researcher does not get to examine the entire population. Instead, measurements on a sample from the population are made. In relation to the imaginary setting we simulate in the example, the typical situation is that the research does not have the complete list of potential measurement evaluations, i.e. the complete list of 100,000 heights in "`pop.1$height`", but only a sample of measurements, namely the list of 100 numbers that are stored in

"`X.samp`" and are presented above. The role of statistics is to make inference on the parameters of the unobserved population based on the information that is obtained from the sample.

For example, we may be interested in estimating the mean value of the heights in the population. A reasonable proposal is to use the sample average to serve as an estimate:

```
> mean(X.samp)
[1] 170.73
```

In our artificial example we can actually compute the true population mean:

```
> mean(pop.1$height)
[1] 170.035
```

Hence, we may see that although the match between the estimated value and the actual value is not perfect still they are close enough.

The actual estimate that we have obtained resulted from the specific sample that was collected. Had we collected a different subset of 100 individuals we would have obtained different numerical value for the estimate. Consequently, one may wonder: Was it pure luck that we got such good estimates? How likely is it to get estimates that are close to the target parameter?

Notice that in realistic settings we do not know the actual value of the target population parameters. Nonetheless, we would still want to have at least a probabilistic assessment of the distance between our estimates and the parameters they try to estimate. The sampling distribution is the vehicle that may enable us to address these questions.

In order to illustrate the concept of the sampling distribution let us select another sample and compute its average:

```
> X.samp <- sample(pop.1$height,100)
> X.bar <- mean(X.samp)
> X.bar
[1] 171.87
```

and do it once more:

```
> X.samp <- sample(pop.1$height,100)
> X.bar <- mean(X.samp)
> X.bar
[1] 171.02
```

In each case we got a different value for the sample average. In the first of the last two iterations the result was more than 1 centimeter away from the population average, which is equal to 170.035, and in the second it was within the range of 1 centimeter. Can we say, prior to taking the sample, what is the probability of falling within 1 centimeter of the population mean?

Chapter 4 discussed the random variable that emerges by randomly sampling a single number from the population presented by the sequence "`pop.1$height`". The distribution of the random variable resulted from the assignment of the probability 1/100,000 to each one of the 100,000 possible outcomes. The same principle applies when we randomly sample 100 individuals. Each possible outcome is a collection of 100 numbers and each collection is assigned equal probability. The resulting distribution is called *the sampling distribution*.

The distribution of the average of the sample emerges from this distribution: With each sample one may associate the average of that sample. The probability assigned to that average outcome is the probability of the sample. Hence, one may assess the probability of falling within 1 centimeter of the population mean using the sampling distribution. Each sample produces an average that either falls within the given range or not. The probability of the sample average falling within the given range is the proportion of samples for which this event happens among the entire collection of samples.

However, we face a technical difficulty when we attempt to assess the sampling distribution of the average and the probability of falling within 1 centimeter of the population mean. Examination of the distribution of a sample of a single individual is easy enough. The total number of outcomes, which is 100,000 in the given example, can be handled with no effort by the computer. However, when we consider samples of size 100 we get that the total number of ways to select 100 number out of 100,000 numbers is in the order of 10^{342} (1 followed by 342 zeros) and cannot be handled by any computer. Thus, the probability cannot be computed.

As a compromise we will approximate the distribution by selecting a large number of samples, say 100,000, to represent the entire collection, and use the resulting distribution as an approximation of the sampling distribution. Indeed, the larger the number of samples that we create the more accurate the approximation of the distribution is. Still, taking 100,000 repeats should produce approximations which are good enough for our purposes.

Consider the sampling distribution of the sample average. We simulated above a few examples of the average. Now we would like to simulate 100,000 such examples. We do this by creating first a sequence of the length of the number of evaluations we seek (100,000) and then write a small program that produces each time a new random sample of size 100 and assigns the value of the average of that sample to the appropriate position in the sequence. Do first and explain later[1]:

```
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- sample(pop.1$height,100)
+     X.bar[i] <- mean(X.samp)
+ }
> hist(X.bar)
```

In the first line we produce a sequence of length 100,000 that contains zeros. The function "`rep`" creates a sequence that contains repeats of its first argument a number of times that is specified by its second argument. In this example, the numerical value 0 is repeated 100,000 times to produce a sequence of zeros of the length we seek.

[1]Running this simulation, and similar simulations of the same nature that will be considered in the sequel, demands more of the computer's resources than the examples that were considered up until now. Beware that running times may be long and, depending on the strength of your computer and your patience, too long. You may save time by running less iterations, replacing, say, "`10^5`" by "`10^4`". The results of the simulation will be less accurate, but will still be meaningful.

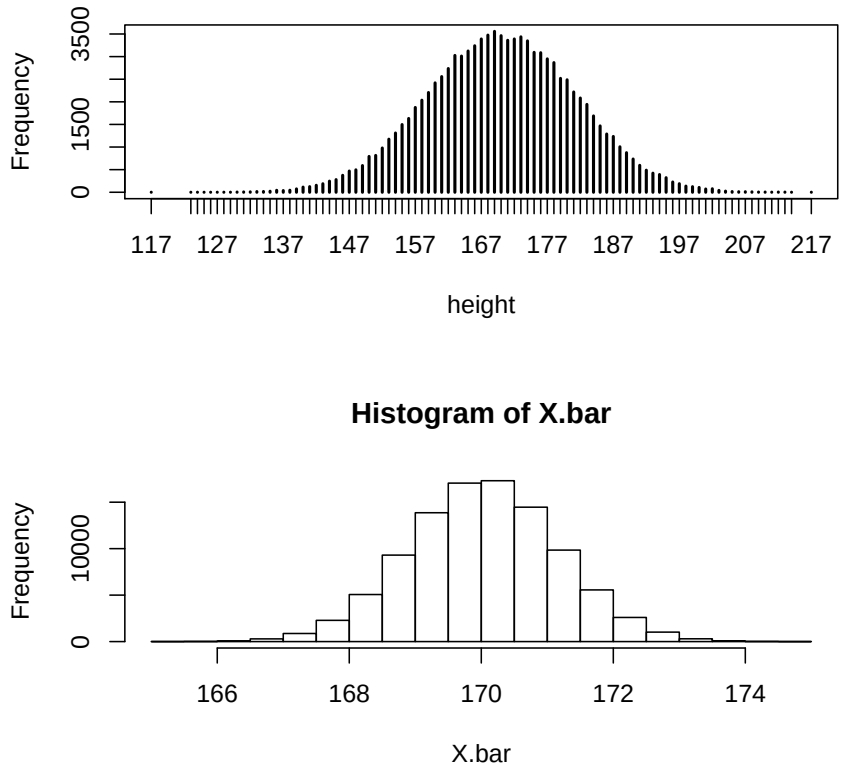

Figure 7.1: Distribution of Height and the Sampling Distribution of Averages

The main part of the program is a "`for`" loop. The argument of the function "`for`" takes the special form: "*index.name* `in` *index.values*", where *index.name* is the name of the running index and *index.values* is the collection of values over which the running index is evaluated. In each iteration of the loop the running index is assigned a value from the collection and the expression that follows the brackets of the "`for`" function is evaluated with the given value of the running index.

In the given example the collection of values is produced by the expression "`1:n`". Recall that the expression "`1:n`" produces the collection of integers between `1` and `n`. Here, `n` = 100,000. Hence, in the given application the collection of values is a sequence that contains the integers between 1 and 100,000. The running index is called "`i`". the expression is evaluated 100,000 times, each time with a different integer value for the running index "`i`".

The `R` system treats a collection of expressions enclosed within curly brackets as one entity. Therefore, in each iteration of the "`for`" loop, the lines that are within the curly brackets are evaluated. In the first line a random sample of size 100 is produced and in the second line the average of the sample is computed and stored in the i-th position of the sequence "`X.bar`". Observe that the specific

position in the sequence is referred to by using square brackets.

The program changes the original components of the sequence, from 0 to the average of a random sample, one by one. When the loop ends all values are changed and the sequence "`X.bar`" contains 100,000 evaluations of the sample average. The last line, which is outside the curly brackets and is evaluated after the "`for`" loop ends, produces an histogram of the averages that were simulated. The histogram is presented in the lower panel of Figure 7.1.

Compare the distribution of the sample average to the distribution of the heights in the population that was presented first in Figure 4.1 and is currently presented in the upper panel of Figure 7.1. Observe that both distributions are centered at about 170 centimeters. Notice, however, that the range of values of the sample average lies essentially between 166 and 174 centimeters, whereas the range of the distribution of heights themselves is between 127 and 217 centimeter. Broadly speaking, the sample average and the original measurement are centered around the same location but the sample average is less spread.

Specifically, let us compare the expectation and standard deviation of the sample average to the expectation and standard deviation of the original measurement:

```
> mean(pop.1$height)
[1] 170.035
> sd(pop.1$height)
[1] 11.23205
> mean(X.bar)
[1] 170.037
> sd(X.bar)
[1] 1.122116
```

Observe that the expectation of the population and the expectation of the sample average, are practically the same, the standard deviation of the sample average is about 10 times smaller than the standard deviation of the population. This result is not accidental and actually reflects a general phenomena that will be seen below in other examples.

We may use the simulated sampling distribution in order to compute an approximation of the probability of the sample average falling within 1 centimeter of the population mean. Let us first compute the relevant probability and then explain the details of the computation:

```
> mean(abs(X.bar - mean(pop.1$height)) <= 1)
[1] 0.62589
```

Hence we get that the probability of the given event is about 62.6%.

The object "`X.bar`" is a sequence of length 100,000 that contains the simulated sample averages. This sequence represents the distribution of the sample average. The expression "`abs(X.bar - mean(pop.1$height)) <= 1`" produces a sequence of logical "`TRUE`" or "`FALSE`" values, depending on the value of the sample average being less or more than one unit away from the population mean. The application of the function "`mean`" to the output of the last expression results in the computation of the relative frequency of `TRUE`s, which corresponds to the probability of the event of interest.

Example 7.1. *A poll for the determination of the support in the population for a candidate was describe in Example 5.1. The proportion in the population of supporters was denoted by p. A sample of size $n = 300$ was considered in order to estimate the size of p. We identified that the distribution of X, the number of supporters in the sample, is* Binomial$(300, p)$. *This distribution is the sampling distribution*[2] *of X. One may use the proportion in the sample of supporters, the number of supporters in the sample divided by 300, as an estimate to the parameter p. The sampling distribution of this quantity, $X/300$, may be considered in order to assess the discrepancy between the estimate and the actual value of the parameter.*

7.2.3 Theoretical Models

Sampling distribution can also be considered in the context of theoretical distribution models. For example, take a measurement $X \sim$ Binomial$(10, 0.5)$ from the Binomial distribution. Assume 64 independent measurements are produced with this distribution: $X_1, X_2, \ldots, X_{64}$. The sample average in this case corresponds to the distribution of the random variable produced by averaging these 64 random variables:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{64}}{64} = \frac{1}{64} \sum_{i=1}^{64} X_i \ .$$

Again, one may wonder what is the distribution of the sample average $\bar{X}$ in this case?

We can approximate the distribution of the sample average by simulation. The function "`rbinom`" produces a random sample from the Binomial distribution. The first argument to the function is the sample size, which we take in this example to be equal to 64. The second and third arguments are the parameters of the Binomial distribution, 10 and 0.5 in this case. We can use this function in the simulation:

```
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- rbinom(64,10,0.5)
+     X.bar[i] <- mean(X.samp)
+ }
```

Observe that in this code we created a sequence of length 100,000 with evaluations of the sample average of 64 Binomial random variables. We start with a sequence of zeros and in each iteration of the "`for`" loop a zero is replaced by the average of a random sample of 64 Binomial random variables.

Examine the sampling distribution of the Binomial average:

```
> hist(X.bar)
```

[2] Mathematically speaking, the Binomial distribution is only an approximation to the sampling distribution of X. Actually, the Binomial is an exact description to the distribution only in the case where each subject has the chance be represented in the sample more than once. However, only when the size of the sample is comparable to the size of the population would the Binomial distribution fail to be an adequate approximation to the sampling distribution.

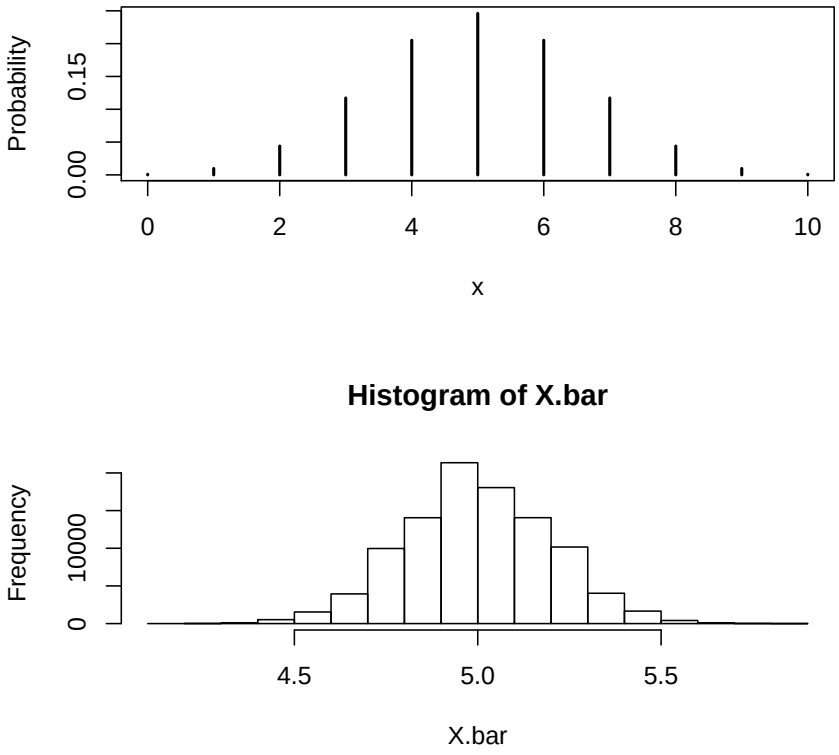

Figure 7.2: Distributions of an Average and a Single Binomial(10,0.5)

```
> mean(X.bar)
[1] 4.999074
> sd(X.bar)
[1] 0.1982219
```

The histogram of the sample average is presented in the lower panel of Figure 7.2. Compare it to the distribution of a single Binomial random variable that appears in the upper panel. Notice, once more, that the center of the two distributions coincide but the spread of the sample average is smaller. The sample space of a single Binomial random variable is composed of integers. The sample space of the average of 64 Binomial random variables, on the other hand, contains many more values and is closer to the sample space of a random variable with a continuous distribution.

Recall that the expectation of a Binomial$(10, 0.5)$ random variable is $\mathrm{E}(X) = 10 \cdot 0.5 = 5$ and the variance is $\mathrm{Var}(X) = 10 \cdot 0.5 \cdot 0.5 = 2.5$ (thus, the standard deviation is $\sqrt{2.5} = 1.581139$). Observe that the expectation of the sample average that we got from the simulation is essentially equal to 5 and the standard deviation is 0.1982219.

One may prove mathematically that the expectation of the sample mean is equal to the theoretical expectation of its components:

$$\mathrm{E}(\bar{X}) = \mathrm{E}(X) \ .$$

The results of the simulation for the expectation of the sample average are consistent with the mathematical statement. The mathematical theory of probability may also be used in order to prove that the variance of the sample average is equal to the variance of each of the components, divided by the sample size:

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

here n is the number of observations in the sample. Specifically, in the Binomial example we get that $\mathrm{Var}(\bar{X}) = 2.5/64$, since the variance of a Binomial component is 2.5 and there are 64 observations. Consequently, the standard deviation is $\sqrt{2.5/64} = 0.1976424$, in agreement, more or less, with the results of the simulation (that produced 0.1982219 as the standard deviation).

Consider the problem of identifying the central interval that contains 95% of the distribution. In the Normal distribution we were able to use the function "`qnorm`" in order to compute the percentiles of the theoretical distribution. A function that can be used for the same purpose for simulated distribution is the function "`quantile`". The first argument to this function is the sequence of simulated values of the statistic, "`X.bar`" in the current case. The second argument is a number between 0 and 1, or a sequence of such numbers:

```
> quantile(X.bar,c(0.025,0.975))
    2.5%    97.5%
4.609375 5.390625
```

We used the sequence "`c(0.025,0.975)`" as the input to the second argument. As a result we obtained the output 4.609375, which is the 2.5%-percentile of the sampling distribution of the average, and 5.390625, which is the 97.5%-percentile of the sampling distribution of the average.

Of interest is to compare these percentiles to the parallel percentiles of the Normal distribution with the same expectation and the same standard deviation as the average of the Binomials:

```
> qnorm(c(0.025,0.975),mean(X.bar),sd(X.bar))
[1] 4.611456 5.389266
```

Observe the similarity between the percentiles of the distribution of the average and the percentiles of the Normal distribution. This similarity is a reflection of the Normal approximation of the sampling distribution of the average, which is formulated in the next section under the title: *The Central Limit Theorem.*

Example 7.2. *The distribution of the number of events of radio active decay in a second was modeled in Example 5.3 according to the Poisson distribution. A quantity of interest is λ, the expectation of that Poisson distribution. This quantity may be estimated by measuring the total number of decays over a period of time and dividing the outcome by the number of seconds in that period of time. Let n be this number of second. The procedure just described corresponds to taking the sample average of* Poisson(λ) *observations for a sample of size n.*

The expectation of the sample average is λ and the variance is λ/n, leading to a standard deviation of size $\sqrt{\lambda/n}$. The Central Limit Theorem states that the sampling distribution of this average corresponds, approximately, to the Normal distribution with this expectation and standard deviation.

7.3 Law of Large Numbers and Central Limit Theorem

The Law of Large Numbers and the Central Limit Theorem are mathematical theorems that describe the sampling distribution of the average for large samples.

7.3.1 The Law of Large Numbers

The Law of Large Numbers states that, as the sample size becomes larger, the sampling distribution of the sample average becomes more and more concentrated about the expectation.

Let us demonstrate the Law of Large Numbers in the context of the Uniform distribution. Let the distribution of the measurement X be Uniform$(3, 7)$. Consider three different sample sizes n: $n = 10$, $n = 100$, and $n = 1000$. Let us carry out a simulation similar to the simulations of the previous section. However, this time we run the simulation for the three sample sizes in parallel:

```
> unif.10 <- rep(0,10^5)
> unif.100 <- rep(0,10^5)
> unif.1000 <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp.10 <- runif(10,3,7)
+     unif.10[i] <- mean(X.samp.10)
+     X.samp.100 <- runif(100,3,7)
+     unif.100[i] <- mean(X.samp.100)
+     X.samp.1000 <- runif(1000,3,7)
+     unif.1000[i] <- mean(X.samp.1000)
+ }
```

Observe that we have produced 3 sequences of length 100,000 each: "`unif.10`", "`unif.100`", and "`unif.1000`". The first sequence is an approximation of the sampling distribution of an average of 10 independent Uniform measurements, the second approximates the sampling distribution of an average of 100 measurements and the third the distribution of an average of 1000 measurements. The distribution of single measurement in each of the examples is Uniform$(3, 7)$.

Consider the expectation of sample average for the three sample sizes:

```
> mean(unif.10)
[1] 4.999512
> mean(unif.100)
[1] 4.999892
> mean(unif.1000)
[1] 4.99996
```

For all sample size the expectation of the sample average is equal to 5, which is the expectation of the Uniform$(3, 7)$ distribution.

Recall that the variance of the Uniform(a, b) distribution is $(b - a)^2/12$. Hence, the variance of the given Uniform distribution is $\mathrm{Var}(X) = (7-3)^2/12 = 16/12 \approx 1.3333$. The variances of the sample averages are:

```
> var(unif.10)
[1] 0.1331749
> var(unif.100)
[1] 0.01333089
> var(unif.1000)
[1] 0.001331985
```

Notice that the variances decrease with the increase of the sample sizes. The decrease is according to the formula $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$.

The variance is a measure of the spread of the distribution about the expectation. The smaller the variance the more concentrated is the distribution around the expectation. Consequently, in agreement with the Law of Large Numbers, the larger the sample size the more concentrated is the sampling distribution of the sample average about the expectation.

7.3.2 The Central Limit Theorem (CLT)

The Law of Large Numbers states that the distribution of the sample average tends to be more concentrated as the sample size increases. The Central Limit Theorem (CLT in short) provides an approximation of this distribution.

The deviation between the sample average and the expectation of the measurement tend to decreases with the increase in sample size. In order to obtain a refined assessment of this deviation one needs to magnify it. The appropriate way to obtain the magnification is to consider the standardized sample average, in which the deviation of the sample average from its expectation is divided by the standard deviation of the sample average:

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\sqrt{\mathrm{Var}(\bar{X})}} \,.$$

Recall that the expectation of the sample average is equal to the expectation of a single random variable ($\mathrm{E}(\bar{X}) = \mathrm{E}(X)$) and that the variance of the sample average is equal to the variance of a single observation, divided by the sample size ($\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$). Consequently, one may rewrite the standardized sample average in the form:

$$Z = \frac{\bar{X} - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)/n}} = \frac{\sqrt{n}(\bar{X} - \mathrm{E}(X))}{\sqrt{\mathrm{Var}(X)}} \,.$$

The second equality follows from placing in the numerator the square root of n which *divides* the term in the denominator. Observe that with the increase of the sample size the decreasing difference between the average and the expectation is magnified by the square root of n.

The Central Limit Theorem states that, with the increase in sample size, the sample average converges (after standardization) to the standard Normal distribution.

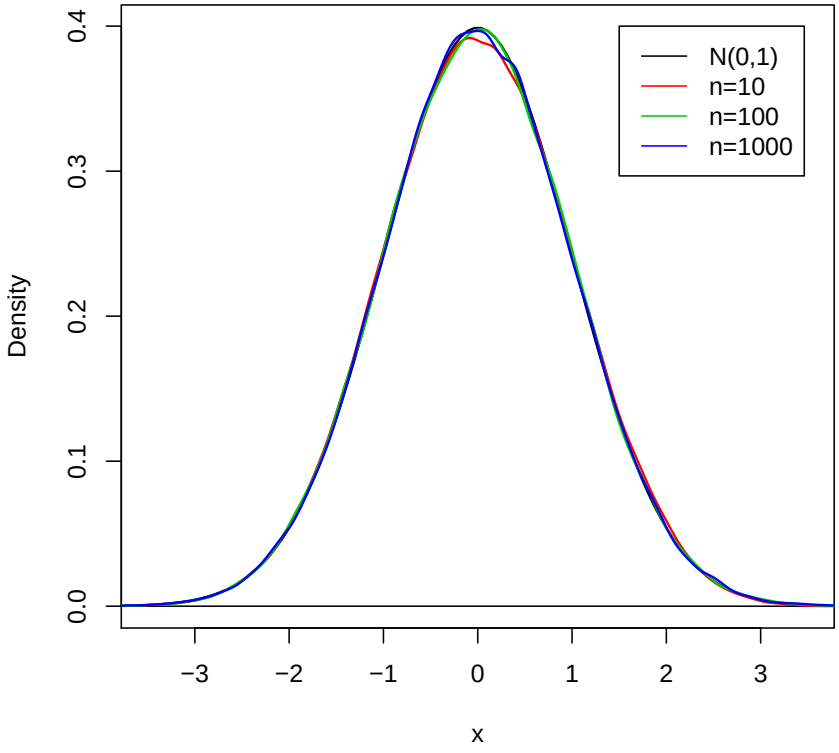

Figure 7.3: The CLT for the Uniform(3,7) Distribution

Let us examine the Central Normal Theorem in the context of the example of the Uniform measurement. In Figure 7.3 you may find the (approximated) density of the standardized average for the three sample sizes based on the simulation that we carried out previously (as *red*, *green*, and *blue* lines). Along side with these densities you may also find the theoretical density of the standard Normal distribution (as a *black* line). Observe that the four curves are almost one on top of the other, proposing that the approximation of the distribution of the average by the Normal distribution is good even for a sample size as small as $n = 10$.

However, before jumping to the conclusion that the Central Limit Theorem applies to any sample size, let us consider another example. In this example we repeat the same simulation that we did with the Uniform distribution, but this time we take Exponential(0.5) measurements instead:

```
> exp.10 <- rep(0,10^5)
> exp.100 <- rep(0,10^5)
> exp.1000 <- rep(0,10^5)
> for(i in 1:10^5)
```

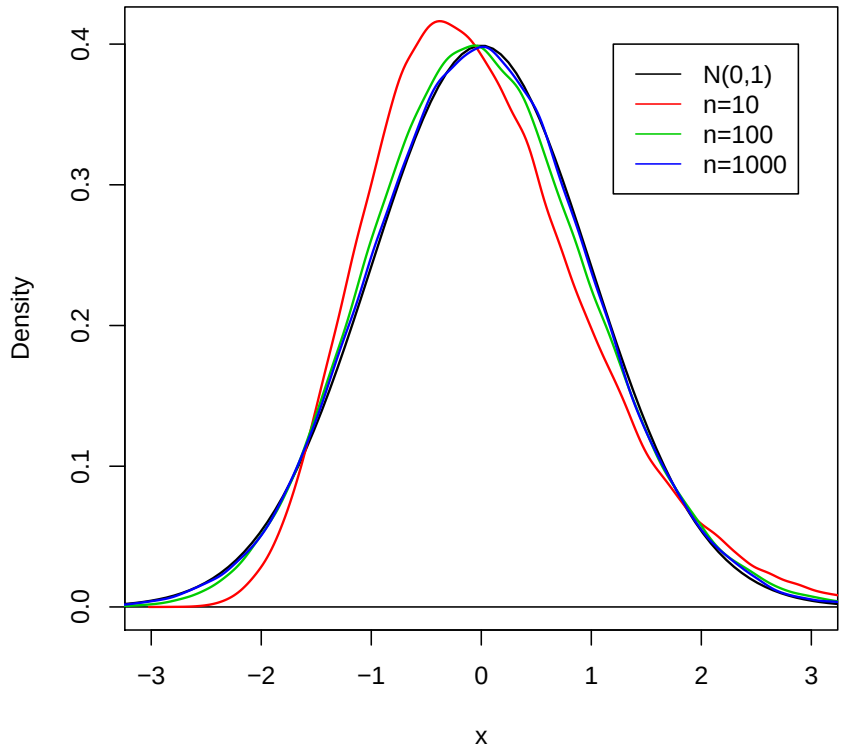

Figure 7.4: The CLT for the Exponential(0.5) Distribution

```
+ {
+     X.samp.10 <- rexp(10,0.5)
+     exp.10[i] <- mean(X.samp.10)
+     X.samp.100 <- rexp(100,0.5)
+     exp.100[i] <- mean(X.samp.100)
+     X.samp.1000 <- rexp(1000,0.5)
+     exp.1000[i] <- mean(X.samp.1000)
+ }
```

The expectation of an Exponential(0.5) random variable is $\mathrm{E}(X) = 1/\lambda = 1/0.5 = 2$ and the variance is $\mathrm{Var}(X) = 1/\lambda^2 = 1/(0.5)^2 = 4$. Observe below that the expectations of the sample averages are equal to the expectation of the measurement and the variances of the sample averages follow the relation $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$:

```
> mean(exp.10)
[1] 1.999888
> mean(exp.100)
[1] 2.000195
```

```
> mean(exp.1000)
[1] 1.999968
```

So the expectations of the sample average are all equal to 2. For the variance we get:

```
> var(exp.10)
[1] 0.4034642
> var(exp.100)
[1] 0.03999479
> var(exp.1000)
[1] 0.004002908
```

Which is in agreement with the decrease proposed by the theory,

However, when one examines the densities of the sample averages in Figure 7.4 one may see a clear distinction between the sampling distribution of the average for a sample of size 10 and the normal distribution (compare the *red* curve to the *black* curve. The match between the *green* curve that corresponds to a sample of size $n = 100$ and the *black* line is better, but not perfect. When the sample size is as large as $n = 1000$ (the *blue* curve) then the agreement with the normal curve is very good.

7.3.3 Applying the Central Limit Theorem

The conclusion of the Central Limit Theorem is that the sampling distribution of the sample average can be approximated by the Normal distribution, regardless what is the distribution of the original measurement, but provided that the sample size is large enough. This statement is very important, since it allows us, in the context of the sample average, to carry out probabilistic computations using the Normal distribution even if we do not know the actual distribution of the measurement. All we need to know for the computation are the expectation of the measurement, its variance (or standard deviation) and the sample size.

The theorem can be applied whenever probability computations associated with the sampling distribution of the average are required. The computation of the approximation is carried out by using the Normal distribution with the same expectation and the same standard deviation as the sample average.

An example of such computation was conducted in Subsection 7.2.3 where the central interval that contains 95% of the sampling distribution of a Binomial average was required. The 2.5%- and the 97.5%-percentiles of the Normal distribution with the same expectation and variance as the sample average produced boundaries for the interval. These boundaries were in good agreement with the boundaries produced by the simulation. More examples will be provided in the Solved Exercises of this chapter and the next one.

With all its usefulness, one should treat the Central Limit Theorem with a grain of salt. The approximation may be valid for large samples, but may be bad for samples that are not large enough. When the sample is small a careless application of the Central Limit Theorem may produce misleading conclusions.

7.4 Solved Exercises

Question 7.1. The file "`pop2.csv`" 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`). The variables in this file are:

id: A numerical variable. A 7 digits number that serves as a unique identifier of the subject.

sex: A factor variable. The sex of each subject. The values are either "`MALE`" or "`FEMALE`".

age: A numerical variable. The age of each subject.

bmi: A numerical variable. The body mass index of each subject.

systolic: A numerical variable. The systolic blood pressure of each subject.

diastolic: A numerical variable. The diastolic blood pressure of each subject.

group: A factor variable. The blood pressure category of each subject. The values are "`NORMAL`" both the systolic blood pressure is within its normal range (between 90 and 139) and the diastolic blood pressure is within its normal range (between 60 and 89). The value is "`HIGH`" if either measurements of blood pressure are above their normal upper limits and it is "`LOW`" if either measurements are below their normal lower limits.

Our goal in this question is to investigate the sampling distribution of the sample average of the variable "`bmi`". We assume a sample of size $n = 150$.

1. Compute the population average of the variable "`bmi`".
2. Compute the population standard deviation of the variable "`bmi`".
3. Compute the expectation of the sampling distribution for the sample average of the variable.
4. Compute the standard deviation of the sampling distribution for the sample average of the variable.
5. Identify, using simulations, the central region that contains 80% of the sampling distribution of the sample average.
6. Identify, using the Central Limit Theorem, an approximation of the central region that contains 80% of the sampling distribution of the sample average.

Solution (to Question 7.1.1): After placing the file "`pop2.csv`" in the working directory one may produce a data frame with the content of the file and compute the average of the variable "`bmi`" using the code:

```
> pop.2 <- read.csv(file="pop2.csv")
> mean(pop.2$bmi)
[1] 24.98446
```

We obtain that the population average of the variable is equal to 24.98446.

Solution (to Question 7.1.2): Applying the function "`sd`" to the sequence of population values produces the population standard deviation:

```
> sd(pop.2$bmi)
[1] 4.188511
```

In turns out that the standard deviation of the measurement is 4.188511.

Solution (to Question 7.1.3): In order to compute the expectation under the sampling distribution of the sample average we conduct a simulation. The simulation produces (an approximation) of the sampling distribution of the sample average. The sampling distribution is represented by the content of the sequence "`X.bar`":

```
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- sample(pop.2$bmi,150)
+     X.bar[i] <- mean(X.samp)
+ }
> mean(X.bar)
[1] 24.98681
```

Initially, we produce a vector of zeros of the given lenght (100,000). In each iteration of the "`for`" loop a random sample of size 150 is selected from the population. The sample average is computed and stored in the sequence "`X.bar`". At the end of all the iterations all the zeros are replaced by evaluations of the sample average.

The expectation of the sampling distribution of the sample average is computed by the application of the function "`mean`" to the sequence that represents the sampling distribution of the sample average. The result for the current is 24.98681, which is vary similar[3] to the population average 24.98446.

Solution (to Question 7.1.4): The standard deviation of the sample average under the sampling distribution is computed using the function "`sd`":

```
> sd(X.bar)
[1] 0.3422717
```

The resulting standard deviation is 0.3422717. Recall that the standard deviation of a single measurement is equal to 4.188511 and that the sample size is $n = 150$. The ratio between the standard deviation of the measurement and the square root of 150 is $4.188511/\sqrt{150} = 0.3419905$, which is similar in value to the standard deviation of the sample average[4].

[3] Theoretically, the two numbers should coincide. The small discrepancy follows from the fact that the sequence "`X.bar`" is only an approximation of the sampling distribution.

[4] It can be shown mathematically that the variance of the sample average, in the case of sampling from a population, is equal to $[(N-n)/(N-1)] \cdot \mathrm{Var}(X)/n$, where $\mathrm{Var}(X)$ is the population variance of the measurement, 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.

Solution (to Question 7.1.5): The central region that contains 80% of the sampling distribution of the sample average can be identified with the aid of the function "`quantile`":

```
> quantile(X.bar,c(0.1,0.9))
     10%      90%
24.54972 25.42629
```

The value 24.54972 is the 10%-percentile of the sampling distribution. To the left of this value are 10% of the distribution. The value 25.42629 is the 90%-percentile of the sampling distribution. To the right of this value are 10% of the distribution. Between these two values are 80% of the sampling distribution.

Solution (to Question 7.1.6): The Normal approximation, which is the conclusion of the Central Limit Theorem substitutes the sampling distribution of the sample average by the Normal distribution with the same expectation and standard deviation. The percentiles are computed with the function "`qnorm`":

```
> qnorm(c(0.1,0.9),mean(X.bar),sd(X.bar))
[1] 24.54817 25.42545
```

Observe that we used the expectation and the standard deviation of the sample average in the function. The resulting interval is $[24.54817, 25.42545]$, which is similar to the interval $[24.54972, 25.42629]$ which was obtained via simulations.

Question 7.2. A subatomic particle hits a linear detector at random locations. The length of the detector is 10 nm and the hits are uniformly distributed. The location of 25 random hits, measured from a specified endpoint of the interval, are marked and the average of the location computed.

1. What is the expectation of the average location?
2. What is the standard deviation of the average location?
3. Use the Central Limit Theorem in order to approximate the probability the average location is in the left-most third of the linear detector.
4. The central region that contains 99% of the distribution of the average is of the form $5 \pm c$. Use the Central Limit Theorem in order to approximate the value of c.

Solution (to Question 7.2.1): Denote by X the distance from the specified endpoint of a random hit. Observe that $X \sim \text{Uniform}(0, 10)$. The 25 hits form a sample $X_1, X_2, \ldots, X_{25}$ from this distribution and the sample average $\bar{X}$ is the average of these random locations. The expectation of the average is equal to the expectation of a single measurement. Since $\text{E}(X) = (a + b)/2 = (0 + 10)/2 = 5$ we get that $\text{E}(\bar{X}) = 5$.

Solution (to Question 7.2.2): The variance of the sample average is equal to the variance of a single measurement, divided by the sample size. The variance of the Uniform distribution is $\text{Var}(X) = (a+b)^2/12 = (10-0)^2/12 = 8.333333$. The standard deviation of the sample average is equal to the standard deviation

of the sample average is equal to the standard deviation of a single measurement, divided by the square root of the sample size. The sample size is $n = 25$. Consequently, the standard deviation of the average is $\sqrt{8.333333/25} = 0.5773503$.

Solution (to Question 7.2.3): The left-most third of the detector is the interval to the left of 10/3. The distribution of the sample average, according to the Central Limit Theorem, is Normal. The probability of being less than 10/3 for the Normal distribution may be computed with the function "`pnorm`":

```
> mu <- 5
> sig <- sqrt(10^2/(12*25))
> pnorm(10/3,mu,sig)
[1] 0.001946209
```

The expectation and the standard deviation of the sample average are used in computation of the probability. The probability is 0.001946209, about 0.2%.

Solution (to Question 7.2.3): The central region in the Normal(μ, σ^2) distribution that contains 99% of the distribution is of the form $\mu \pm$ `qnorm(0.995)` $\cdot \sigma$, where "`qnorm(0.995)`" is the 99.5%-percentile of the Standard Normal distribution. Therefore, $c =$ `qnorm(0.995)` $\cdot \sigma$:

```
> qnorm(0.995)*sig
[1] 1.487156
```

We get that $c = 1.487156$.

7.5 Summary

Glossary

Random Sample: The probabilistic model for the values of a measurements in the sample, before the measurement is taken.

Sampling Distribution: The distribution of a random sample.

Sampling Distribution of a Statistic: A statistic is a function of the data; i.e. a formula applied to the data. The statistic becomes a random variable when the formula is applied to a random sample. The distribution of this random variable, which is inherited from the distribution of the sample, is its sampling distribution.

Sampling Distribution of the Sample Average: The distribution of the sample average, considered as a random variable.

The Law of Large Numbers: A mathematical result regarding the sampling distribution of the sample average. States that the distribution of the average of measurements is highly concentrated in the vicinity of the expectation of a measurement when the sample size is large.

The Central Limit Theorem: A mathematical result regarding the sampling distribution of the sample average. States that the distribution of the average is approximately Normal when the sample size is large.

Discussion in the Forum

Limit theorems in mathematics deal with the convergence of some property to a limit as some indexing parameter goes to infinity. The Law of Large Numbers and the Central Limit Theorem are examples of limit theorems. The property they consider is the sampling distribution of the sample average. The indexing parameter that goes to infinity is the sample size n.

Some people say that the Law of Large Numbers and the Central Limit Theorem are useless for practical purposes. These theorems deal with a sample size that goes to infinity. However, all sample sizes one finds in reality are necessarily finite. What is your opinion?

When forming your answer to this question you may give an example of a situation from your own field of interest in which conclusions of an abstract mathematical theory are used in order to solve a practical problem. Identify the merits and weaknesses of the application of the mathematical theory.

For example, in making statistical inference one frequently needs to make statements regarding the sampling distribution of the sample average. For instant, one may want to identify the central region that contains 95% of the distribution. The Normal distribution is used in the computation. The justification is the Central Limit Theorem.

Summary of Formulas

Expectation of the sample average: $\mathrm{E}(\bar{X}) = \mathrm{E}(X)$

Variance of the sample average: $\mathrm{V}ar(\bar{X}) = \mathrm{V}ar(X)/n$

Chapter 8

Overview and Integration

8.1 Student Learning Objective

This section provides an overview of the concepts and methods that where presented in the first part of the book. We attempt to relate them to each other and put them in prospective. Some problems are provided. The solutions to these problems require combinations of many of the tools that were presented in previous chapters. By the end of this chapter, the student should be able to:

- Have a better understanding of the relation between descriptive statistics, probability, and inferential statistics.
- Distinguish between the different uses of the concept of variability.
- Integrate the tools that were given in the first part of the book in order to solve complex problems.

8.2 An Overview

The purpose of the first part of the book was to introduce the fundamentals of statistics and teach the concepts of probability which are essential for the understanding of the statistical procedures that are used to analyze data. These procedures are presented and discussed in the second part of the book.

Data is typically obtained by selecting a sample from a population and taking measurements on the sample. There are many ways to select a sample, but all methods for such selection should not violate the most important characteristic that a sample should posses, namely that it represents the population it came from. In this book we concentrate on simple random sampling. However, the reader should be aware of the fact that other sampling designs exist and may be more appropriate in specific applications. Given the sampled data, the main concern of the science of statistics is in making inference on the parameter of the population on the basis of the data collected. Such inferences are carried out with the aid of statistics, which are functions of the data.

Data is frequently stored in the format of a data frame, in which columns are the measured variable and the rows are the observations associated with the selected sample. The main types of variables are numeric, either discrete or not,

and factors. We learned how one can produce data frames and read data into R for further analysis.

Statistics is geared towards dealing with variability. Variability may emerge in different forms and for different reasons. It can be summarized, analyzed and handled with many tools. Frequently, the same tool, or tools that have much resemblance to each other, may be applied in different settings and for different forms of variability. In order not to loose track it is important to understand in each scenario the source and nature of the variability that is being examined.

An important split in term of the source of variability is between descriptive statistics and probability. Descriptive statistics examines the distribution of data. The frame of reference is the data itself. Plots, such as the bar plots, histograms and box plot; tables, such as the frequency and relative frequency as well as the cumulative relative frequency; and numerical summaries, such as the mean, median and standard deviation, can all serve in order to understand the distribution of the given data set.

In probability, on the other hand, the frame of reference is not the data at hand but, instead, it is all data sets that could have been sampled (the sample space of the sampling distribution). One may use similar plots, tables, and numerical summaries in order to analyze the distribution of functions of the sample (statistics), but the meaning of the analysis is different. As a matter of fact, the relevance of the probabilistic analysis to the data actually sampled is indirect. The given sample is only one realization within the sample space among all possible realizations. In the probabilistic context there is no special role to the observed realization in comparison to all other potential realizations.

The fact that the relation between probabilistic variability and the observed data is not direct does not make the relation unimportant. On the contrary, this indirect relation is the basis for making statistical inference. In statistical inference the characteristics of the data may be used in order to extrapolate from the sampled data to the entire population. Probabilistic description of the distribution of the sample is then used in order to assess the reliability of the extrapolation. For example, one may try to estimate the value of population parameters, such as the population average and the population standard deviation, on the basis of the parallel characteristics of the data. The variability of the sampling distribution is used in order to quantify the accuracy of this estimation. (See Example 5 below.)

Statistics, like many other empirically driven forms of science, uses theoretical modeling for assessing and interpreting observational data. In statistics this modeling component usually takes the form of a probabilistic model for the measurements as random variables. In the first part of this book we have encountered several such models. The model of simple sampling assumed that each subset of a given size from the population has equal probability to be selected as the sample. Other, more structured models, assumed a specific form to the distribution of the measurements. The examples we considered were the Binomial, the Poisson, the Uniform, the Exponential and the Normal distributions. Many more models may be found in the literature and may be applied when appropriate. Some of these other models have R functions that can be used in order to compute the distribution and produce simulations.

A statistic is a function of sampled data that is used for making statistical inference. When a statistic, such as the average, is computed on a random sample then the outcome, from a probabilistic point of view, is a random vari-

able. The distribution of this random variable depends on the distribution of the measurements that form the sample but is not identical to that distribution. Hence, for example, the distribution of an average of a sample from the Uniform distribution does not follow the Uniform distribution. In general, the relation between the distribution of a measurement and the distribution of a statistic computed from a sample that is generated from that distribution may be complex. Luckily, in the case of the sample average the relation is rather simple, at least for samples that are large enough.

The Central Limit Theorem provides an approximation of the distribution of the sample average that typically improves with the increase in sample size. The expectation of the sample average is equal to the expectation of a single measurement and the variance is equal to the variance of a single measurement, divided by the sample size. The Central Limit Theorem adds to this observation the statement that the distribution of the sample average may be approximated by the Normal distribution (with the same expectation and standard deviation as those of the sample average). This approximation is valid for practically any distribution of the measurement. The conclusion is, at least in the case of the sample average, that the distribution of the statistic depends on the underlying distribution of the measurements only through their expectation and variance but not through other characteristics of the distribution.

The conclusion of the theorem extends to quantities proportional to the sample average. Therefore, since the sum of the sample is obtained by multiplying the sample average by the sample size n, we get that the theorem can be used in order to approximate the distribution of sums. As a matter of fact, the theorem may be generalized much further. For example, it may be shown to hold for a smooth function of the sample average, thereby increasing the applicability of the theorem and its importance.

In the next section we will solve some practical problems. In order to solve these problems you are required to be familiar with the concepts and tools that were introduced throughout the first part of the book. Hence, we strongly recommend that you read again and review all the chapters of the book that preceded this one before moving on to the next section.

8.3 Integrated Applications

The main message of the Central Limit Theorem is that for the sample average we may compute probabilities based on the Normal distribution and obtain reasonable approximations, provided that the sample size is not too small. All we need to figure out for the computations are the expectation and variance of the underlying measurement. Otherwise, the exact distribution of that measurement is irrelevant. Let us demonstrate the applicability of the Central Limit Theorem in two examples.

8.3.1 Example 1

A study involving stress is done on a college campus among the students. The stress scores follow a (continuous) Uniform distribution with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:

1. The probability that the average stress score for the 75 students is less than 2.
2. The 90th percentile for the average stress score for the 75 students.
3. The probability that the total of the 75 stress scores is less than 200.
4. The 90th percentile for the total stress score for the 75 students.

Solution:

Denote by X the stress score of a random student. We are given that $X \sim \text{Uniform}(1, 5)$. We use the formulas $\text{E}(X) = (a+b)/2$ and $\text{Var}(X) = (b-a)^2/12$ in order to obtain the expectation and variance of a single observation and then we use the relations $\text{E}(\bar{X}) = \text{E}(X)$ and $\text{Var}(\bar{X}) = \text{Var}(X)/n$ to translated these results to the expectation and variance of the sample average:

```
> a <- 1
> b <- 5
> n <- 75
> mu.bar <- (a+b)/2
> sig.bar <- sqrt((b-a)^2/(12*n))
> mu.bar
[1] 3
> sig.bar
[1] 0.1333333
```

After obtaining the expectation and the variance of the sample average we can forget about the Uniform distribution and proceed only with the `R` functions that are related to the Normal distribution. By the Central Limit Theorem we get that the distribution of the sample average is approximately Normal(μ, σ^2), with $\mu =$ `mu.bar` and $\sigma =$ `sig.bar`.

In the Question 1.1 we are asked to find the value of the cumulative distribution function of the sample average at $x = 2$:

```
> pnorm(2,mu.bar,sig.bar)
[1] 3.190892e-14
```

The goal of Question 1.2 is to identify the 90%-percentile of the sample average:

```
> qnorm(0.9,mu.bar,sig.bar)
[1] 3.170874
```

The sample average is equal to the total sum divided by the number of observations, $n = 75$ in this example. The total sum is less than 200 if, and only if the average is less than $200/n$. Therefore, for Question 1.3:

```
> pnorm(200/n,mu.bar,sig.bar)
[1] 0.006209665
```

Finally, if 90% of the distribution of the average is less that 3.170874 then 90% of the distribution of the total sum is less than $3.170874\,n$. In Question 1.4 we get:

```
> n*qnorm(0.9,mu.bar,sig.bar)
[1] 237.8155
```

8.3.2 Example 2

Consider again the same stress study that was described in Example 1 and answer the same questions. However, this time assume that the stress score may obtain only the values 1, 2, 3, 4 or 5, with the same likelihood for obtaining each of the values.

Solution:

Denote again by X the stress score of a random student. The modified distribution states that the sample space of X are the integers $\{1, 2, 3, 4, 5\}$, with equal probability for each value. Since the probabilities must sum to 1 we get that $\mathrm{P}(X = x) = 1/5$, for all x in the sample space. In principle we may repeat the steps of the solution of previous example, substituting the expectation and standard deviation of the continuous measurement by the discrete counterpart:

```
> x <- 1:5
> p <- rep(1/5,5)
> n <- 75
> mu.X <- sum(x*p)
> sig.X <- sum((x-mu.X)^2*p)
> mu.bar <- mu.X
> sig.bar <- sqrt(sig.X/n)
> mu.bar
[1] 3
> sig.bar
[1] 0.1632993
```

Notice that the expectation of the sample average is the same as before but the standard deviation is somewhat larger due to the larger variance in the distribution of a single response.

We may apply the Central Limit Theorem again in order to conclude that distribution of the average is approximately Normal(μ, σ^2), with $\mu =$ `mu.bar` as before and for the new $\sigma =$ `sig.bar`.

For Question 2.1 we compote that the cumulative distribution function of the sample average at $x = 2$ is approximately equal:

```
> pnorm(2,mu.bar,sig.bar)
[1] 4.570649e-10
```

and the 90%-percentile is:

```
> qnorm(0.9,mu.bar,sig.bar)
[1] 3.209276
```

which produces the answer to Question 2.2.

Similarly to the solution of Question 1.3 we may conclude that the total sum is less than 200 if, and only if the average is less than $200/n$. Therefore, for Question 2.3:

```
> pnorm(200/n,mu.bar,sig.bar)
[1] 0.02061342
```

Observe that in the current version of the question we have the score is integer-valued. Clearly, the sum of scores is also integer valued. Hence we may choose to apply the continuity correction for the Normal approximation whereby we approximate the probability that the sum is less than 200 (i.e. is less than or equal to 199) by the probability that a Normal random variable is less than or equal to 199.5. Translating this event back to the scale of the average we get the approximation[1]:

```
> pnorm(199.5/n,mu.bar,sig.bar)
[1] 0.01866821
```

Finally, if 90% of the distribution of the average is less that 3.170874 then 90% of the distribution of the total sum is less than $3.170874n$. Therefore:

```
> n*qnorm(0.9,mu.bar,sig.bar)
[1] 240.6957
```

or, after rounding to the nearest integer we get for Question 2.4 the answer 241.

8.3.3 Example 3

Suppose that a market research analyst for a cellular phone company conducts a study of their customers who exceed the time allowance included on their basic cellular phone contract. The analyst finds that for those customers who exceed the time included in their basic contract, the excess time used follows an exponential distribution with a mean of 22 minutes. Consider a random sample of 80 customers and find

1. The probability that the average excess time used by the 80 customers in the sample is longer than 20 minutes.

2. The 95th percentile for the average excess time for samples of 80 customers who exceed their basic contract time allowances.

Solution:

Let X be the excess time for customers who exceed the time included in their basic contract. We are told that $X \sim$ Exponential(λ). For the Exponential distribution $\mathrm{E}(X) = 1/\lambda$. Hence, given that $\mathrm{E}(X) = 22$ we can conclude that $\lambda = 1/22$. For the Exponential we also have that $\mathrm{Var}(X) = 1/\lambda^2$. Therefore:

```
> lam <- 1/22
> n <- 80
> mu.bar <- 1/lam
> sig.bar <- sqrt(1/(lam^2*n))
> mu.bar
[1] 22
> sig.bar
[1] 2.459675
```

[1] As a matter of fact, the continuity correction could have been applied in the previous two sections as well, since the sample average has a discrete distribution.

Like before, we can forget at this stage about the Exponential distribution and refer henceforth to the Normal Distribution. In Question 2.1 we are asked to compute the probability above $x = 20$. The total probability is 1. Hence, the required probability is the difference between 1 and the probability of being less or equal to $x = 20$:

```
> 1-pnorm(20,mu.bar,sig.bar)
[1] 0.7919241
```

The goal in Question 2.2 is to find the 95%-percentile of the sample average:

```
> qnorm(0.95,mu.bar,sig.bar)
[1] 26.04580
```

8.3.4 Example 4

A beverage company produces cans that are supposed to contain 16 ounces of beverage. Under normal production conditions the expected amount of beverage in each can is 16.0 ounces, with a standard deviation of 0.10 ounces.

As a quality control measure, each hour the QA department samples 50 cans from the production during the previous hour and measures the content in each of the cans. If the average content of the 50 cans is below a control threshold then production is stopped and the can filling machine is re-calibrated.

1. Compute the probability that the amount of beverage in a random can is below 15.95.

2. Compute the probability that the amount of beverage in a sample average of 50 cans is below 15.95.

3. Find a threshold with the property that the probability of stopping the machine in a given hour is 5% when, in fact, the production conditions are normal.

4. Consider the data in the file "`QC.csv`"[2]. It contains measurement results of 8 hours. Assume that we apply the threshold that was obtained in Question 4.3. At the end of which of the hours the filling machine needed re-calibration?

5. Based on the data in the file "`QC.csv`", which of the hours contains measurements which are suspected outliers in comparison to the other measurements conducted during that hour?

Solution

The only information we have on the distribution of each measurement is its expectation (16.0 ounces under normal conditions) and its standard deviation (0.10, under the same condition). We do not know, from the information provided in the question, the actual distribution of a measurement. (The fact that the production conditions are normal does not imply that the distribution

[2]URL for the file: `http://pluto.huji.ac.il/~msby/StatThink/Datasets/QC.csv`

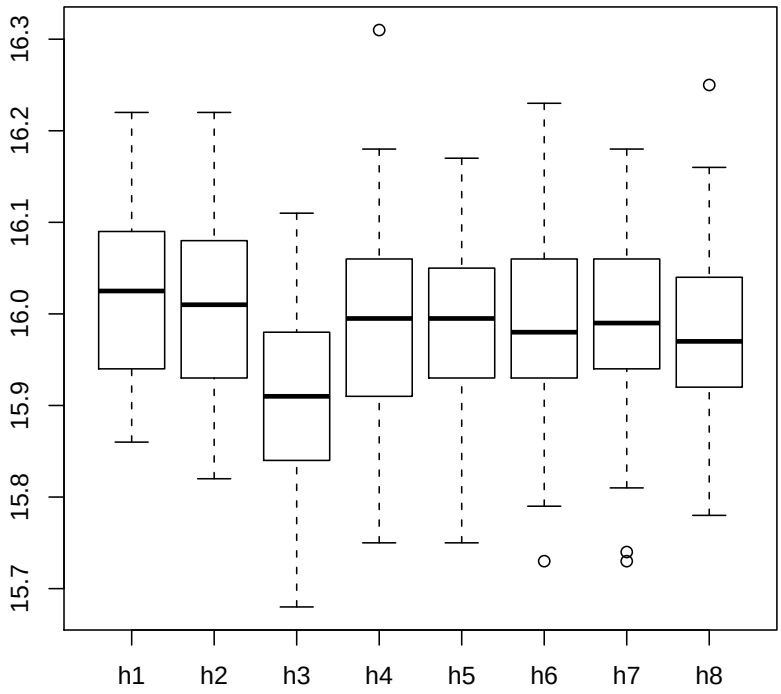

Figure 8.1: Box Plots

of the measurement in the Normal distribution!) Hence, the correct answer to Question 4.1 is that there is not enough information to calculate the probability.

When we deal with the sample average, on the other hand, we may apply the Central Limit Theorem in order to obtain at least an approximation of the probability. Observe that the expectation of the sample average is 16.0 ounces and the standard deviation is $0.1/\sqrt{50}$. The distribution of the average is approximately the Normal distribution:

```
> pnorm(15.95,16,0.1/sqrt(50))
[1] 0.000203476
```

Hence, we get that the probability of the average being less than 15.95 ounces is (approximately) 0.0002, which is a solution to Question 4.2.

In order to solve Question 4.3 we may apply the function "`qnorm`" in order to compute the 5%-percentile of the distribution of the average:

```
> qnorm(0.05,16,0.1/sqrt(50))
[1] 15.97674
```

Consider the data in the file "QC.csv". Let us read the data into a data frame by the by the name "QC" and apply the function "summary" to obtain an overview of the content of the file:

```
> QC <- read.csv("QC.csv")
> summary(QC)
       h1              h2              h3              h4
 Min.   :15.86   Min.   :15.82   Min.   :15.68   Min.   :15.75
 1st Qu.:15.94   1st Qu.:15.93   1st Qu.:15.84   1st Qu.:15.91
 Median :16.02   Median :16.01   Median :15.91   Median :15.99
 Mean   :16.02   Mean   :16.01   Mean   :15.91   Mean   :15.99
 3rd Qu.:16.09   3rd Qu.:16.08   3rd Qu.:15.98   3rd Qu.:16.06
 Max.   :16.22   Max.   :16.22   Max.   :16.11   Max.   :16.31
       h5              h6              h7              h8
 Min.   :15.75   Min.   :15.73   Min.   :15.73   Min.   :15.78
 1st Qu.:15.93   1st Qu.:15.93   1st Qu.:15.94   1st Qu.:15.92
 Median :15.99   Median :15.98   Median :15.99   Median :15.97
 Mean   :15.99   Mean   :15.98   Mean   :15.99   Mean   :15.97
 3rd Qu.:16.05   3rd Qu.:16.06   3rd Qu.:16.05   3rd Qu.:16.04
 Max.   :16.17   Max.   :16.23   Max.   :16.18   Max.   :16.25
```

Observe that the file contains 8 quantitative variables that are given the names h1, ..., h8. Each of these variables contains the 50 measurements conducted in the given hour.

Observe that the mean is computed as part of the summary. The threshold that we apply to monitor the filling machine is 15.97674. Clearly, the average of the measurements at the third hour "h3" is below the threshold. Not enough significance digits of the average of the 8th hour are presented to be able to say whether the average is below or above the threshold. A more accurate presentation of the computed mean is obtained by the application of the function "mean" directly to the data:

```
> mean(QC$h8)
[1] 15.9736
```

Now we can see that the average is below the threshold. Hence, the machine required re-calibration after the 3rd and the 8th hours, which is the answer to Question 4.4.

In Chapter 3 it was proposed to use box plots in order to identify points that are suspected to be outliers. We can use the expression "boxplot(QC$h1)" in order to obtain the box plot of the data of the first hour and go through the names of the variable one by one in order to screen all variable. Alternatively, we may apply the function "boxplot" directly to the data frame "QC" and get a plot with box plots of all the variables in the data frame plotted side by side (see Figure 8.1):

```
> boxplot(QC)
```

Examining the plots we may see that evidence for the existence of outliers can be spotted on the 4th, 6th, 7th, and 8th hours, providing an answer to Question 4.5

8.3.5 Example 5

A measurement follows the Uniform$(0, b)$, for an unknown value of b. Two statisticians propose two distinct ways to estimate the unknown quantity b with the aid of a sample of size $n = 100$. Statistician A proposes to use twice the sample average ($2\bar{X}$) as an estimate. Statistician B proposes to use the largest observation instead.

The motivation for the proposal made by Statistician A is that the expectation of the measurement is equal to $\mathrm{E}(X) = b/2$. A reasonable way to estimate the expectation is to use the sample average $\bar{X}$. Thereby, a reasonable way to estimate b, twice the expectation, is to use $2\bar{X}$. A motivation for the proposal made by Statistician B is that although the largest observation is indeed smaller that b, still it may not be much smaller than that value.

In order to choose between the two options they agreed to prefer the statistic that tends to have values that are closer to b. (with respect to the sampling distribution). They also agreed to compute the expectation and variance of each statistic. The performance of a statistic is evaluated using the *mean square error* (MSE), which is defined as the sum of the variance and the squared difference between the expectation and b. Namely, if T is the statistic (either the one proposed by Statistician A or Statistician B) then

$$MSE = \mathrm{Var}(T) + (\mathrm{E}(T) - b)^2 \ .$$

A smaller mean square error corresponds to a better, more accurate, statistic.

1. Assume that the actual value of b is 10 ($b = 10$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician A.

2. Assume that the actual value of b is 10 ($b = 10$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician B. (Hint: the maximal value of a sequence can be computed with the function "`max`".)

3. Assume that the actual value of b is 13.7 ($b = 13.7$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician A.

4. Assume that the actual value of b is 13.7 ($b = 13.7$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician B. (Hint: the maximal value of a sequence can be computed with the function "`max`".)

5. Based on the results in Questions 5.1–4, which of the two statistics seems to be preferable?

Solution

In Questions 5.1 and 5.2 we take the value of b to be equal to 10. Consequently, the distribution of a measurement is Uniform$(0, 10)$. In order to generate the sampling distributions we produce two sequences, "`A`" and "`B`", both of length 100,000, with the evaluations of the statistics:

```
> A <- rep(0,10^5)
> B <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- runif(100,0,10)
+     A[i] <- 2*mean(X.samp)
+     B[i] <- max(X.samp)
+ }
```

Observe that in each iteration of the "`for`" loop a sample of size $n = 100$ from the Uniform$(0, 10)$ distribution is generated. The statistic proposed by Statistician A ("`2*mean(X.samp)`") is computed and stored in sequence "`A`" and the statistic proposed by Statistician B ("`max(X.samp)`") is computed and stored in sequence "`B`".

Consider the statistic proposed by Statistician A:

```
> mean(A)
[1] 9.99772
> var(A)
[1] 0.3341673
> var(A) + (mean(A)-10)^2
[1] 0.3341725
```

The expectation of the statistic is 9.99772 and the variance is 0.3341673. Consequently, we get that the mean square error is equal to

$$0.3341673 + (9.99772 - 10)^2 = 0.3341725 \, .$$

Next, deal with the statistic proposed by Statistician B:

```
> mean(B)
[1] 9.901259
> var(B)
[1] 0.00950006
> var(B) + (mean(B)-10)^2
[1] 0.01924989
```

The expectation of the statistic is 9.901259 and the variance is 0.00950006. Consequently, we get that the mean square error is equal to

$$0.00950006 + (9.901259 - 10)^2 = 0.01924989 \, .$$

Observe that the mean square error of the statistic proposed by Statistician B is smaller.

For Questions 5.3 and 5.4 we run the same type of simulations. All we change is the value of b (from 10 to 13.7):

```
> A <- rep(0,10^5)
> B <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- runif(100,0,13.7)
+     A[i] <- 2*mean(X.samp)
+     B[i] <- max(X.samp)
+ }
```

Again, considering the statistic proposed by Statistician A we get:

```
> mean(A)
[1] 13.70009
> var(A)
[1] 0.6264204
> var(A) + (mean(A)-13.7)^2
[1] 0.6264204
```

The expectation of the statistic in this setting is 13.70009 and the variance is 0.6264204. Consequently, we get that the mean square error is equal to

$$0.6264204 + (13.70009 - 13.7)^2 = 0.6264204 \ .$$

For the statistic proposed by Statistician B we obtain:

```
> mean(B)
[1] 13.56467
> var(B)
[1] 0.01787562
> var(B) + (mean(B)-13.7)^2
[1] 0.03618937
```

The expectation of the statistic is 13.56467 and the variance is 0.01787562. Consequently, we get that the mean square error is equal to

$$0.01787562 + (13.56467 - 13.7)^2 = 0.03618937 \ .$$

Once more, the mean square error of the statistic proposed by Statistician B is smaller.

Considering the fact that the mean square error of the statistic proposed by Statistician B is smaller in both cases we may conclude that this statistic seems to be better for estimation of b in this setting of Uniformly distributed measurements[3].

Discussion in the Forum

In this course we have learned many subjects. Most of these subjects, especially for those that had no previous exposure to statistics, were unfamiliar. In this forum we would like to ask you to share with us the difficulties that you encountered.

What was the topic that was most difficult for you to grasp? In your opinion, what was the source of the difficulty?

When forming your answer to this question we will appreciate if you could elaborate and give details of what the problem was. Pointing to deficiencies in the learning material and confusing explanations will help us improve the presentation for the future application of this course.

[3] As a matter of fact, it can be proved that the statistic proposed by Statistician B has a smaller mean square error than the statistic proposed by Statistician A, for *any* value of b