Chapter 11

Confidence Intervals

11.1 Student Learning Objectives

A confidence interval is an estimate of an unknown parameter by a range of values. This range contains the value of the parameter with a prescribed probability, called the confidence level. In this chapter we discuss the construction of confidence intervals for the expectation and for the variance of a measurement as well as for the probability of an event. In some cases the construction will apply the Normal approximation suggested by the Central Limit Theorem. This approximation is valid when the sample size is large enough. The construction of confidence intervals for a small sample is considered in the context of Normal measurements. By the end of this chapter, the student should be able to:

- Define confidence intervals and confidence levels.
- Construct a confidence interval for the expectation of a measurement and for the probability of an event.
- Construct a confidence interval for expectation and for the variance of a Normal measurement.
- Compute the sample size that will produce a confidence interval of a given width.

11.2 Intervals for Mean and Proportion

A confidence interval, like a point estimator, is a method for estimating the unknown value of a parameter. However, instead of producing a single number, the confidence interval is an interval of numbers. The interval of values is calculated from the data. The confidence interval is likely to include the unknown population parameter. The probability of the event of inclusion is denoted as the *confidence level* of the confidence intervals.

This section presents a method for the computation of confidence intervals for the expectation of a measurement and a similar method for the computation of a confidence interval for the probability of an event. These methods rely on the application of the Central Limit Theorem to the sample average in the one case, and to the sample proportion in the other case.

In the first subsection we compute a confidence interval for the expectation of the variable "price" and a confidence interval for the proportion of diesel cars. The confidence intervals are computed based on the data in the file "cars.csv". In the subsequent subsections we discuss the theory behind the computation of the confidence intervals and explain the meaning of the confidence level. Subsection 11.2.2 does so with respect to the confidence interval for the expectation and Subsection 11.2.3 with respect to the confidence interval for the proportion.

11.2.1 Examples of Confidence Intervals

A point estimator of the expectation of a measurement is the sample average of the variable that is associated with the measurement. A confidence interval is an interval of numbers that is likely to contain the parameter value. A natural interval to consider is an interval centered at the sample average $\bar{x}$. The interval is set to have a width that assures the inclusion of the parameter value in the prescribed probability, namely the confidence level.

Consider the confidence interval for the expectation. The structure of the confidence interval of confidence level 95% is $[\bar{x} - 1.96 \cdot s/\sqrt{n}, \bar{x} + 1.96 \cdot s/\sqrt{n}]$, where s is the estimated standard deviation of the measurement (namely, the sample standard deviation) and n is the sample size. This interval may be expressed in the form:

$$\bar{x} \pm 1.96 \cdot s/\sqrt{n} \, .$$

As an illustration, let us construct a 0.95-confidence interval for the expected price of a car. :

```
> cars <- read.csv("cars.csv")
> x.bar <- mean(cars$price,na.rm=TRUE)
> s <- sd(cars$price,na.rm=TRUE)
> n <- 201
```

In the first line of code the data in the file "cars.csv" is stored in a data frame called "cars". In the second line the average $\bar{x}$ is computed for the variable "price" in the data frame "cars". This average is stored under the name "x.bar". Recall that the variable "price" contains 4 missing values. Hence, in order to compute the average of the non-missing values we should set a "TRUE" value to the argument "na.rm". The sample standard deviation "s" is computed in the third line by the application of the function "sd". We set once more the argument "na.rm=TRUE" in order to deal with the missing values. Finally, in the last line we store the sample size "n", the number of non-missing values.

Let us compute the lower and the upper limits of the confidence interval for the expectation of the price:

```
> x.bar - 1.96*s/sqrt(n)
[1] 12108.47
> x.bar + 1.96*s/sqrt(n)
[1] 14305.79
```

The lower limit of the confidence interval turns out to be $12,108.47 and the upper limit is $14,305.79. The confidence interval is the range of values between these two numbers.

Consider, next, a confidence interval for the probability of an event. The estimate of the probability p is $\hat{p}$, the relative proportion of occurrences of the event in the sample. Again, we construct an interval about this estimate. In this case, a confidence interval of confidence level 95% is of the form $\left[\hat{p} - 1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n}\right]$, where n is the sample size. Observe that $\hat{p}$ replaces $\bar{x}$ as the estimate of the parameter and that $\hat{p}(1-\hat{p})/n$ replace s^2/n as the estimate of the variance of the estimator. The confidence interval for the probability may be expressed in the form:

$$\hat{p} \pm 1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n} \ .$$

As an example, let us construct a confidence interval for the proportion of car types that use diesel fuel. The variable "`fuel.type`" is a factor that records the type of fuel the car uses, either diesel or gas:

```
> table(cars$fuel.type)

diesel    gas
    20    185
```

Only 20 of the 205 types of cars are run on diesel in this data set. The point estimation of the probability of such car types and the confidence interval for this probability are:

```
> n <- 205
> p.hat <- 20/n
> p.hat
[1] 0.09756098
> p.hat - 1.96*sqrt(p.hat*(1-p.hat)/n)
[1] 0.05694226
> p.hat + 1.96*sqrt(p.hat*(1-p.hat)/n)
[1] 0.1381797
```

The point estimation of the probability is $\hat{p} = 20/205 \approx 0.098$ and the confidence interval, after rounding up, is $[0.057, 0.138]$.

11.2.2 Confidence Intervals for the Mean

In the previous subsection we computed a confidence interval for the expected price of a car and a confidence interval for the probability that a car runs on diesel. In this subsection we explain the theory behind the construction of confidence intervals for the expectation. The theory provides insight to the way confidence intervals should be interpreted. In the next subsection we will discuss the theory behind the construction of confidence intervals for the probability of an event.

Assume one is interested in a confidence interval for the expectation of a measurement X. For a sample of size n, one may compute the sample average $\bar{X}$, which is the point estimator for the expectation. The expected value of the sample average is the expectation $\mathrm{E}(X)$, for which we are trying to produce the confidence interval. Moreover, the variance of the sample average is $\mathrm{Var}(X)/n$, where $\mathrm{Var}(X)$ is the variance of a single measurement and n is the sample size.

The construction of a confidence interval for the expectation relies on the Central Limit Theorem and on estimation of the variance of the measurement. The Central Limit Theorem states that the distribution of the (standardized) sample average $Z = (\bar{X} - \mathrm{E}(X)/\sqrt{\mathrm{Var}(X)/n}$ is approximately standard Normal for a large enough sample size. The variance of the measurement can be estimated using the sample variance S^2.

Supposed that we are interested in a confidence interval with a confidence level of 95%. The value 1.96 is the 0.975-percentile of the standard Normal. Therefore, about 95% of the distribution of the standardized sample average is concentrated in the range $[-1.96, 1, 96]$:

$$\mathrm{P}\left(\left|\frac{\bar{X} - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)/n}}\right| \leq 1.96\right) \approx 0.95$$

The event, the probability of which is being described in the last display, states that the absolute value of deviation of the sample average from the expectation, divided by the standard deviation of the sample average, is no more than 1.96. In other words, the distance between the sample average and the expectation is at most 1.96 units of standard deviation. One may rewrite this event in a form that puts the expectation within an interval that is centered at the sample average[1]:

$$\left\{|\bar{X} - \mathrm{E}(X)| \leq 1.96 \cdot \sqrt{\mathrm{Var}(X)/n}\right\} \quad \Longleftrightarrow$$
$$\left\{\bar{X} - 1.96 \cdot \sqrt{\mathrm{Var}(X)/n} \leq \mathrm{E}(X) \leq \bar{X} + 1.96 \cdot \sqrt{\mathrm{Var}(X)/n}\right\}.$$

Clearly, the probability of the later event is (approximately) 0.95 since we are considering the same event, each time represented in a different form. The second representation states that the expectation $\mathrm{E}(X)$ belongs to an interval about the sample average: $\bar{X} \pm 1.96\sqrt{\mathrm{Var}(X)/n}$. This interval is, almost, the confidence interval we seek.

The difficulty is that we do not know the value of the variance $\mathrm{Var}(X)$, hence we cannot compute the interval in the proposed form from the data. In order to overcome this difficulty we recall that the unknown variance may nonetheless be estimated from the data:

$$S^2 \approx \mathrm{Var}(X) \quad \Longrightarrow \quad \sqrt{\mathrm{Var}(X)/n} \approx S/\sqrt{n}\,,$$

where S is the sample standard deviation[2].

When the sample size is sufficiently large, so that S is very close to the value of the standard deviation of an observation, we obtain that the interval $\bar{X} \pm 1.96\sqrt{\mathrm{Var}(X)/n}$ and the interval $\bar{X} \pm 1.96 \cdot S/\sqrt{n}$ almost coincide. Therefore:

$$\mathrm{P}\left(\bar{X} - 1.96 \cdot \frac{S}{\sqrt{n}} \leq \mathrm{E}(X) \leq \bar{X} + 1.96 \cdot \frac{S}{\sqrt{n}}\right) \approx 0.95\,.$$

[1] Observe that $|\bar{X} - \mathrm{E}(X)| = |\mathrm{E}(X) - \bar{X}|$ and therefore $\left\{|\bar{X} - \mathrm{E}(X)| \leq 1.96 \cdot \sqrt{\mathrm{Var}(X)/n}\right\} = \left\{|\mathrm{E}(X) - \bar{X}| \leq 1.96 \cdot \sqrt{\mathrm{Var}(X)/n}\right\}$. From the definition of the absolute value we obtain that the last expression is equal to $\left\{-1.96 \cdot \sqrt{\mathrm{Var}(X)/n} \leq \mathrm{E}(X) - \bar{X} \leq 1.96 \cdot \sqrt{\mathrm{Var}(X)/n}\right\}$. Moving the average to the other side of the inequality (for both inequalities involved) produces the representation $\left\{\bar{X} - 1.96 \cdot \sqrt{\mathrm{Var}(X)/n} \leq \mathrm{E}(X) \leq \bar{X} + 1.96 \cdot \sqrt{\mathrm{Var}(X)/n}\right\}$.

[2] The sample variance, that serves as the estimator of the variance, is computed from the sample using the formula: $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$.

Hence, $\bar{X} \pm 1.96 \cdot S/\sqrt{n}$ is an (approximate) confidence interval of the (approximate) confidence level 0.95.

Let us demonstrate the issue of confidence level by running a simulation. We are interested in a confidence interval for the expected price of a car. In the simulation we assume that the distribution of the price is Exponential(1/13000). (Consequently, $\mathrm{E}(X) = 13,000$). We take the sample size to be equal to $n = 201$ and compute the actual probability of the confidence interval containing the value of the expectation:

```
> lam <- 1/13000
> n <- 201
> X.bar <- rep(0,10^5)
> S <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rexp(n,lam)
+    X.bar[i] <- mean(X)
+    S[i] <- sd(X)
+ }
> LCL <- X.bar - 1.96*S/sqrt(n)
> UCL <- X.bar + 1.96*S/sqrt(n)
> mean((13000 >= LCL) & (13000 <= UCL))
[1] 0.94518
```

Below we will go over the code and explain the simulation. But, before doing so, notice that the actual probability that the confidence interval contains the expectation is about 0.945, which is slightly below the nominal confidence level of 0.95. Still quoting the nominal value as the confidence level of the confidence interval is not too far from reality.

Let us look now at the code that produced the simulation. In each iteration of the simulation a sample is generated. The sample average and standard deviations are computed and stored in the appropriate locations of the sequences "`X.bar`" and "`S`". At the end of all the iterations the content of these two sequences represents the sampling distribution of the sample average $\bar{X}$ and the sample standard deviation S, respectively.

The lower and the upper end-points of the confidence interval are computed in the next two lines of code. The lower level of the confidence interval is stored in the object "`LCL`" and the upper level is stored in "`UCL`". Consequently, we obtain the sampling distribution of the confidence interval. This distribution is approximated by 100,000 random confidence intervals that are generated by the sampling distribution. Some of these random intervals contain the value of the expectation, namely 13,000, and some do not. The proportion of intervals that contain the expectation is the (simulated) confidence level. The last expression produces this confidence level, which turns out to be equal to about 0.945.

The last expression involves a new element, the term "`&`", which calls for more explanations. Indeed, let us refer to the last expression in the code. This expression involves the application of the function "`mean`". The input to this function contains two sequences with logical values ("`TRUE`" or "`FALSE`"), separated by the character "`&`". The character "`&`" corresponds to the logical "AND" operator. This operator produces a "`TRUE`" if a "`TRUE`" appears at both sides. Otherwise, it produces a "`FALSE`". (Compare this operator to the operator

"OR", that is expressed in R with the character "|", that produces a "TRUE" if at least one "TRUE" appears at either sides.)

In order to clarify the behavior of the terms "&" and "|" consider the following example:

```
> a <- c(TRUE, TRUE, FALSE, FALSE)
> b <- c(FALSE, TRUE, TRUE, FALSE)
> a & b
[1] FALSE  TRUE FALSE FALSE
> a | b
[1] TRUE TRUE TRUE FALSE
```

The term "&" produces a "TRUE" only if parallel components in the sequences "a" and "b" both obtain the value "TRUE". On the other hand, the term "|" produces a "TRUE" if at least one of the parallel components are "TRUE". Observe, also, that the output of the expression that puts either of the two terms between two sequences with logical values is a sequence of the same length (with logical components as well).

The expression "(13000 >= LCL)" produces a logical sequence of length 100,000 with "TRUE" appearing whenever the expectation is larger than the lower level of the confidence interval. Similarly, the expression "(13000 <= UCL)" produces "TRUE" values whenever the expectation is less than the upper level of the confidence interval. The expectation belongs to the confidence interval if the value in both expressions is "TRUE". Thus, the application of the term "&" to these two sequences identifies the confidence intervals that contain the expectation. The application of the function "mean" to a logical vector produces the relative frequency of TRUE's in the vector. In our case this corresponds to the relative frequency of confidence intervals that contain the expectation, namely the confidence level.

We calculated before the confidence interval $[12108.47, 14305.79]$ for the expected price of a car. This confidence interval was obtained via the application of the formula for the construction of confidence intervals with a 95% confidence level to the variable "price" in the data frame "cars". Casually speaking, people frequently refer to such an interval as an interval that contains the expectation with probability of 95%.

However, one should be careful when interpreting the confidence level as a probabilistic statement. The probability computations that led to the method for constructing confidence intervals were carried out in the context of the sampling distribution. Therefore, probability should be interpreted in the context of all data sets that could have emerged and not in the context of the given data set. No probability is assigned to the statement "The expectation belongs to the interval $[12108.47, 14305.79]$". The probability is assigned to the statement "The expectation belongs to the interval $\bar{X} \pm 1.96 \cdot S/\sqrt{n}$", where $\bar{X}$ and S are interpreted as random variables. Therefore the statement that the interval $[12108.47, 14305.79]$ contains the expectation with probability of 95% is meaningless. What is meaningful is the statement that the given interval was constructed using a procedure that produces, when applied to random samples, intervals that contain the expectation with the assigned probability.

11.2.3 Confidence Intervals for a Proportion

The next issue is the construction of a confidence interval for the probability of an event. Recall that a probability p of some event can be estimated by the observed relative frequency of the event in the sample, denoted $\hat{P}$. The estimation is associated with the Bernoulli random variable X, that obtains the value 1 when the event occurs and the value 0 when it does not. In the estimation problem p is the expectation of X and $\hat{P}$ is the sample average of this measurement. With this formulation we may relate the problem of the construction of a confidence interval for p to the problem of constructing a confidence interval for the expectation of a measurement. The latter problem was dealt with in the previous subsection.

Specifically, the discussion regarding the steps in the construction – staring with an application of the Central Limit Theorem in order to produce an interval that depends on the sample average and its variance and proceeding by the replacement of the unknown variance by its estimate – still apply and may be taken as is. However, in the specific case we have a particular expression for the variance of the estimate $\hat{P}$:

$$\mathrm{Var}(\hat{P}) = p(1-p)/n \approx \hat{P}(1-\hat{P})/n \; .$$

The tradition is to estimate this variance by using the estimator $\hat{P}$ for the unknown p instead of using the sample variance. The resulting confidence interval of significance level 0.95 takes the form:

$$\bar{P} \pm 1.96 \cdot \sqrt{\hat{P}(1-\hat{P})/n} \; .$$

Let us run a simulation in order to assess the confidence level of the confidence interval for the probability. Assume that $n = 205$ and $p = 0.12$. The simulation we run is very similar to the simulation of Subsection 11.2.2. In the first stage we produce the sampling distribution of $\hat{P}$ (stored in the sequence "`P.hat`") and in the second stage we compute the relative frequency in the simulation of the intervals that contain the actual value of p that was used in the simulation:

```
> p <- 0.12
> n <- 205
> P.hat <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rbinom(n,1,p)
+    P.hat[i] <- mean(X)
+ }
> LCL <- P.hat - 1.96*sqrt(P.hat*(1-P.hat)/n)
> UCL <- P.hat + 1.96*sqrt(P.hat*(1-P.hat)/n)
> mean((p >= LCL) & (p <= UCL))
[1] 0.95131
```

In this simulation we obtained that the actual confidence level is approximately 0.951, which is slightly above the nominal confidence level of 0.95.

The formula $\bar{X} \pm 1.96 \cdot S/\sqrt{n}$ that is used for a confidence interval for the expectation and the formula $\hat{P} \pm 1.96 \cdot \{\hat{P}(1-\hat{P})/n\}^{1/2}$ for the probability both

refer to a confidence intervals with confidence level of 95%. If one is interested in a different confidence level then the width of the confidence interval should be adjusted: a wider interval for higher confidence and a narrower interval for smaller confidence level.

Specifically, if we examine the derivation of the formulae for confidence intervals we may notice that the confidence level is used to select the number 1.96, which is the 0.975-percentile of the standard Normal distribution (1.96 =`qnorm(0.975)`). The selected number satisfies that the interval $[-1.96, 1.96]$ contains 95% of the standard Normal distribution by leaving out 2.5% on both tails. For a different confidence level the number 1.96 should be replace by a different number.

For example, if one is interested in a 90% confidence level then one should use 1.645, which is the 0.95-percentile of the standard Normal distribution (`qnorm(0.95)`), leaving out 5% in both tails. The resulting confidence interval for an expectation is $\bar{X} \pm 1.645 \cdot S/\sqrt{n}$ and the confidence interval for a probability is $\hat{P} \pm 1.645 \cdot \{\hat{P}(1-\hat{P})/n\}^{1/2}$.

11.3 Intervals for Normal Measurements

In the construction of the confidence intervals in the previous section it was assumed that the sample size is large enough. This assumption was used both in the application of the Central Limit Theorem and in the substitution of the unknown variance by its estimated value. For a small sample size the reasoning that was applied before may no longer be valid. The Normal distribution may not be a good enough approximation of the sampling distribution of the sample average and the sample variance may differ substantially from the actual value of the measurement variance.

In general, making inference based on small samples requires more detailed modeling of the distribution of the measurements. In this section we will make the assumption that the distribution of the measurements is Normal. This assumption may not fit all scenarios. For example, the Normal distribution is a poor model for the price of a car, which is better modeled by the Exponential distribution. Hence, a blind application of the methods developed in this section to variables such as the price when the sample size is small may produce dubious outcomes and is not recommended.

When the distribution of the measurements is Normal then the method discussed in this section will produce valid confidence intervals for the expectation of the measurement even for a small sample size. Furthermore, we will extend the methodology to enable the construction of confidence intervals for the variance of the measurement.

Before going into the details of the methods let us present an example of inference that involves a small sample. Consider the issue of fuel consumption. Two variables in the "`cars`" data frame describe the fuel consumption. The first, "`city.mpg`", reports the number of miles per gallon when the car is driven in urban conditions and the second, "`highway.mpg`", reports the miles per gallon in highway conditions. Typically, driving in city conditions requires more stopping and change of speed and is less efficient in terms of fuel consumption. Hence, one expects to obtained a reduced number of miles per gallon when driving in urban conditions compared to the number when driving in highway conditions.

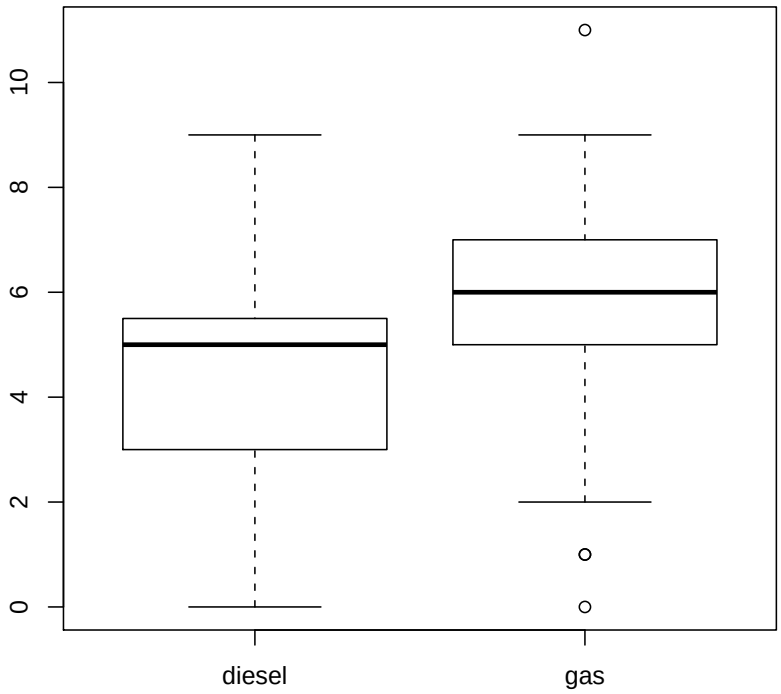

Figure 11.1: Box Plots of Differences in MPG

For each car type we calculate the difference variable that measures the difference between the number of miles per gallon in highway conditions and the number in urban conditions. The cars are sub-divided between cars that run on diesel and cars that run on gas. Our concern is to estimate, for each fuel type, the expectation of difference variable and to estimate the variance of that variable. In particular, we are interested in the construction of a confidence intervals for the expectation and a confidence interval for the variance.

Box plots of the difference in fuel consumption between highway and urban conditions are presented in Figure 11.1. The box plot on the left hand side corresponds to cars that run on diesel and the box plot on the right hand side corresponds to cars that run on gas. Recall that 20 of the 205 car types use diesel and the other 185 car types use gas. One may suspect that the fuel consumption characteristics vary between the two types of fuel. Indeed, the measurement tends to have sightly higher values for vehicles that use gas.

We conduct inference for each fuel type separately. However, since the sample size for cars that run on diesel is only 20, one may have concerns regarding the application of methods that assume a large sample size to a sample size this small.

11.3.1 Confidence Intervals for a Normal Mean

Consider the construction of a confidence interval for the expectation of a Normal measurement. In the previous section, when dealing with the construction of a confidence interval for the expectation, we exploited the Central Limit Theorem in order to identify that the distribution of the standardized sample average $(\bar{X} - \mathrm{E}(X))/\sqrt{\mathrm{Var}(X)/n}$ is, approximately, standard Normal. Afterwards, we substituted the standard deviation of the measurement by the sample standard deviation S, which was an accurate estimator of the former due to the magnitude sample size.

In the case where the measurements themselves are Normally distributed one can identify the exact distribution of the standardized sample average, with the sample variance substituting the variance of the measurement: $(\bar{X} - \mathrm{E}(X))/(S/\sqrt{n})$. This specific distribution is called the Student's t-distribution, or simply the t-distribution.

The t-distribution is bell shaped and symmetric. Overall, it looks like the standard Normal distribution but it has wider tails. The t-distribution is characterized by a parameter called the number of *degrees of freedom*. In the current setting, where we deal with the standardized sample average (with the sample variance substituting the variance of the measurement) the number of degrees of freedom equals the number of observations associated with the estimation of the variance, minus 1. Hence, if the sample size is n and if the measurement is Normally distributed then the standardized sample average (with S substituting the standard deviation of the measurement) has a t-distribution on $(n-1)$ degrees of freedom. We use $t_{(n-1)}$ to denote this t-distribution.

The `R` system contains functions for the computation of the density, the cumulative probability function and the percentiles of the t-distribution, as well as for the simulation of a random sample from this distribution. Specifically, the function "`qt`" computes the percentiles of the t-distribution. The first argument to the function is a probability and the second argument is the number of degrees of freedom. The output of the function is the percentile associated with the probability of the first argument. Namely, it is a value such that the probability that the t-distribution is below the value is equal to the probability in the first argument.

For example, let "`n`" be the sample size. The output of the expression "`qt(0.975,n-1)`" is the 0.975-percentile of the t-distribution on $(n-1)$ degrees of freedom. By definition, 97.5% of the t-distribution are below this value and 2.5% are above it. The symmetry of the t distribution implies that 2.5% of the distribution is below the negative of this value. The middle part of the distribution is bracketed by these two values: $[-$`qt(0.975,n-1)`$,$`qt(0.975,n-1)`$]$, and it contains 95% of the distribution.

Summarizing the above claims in a single formula produces the statement:

$$\frac{\bar{X} - \mathrm{E}(X)}{S/\sqrt{n}} \sim t_{(n-1)} \quad \Longrightarrow \quad \mathrm{P}\left(\left| \frac{\bar{X} - \mathrm{E}(X)}{S/\sqrt{n}} \right| \leq \texttt{qt(0.975,n-1)} \right) = 0.95 \; .$$

Notice that the equation associated with the probability is not an approximation but an exact relation[3]. Rewriting the event that is described in the probability

[3] When the measurement is Normally distributed.

in the form of a confidence interval, produces

$$\bar{X} \pm \texttt{qt(0.975,n-1)} \cdot S/\sqrt{n}$$

as a confidence interval for the expectation of the Normal measurement with a confidence level of 95%.

The structure of the confidence interval for the expectation of a Normal measurement is essentially identical to the structure proposed in the previous section. The only difference is that the number 1.96, the percentile of the standard Normal distribution, is substituted by the percentile of the t-distribution.

Consider the construction of a confidence interval for the expected difference in fuel consumption between highway and urban driving conditions. In order to save writing we created two new variables; a factor called "`fuel`" that contains the data on the fuel type of each car, and a numerical vector called "`dif.mpg`" that contains the difference between highway and city fuel consumption for each car type:

```
> fuel <- cars$fuel.type
> dif.mpg <- cars$highway.mpg - cars$city.mpg
```

We are interested in confidence intervals based on the data stored in the variable "`dif.mpg`". One confidence interval will be associated with the level "`diesel`" of the factor "`fuel`" and the other will be associated with the level "`gas`" of the same factor.

In order to compute these confidence intervals we need to compute, for each level of the factor "`fuel`", the sample average and the sample standard deviation of the data points of the variable "`dif.mpg`" that are associated with that level.

It is convenient to use the function "`tapply`" for this task. This function uses three arguments. The first argument is the sequence of values over which we want to carry out some computation. The second argument is a factor. The third argument is a name of a function that is used for the computation. The function "`tapply`" applies the function in the third argument to each sub-collection of values of the first argument. The sub-collections are determined by the levels of the second argument.

Sounds complex but it is straightforward enough to apply:

```
> tapply(dif.mpg,fuel,mean)
  diesel      gas
4.450000 5.648649
> tapply(dif.mpg,fuel,sd)
  diesel      gas
2.781045 1.433607
```

Sample averages are computed in the first application of the function "`tapply`". Observe that an average was computed for cars that run on diesel and an average was computed for cars that run on gas. In both cases the average corresponds to the difference in fuel consumption. Similarly, the standard deviations were computed in the second application of the function. We obtain that the point estimates of the expectation for diesel and gas cars are 4.45 and 5.648649, respectively and the point estimates for the standard deviation of the variable are 2.781045 and 1.433607.

Let us compute the confidence interval for each type of fuel:

```
> x.bar <- tapply(dif.mpg,fuel,mean)
> s <- tapply(dif.mpg,fuel,sd)
> n <- c(20,185)
> x.bar - qt(0.975,n-1)*s/sqrt(n)
  diesel      gas
3.148431 5.440699
> x.bar + qt(0.975,n-1)*s/sqrt(n)
  diesel      gas
5.751569 5.856598
```

The objects "`x.bar`" and "`s`" contain the sample averages and sample standard deviations, respectively. Both are sequences of length two, with the first component referring to "`diesel`" and the second component referring to "`gas`". The object "`n`" contains the two sample sizes, 20 for "`diesel`" and 185 for "`gas`". In the expression next to last the lower boundary for each of the confidence intervals is computed and in the last expression the upper boundary is computed. The confidence interval for the expected difference in diesel cars is $[3.148431, 5.751569]$. and the confidence interval for cars using gas is $[5.440699, 5.856598]$.

The 0.975-percentiles of the t-distributions are computed with the expressions "`qt(0.025,n-1)`":

```
> qt(0.975,n-1)
[1] 2.093024 1.972941
```

The second argument of the function "`qt`" is a sequence with two components, the number 19 and the number 184. Accordingly, The first position in the output of the function is the percentile associated with 19 degrees of freedom and the second position is the percentile associated to 184 degrees of freedom.

Compare the resulting percentiles to the 0.975-percentile of the standard Normal distribution, which is essentially equal to 1.96. When the sample size is small, 20 for example, the percentile of the t-distribution is noticeably larger than the percentile of the standard Normal. However, for a larger sample size the percentiles, more or less, coincide. It follows that for a large sample the method proposed in Subsection 11.2.2 and the method discussed in this subsection produce essentially the same confidence intervals.

11.3.2 Confidence Intervals for a Normal Variance

The next task is to compute confidence intervals for the variance of a Normal measurement. The main idea in the construction of a confidence interval is to identify the distribution of a random variable associated with the parameter of interest. A region that contains 95% of the distribution of the random variable (or, more generally, the central part of the distribution of probability equal to the confidence level) is identified. The confidence interval results from the reformulation of the event associated with that region. The new formulation puts the parameter between a lower limit and an upper limit. These lower and the upper limits are computed from the data and they form the boundaries of the confidence interval.

We start with the sample variance, $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, which serves as a point estimator of the parameter of interest. When the measurements are

Normally distributed then the random variable $(n-1)S^2/\sigma^2$ possesses a special distribution called the chi-square distribution. (Chi is the Greek letter χ, which is read "Kai".) This distribution is associated with the sum of squares of Normal variables. It is parameterized, just like the t-distribution, by a parameter called the number of degrees of freedom. This number is equal to $(n-1)$ in the situation we discuss. The chi-square distribution on $(n-1)$ degrees of freedom is denoted with the symbol $\chi^2_{(n-1)}$.

The `R` system contains functions for the computation of the density, the cumulative probability function and the percentiles of the chi-square distribution, as well as for the simulation of a random sample from this distribution. Specifically, the percentiles of the chi-square distribution are computed with the aid of the function "`qchisq`". The first argument to the function is a probability and the second argument is the number of degrees of freedom. The output of the function is the percentile associated with the probability of the first argument. Namely, it is a value such that the probability that the chi-square distribution is below the value is equal to the probability in the first argument.

For example, let "`n`" be the sample size. The output of the expression "`qt(0.975,n-1)`" is the 0.975-percentile of the chi-square distribution. By definition, 97.5% of the chi-square distribution are below this value and 2.5% are above it. Similarly, the expression "`qchisq(0.025,n-1)`" is the 0.025-percentile of the chi-square distribution, with 2.5% of the distribution below this value. Notice that between these two percentiles, namely within the interval [`qchisq(0.025,n-1)`, `qchisq(0.975,n-1)`], are 95% of the chi-square distribution.

We may summarize that for Normal measurements:

$$(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)} \quad \Longrightarrow$$
$$\mathrm{P}\big(\texttt{qchisq(0.025,n-1)} \leq (n-1)S^2/\sigma^2 \leq \texttt{qchisq(0.975,n-1)}\big) = 0.95\ .$$

The chi-square distribution is not symmetric. Therefore, in order to identify the region that contains 95% of the distribution region we have to compute both the 0.025- and the 0.975-percentiles of the distribution.

The event associated with the 95% region is rewritten in a form that puts the parameter σ^2 in the center:

$$\big\{(n-1)S^2/\texttt{qchisq(0.975,n-1)} \leq \sigma^2 \leq (n-1)S^2/\texttt{qchisq(0.025,n-1)}\big\}\ .$$

The left most and the right most expressions in this event mark the end points of the confidence interval. The structure of the confidence interval is:

$$\big[\{(n-1)/\texttt{qchisq(0.975,n-1)}\} \times S^2,\ \{(n-1)/\texttt{qchisq(0.025,n-1)}\} \times S^2\big]\ .$$

Consequently, the confidence interval is obtained by the multiplication of the estimator of the variance by a ratio between the number of degrees of freedom $(n-1)$ and an appropriate percentile of the chi-square distribution. The percentile on the left hand side is associated with the larger probability (making the ratio smaller) and the percentile on the right hand side is associated with the smaller probability (making the ratio larger).

Consider, specifically, the confidence intervals for the variance of the measurement "`diff.mpg`" for cars that run on diesel and for cars that run on gas. Here, the size of the samples is 20 and 185, respectively:

```
> (n-1)/qchisq(0.975,n-1)
[1] 0.5783456 0.8234295
> (n-1)/qchisq(0.025,n-1)
[1] 2.133270 1.240478
```

The ratios that are used in the left hand side of the intervals are 0.5783456 and 0.8234295, respectively. Both ratios are less than one. On the other hand, the ratios associated with the other end of the intervals, 2.133270 and 1.240478, are both larger than one.

Let us compute the point estimates of the variance and the associated confidence intervals. Recall that the object "s" contains the sample standard deviations of the difference in fuel consumption for diesel and for gas cars. The object "n" contains the two sample sizes:

```
> s^2
  diesel      gas
7.734211 2.055229
> (n-1)*s^2/qchisq(0.975,n-1)
  diesel      gas
4.473047 1.692336
> (n-1)*s^2/qchisq(0.025,n-1)
   diesel       gas
16.499155  2.549466
```

The variance of the difference in fuel consumption for diesel cars is estimated to be 7.734211 with a 95%-confidence interval of $[4.473047, 16.499155]$ and for cars that use gas the estimated variance is 2.055229, with a confidence interval of $[1.692336, 2.549466]$.

As a final example in this section let us simulate the confidence level for a confidence interval for the expectation and for a confidence interval for the variance of a Normal measurement. In this simulation we assume that the expectation is equal to $\mu = 3$ and the variance is equal to $\sigma^2 = 3^2 = 9$. The sample size is taken to be $n = 20$. We start by producing the sampling distribution of the sample average $\bar{X}$ and of the sample standard deviation S:

```
> mu <- 4
> sig <- 3
> n <- 20
> X.bar <- rep(0,10^5)
> S <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rnorm(n,mu,sig)
+    X.bar[i] <- mean(X)
+    S[i] <- sd(X)
+ }
```

Consider first the confidence interval for the expectation:

```
> mu.LCL <- X.bar - qt(0.975,n-1)*S/sqrt(n)
> mu.UCL <- X.bar + qt(0.975,n-1)*S/sqrt(n)
> mean((mu >= mu.LCL) & (mu <= mu.UCL))
[1] 0.95033
```

The nominal significance level of the confidence interval is 95%, which is practically identical to the confidence level that was computed in the simulation.

The confidence interval for the variance is obtained in a similar way. The only difference is that we apply now different formulae for the computation of the upper and lower confidence limits:

```
> var.LCL <- (n-1)*S^2/qchisq(0.975,n-1)
> var.UCL <- (n-1)*S^2/qchisq(0.025,n-1)
> mean((sig^2 >= var.LCL) & (sig^2 <= var.UCL))
[1] 0.94958
```

Again, we obtain that the nominal confidence level of 95% coincides with the confidence level computed in the simulation.

11.4 Choosing the Sample Size

One of the more important contributions of Statistics to research is providing guidelines for the design of experiments and surveys. A well planed experiment may produce accurate enough answers to the research questions while optimizing the use of resources. On the other hand, poorly planed trials may fail to produce such answers or may waste valuable resources.

Unfortunately, in this book we do not cover the subject of experiment design. Still, we would like to give a brief discussion of a narrow aspect in design: The selection of the sample size.

An important consideration at the stage of the planning of an experiment or a survey is the number of observations that should be collected. Indeed, having a larger sample size is usually preferable from the statistical point of view. However, an increase in the sample size typically involves an increase in expenses. Thereby, one would prefer to collect the minimal number of observations that is still sufficient in order to reach a valid conclusion.

As an example, consider an opinion poll aimed at the estimation of the proportion in the population of those that support a specific candidate that considers running for an office. How large the sample must be in order to assure, with high probability, that the percentage in the sample of supporters is within 0.5% of the percentage in the population? Within 0.25%?

A natural way to address this problem is via a confidence interval for the proportion. If the range of the confidence interval is no more than 0.05 (or 0.025 in the other case) then with a probability equal to the confidence level it is assured that the population relative frequency is within the given distance from the sample proportion.

Consider a confidence level of 95%. Recall that the structure of the confidence interval for the proportion is $\hat{P} \pm 1.96 \cdot \{\hat{P}(1-\hat{P})/n\}^{1/2}$. The range of the confidence interval is $1.96 \cdot \{\hat{P}(1-\hat{P})/n\}^{1/2}$. How large should n be in order to guarantee that the range is no more than 0.05?

The answer to this question depends on the magnitude of $\hat{P}(1-\hat{P})$, which is a random quantity. Luckily, one may observe that the maximal value[4] of the quadratic function $f(p) = p(1-p)$ is 1/4. It follows that

$$1.96 \cdot \{\hat{P}(1-\hat{P})/n\}^{1/2} \leq 1.96 \cdot \{0.25/n\}^{1/2} = 0.98/\sqrt{n}\,.$$

[4]The derivative is $f'(p) = 1 - 2p$. Solving $f'(p) = 0$ produces $p = 1/2$ as the maximizer. Plugging this value in the function gives 1/4 as the maximal value of the function.

Finally,

$$0.98/\sqrt{n} \leq 0.05 \quad \Longrightarrow \quad \sqrt{n} \geq 0.98/0.05 = 19.6 \quad \Longrightarrow \quad n \geq (19.6)^2 = 384.16 \ .$$

The conclusion is that n should be larger than 384 in order to assure the given range. For example, $n = 385$ should be sufficient.

If the request is for an interval of range 0.025 then the last line of reasoning should be modified accordingly:

$$0.98/\sqrt{n} \leq 0.025 \quad \Longrightarrow \quad \sqrt{n} \geq \frac{0.98}{0.025} = 39.2 \quad \Longrightarrow \quad n \geq (39.2)^2 = 1536.64 \ .$$

Consequently, $n = 1537$ will do. Increasing the accuracy by 50% requires a sample size that is 4 times larger.

More examples that involve selection of the sample size will be considered as part of the homework.

11.5 Solved Exercises

Question 11.1. This exercise deals with an experiment that was conducted among students. The aim of the experiment was to assess the effect of rumors and prior reputation of the instructor on the evaluation of the instructor by her students. The experiment was conducted by Towler and Dipboye[5]. 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 "Instructor Reputation and Teacher Ratings" that is presented in that site.

The experiment involved 49 students that were randomly assigned to one of two conditions. Before viewing the lecture, students were give one of two "summaries" of the instructor's prior teaching evaluations. The first type of summary, i.e. the first condition, described the lecturer as a charismatic instructor. The second type of summary (second condition) described the lecturer as a punitive instructor. We code the first condition as "`C`" and the second condition as "`P`". All subjects watched the same twenty-minute lecture given by the exact same lecturer. Following the lecture, subjects rated the lecturer.

The outcomes are stored in the file "`teacher.csv`". The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/teacher.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. Identify, for each variable in the file "`teacher.csv`", the name and the type of the variable (factor or numeric).

2. Estimate the expectation and the standard deviation among all students of the rating of the teacher.

3. Estimate the expectation and the standard deviation of the rating only for students who were given a summary that describes the teacher as charismatic.

[5] Towler, A. and Dipboye, R. L. (1998). The effect of instructor reputation and need for cognition on student behavior (poster presented at American Psychological Society conference, May 1998).

4. Construct a confidence interval of 99% confidence level for the expectation of the rating among students who were given a summary that describes the teacher as charismatic. (Assume the ratings have a Normal distribution.)

5. Construct a confidence interval of 90% confidence level for the variance of the rating among students who were given a summary that describes the teacher as charismatic. (Assume the ratings have a Normal distribution.)

Solution (to Question 11.1.1): We read the content of the file "`teacher.csv`" into a data frame by the name "`teacher`" and produce a summary of the content of the data frame:

```
> teacher <- read.csv("teacher.csv")
> summary(teacher)
 condition     rating
 C:25      Min.   :1.333
 P:24      1st Qu.:2.000
           Median :2.333
           Mean   :2.429
           3rd Qu.:2.667
           Max.   :4.000
```

There are two variables: The variable "`condition`" is a factor with two levels, "`C`" that codes the Charismatic condition and "`P`" that codes the Punitive condition. The second variable is "`rating`", which is a numeric variable.

Solution (to Question 11.1.2): The sample average for the variable "`rating`" can be obtained from the summary or from the application of the function "`mean`" to the variable. The standard deviation is obtained from the application of the function "`sd`" to the variable:

```
> mean(teacher$rating)
[1] 2.428567
> sd(teacher$rating)
[1] 0.5651949
```

Observe that the sample average is equal to 2.428567 and the sample standard deviation is equal to 0.5651949.

Solution (to Question 11.1.3): The sample average and standard deviation for each sub-sample may be produced with the aid of the function "`tapply`". We apply the function in the third argument, first "`mean`" and then "`sd`" to the variable rating, in the first argument, over each level of the factor "`condition`" in the second argument:

```
> tapply(teacher$rating,teacher$condition,mean)
       C        P
2.613332 2.236104
> tapply(teacher$rating,teacher$condition,sd)
        C         P
0.5329833 0.5426667
```

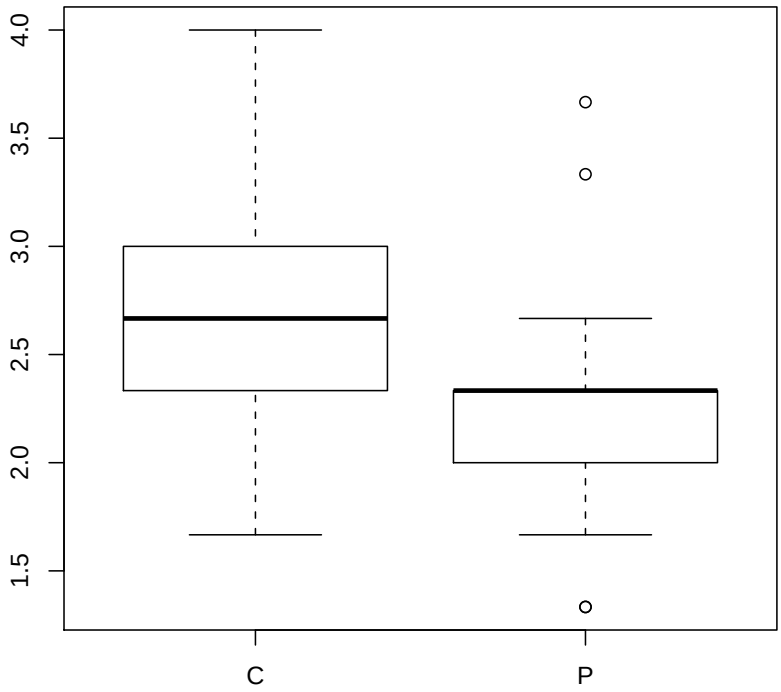

Figure 11.2: Box Plots of Ratings

Obtain that average for the condition "`C`" is 2.613332 and the standard deviation is 0.5329833.

You may note that the rating given by students that were exposed to the description of the lecturer as charismatic is higher on the average than the rating given by students that were exposed to a less favorable description. The box plots of the ratings for the two conditions are presented in Figure 11.2.

Solution (to Question 11.1.4): The 99% confidence interval for the expectation is computed by the formula $\bar{x} \pm$ `qt(0.995,n-1)` $\cdot s/\sqrt{n}$. Only 0.5% of the t-distribution on $(n-1)$ degrees of freedom resides above the percentile "`qt(0.995,n-1)`". Using this percentile leaves out a total of 1% in both tails and keeps 99% of the distribution inside the central interval.

For the students that were exposed to Condition "`C`", $\bar{x} = 2.613332$, $s = 0.5329833$, and $n = 25$:

```
> 2.613332 - qt(0.995,24)*0.5329833/sqrt(25)
[1] 2.315188
> 2.613332 + qt(0.995,24)*0.5329833/sqrt(25)
```

```
[1] 2.911476
```

The confidence interval for the expectation is $[2.315188, 2.911476]$.

Solution (to Question 11.1.5): The 90% confidence interval for the variance is computed by the formula $\left[\frac{n-1}{\texttt{qchisq(0.95,n-1)}}s^2,\ \frac{n-1}{\texttt{qchisq(0.05,n-1)}}s^2\right]$. Observe that 5% of the chi-square distribution on $(n-1)$ degrees of freedom is above the percentile "`qchisq(0.95,n-1)`" and 5% are below the percentile "`qchisq(0.05,n-1)`".

For the students that were exposed to Condition "`C`", $s = 0.5329833$, and $n = 25$:

```
> (24/qchisq(0.95,24))*0.5329833^2
[1] 0.1872224
> (24/qchisq(0.05,24))*0.5329833^2
[1] 0.4923093
```

The point estimate of the variance is $s^2 = 0.5329833^2 = 0.2840712$. The confidence interval for the variance is $[0.18722243, 0.4923093]$.

Question 11.2. Twenty observations are used in order to construct a confidence interval for the expectation. In one case, the construction is based on the Normal approximation of the sample average and in the other case it is constructed under the assumption that the observations are Normally distributed. Assume that in reality the measurement is distributed Exponential(1/4).

1. Compute, via simulation, the actual confidence level for the first case of a confidence interval with a nominal confidence level of 95%.
2. Compute, via simulation, the actual confidence level for the second case of a confidence interval with a nominal confidence level of 95%.
3. Which of the two approaches would you prefer?

Solution (to Question 11.2.1): Let us produce the sampling distribution of the sample average and the sample standard deviation for 20 observations from the Exponential(1/4) distribution:

```
> lam <- 1/4
> n <- 20
> X.bar <- rep(0,10^5)
> S <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rexp(n,lam)
+    X.bar[i] <- mean(X)
+    S[i] <- sd(X)
+ }
```

We compute the confidence level for a confidence interval with a nominal confidence level of 95%. Observe that using the Normal approximation of the sample average corresponds to the application of the Normal percentile in the construction of the confidence interval.

```
> norm.LCL <- X.bar - qnorm(0.975)*S/sqrt(n)
> norm.UCL <- X.bar + qnorm(0.975)*S/sqrt(n)
> mean((4 >= norm.LCL) & (4 <= norm.UCL))
[1] 0.9047
```

The expectation of the measurement is equal to 4. This number belongs to the confidence interval 90.47% of the times. Consequently, the actual confidence level is 90.47%.

Solution (to Question 11.2.2): Using the same sampling distribution that was produced in the solution to Question 1 we now compute the actual confidence level of a confidence interval that is constructed under the assumption that the measurement has a Normal distribution:

```
> t.LCL <- X.bar - qt(0.975,n-1)*S/sqrt(n)
> t.UCL <- X.bar + qt(0.975,n-1)*S/sqrt(n)
> mean((4 >= t.LCL) & (4 <= t.UCL))
[1] 0.91953
```

Based on the assumption we used the percentiles of the t-distribution. The actual significance level is $91.953\% \approx 92\%$, still short of the nominal 95% confidence level.

Solution (to Question 11.2.3): It would be preferable to use the (incorrect) assumption that the observations have a Normal distribution than to apply the Normal approximation to such a small sample. In the current setting the former produced a confidence level that is closer to the nominal one. In general, using the percentiles of the t-distribution will produce wider and more conservative confidence intervals than those produces under the Normal approximation of the average. To be on the safer size, one typically prefers the more conservative confidence intervals.

Question 11.3. Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

1. When designing a study to determine this proportion, what is the minimal sample size that is required for a 99% confident that the population proportion is accurately estimated, up to an error of 0.03?
2. Suppose that the insurance companies did conduct the study by surveying 400 drivers. They found that 320 of the drives claim to always buckle up. Produce an 80% confidence interval for the population proportion of drivers who claim to always buckle up.

Solution (to Question 11.3.1): The range of the confidence interval with 99% confidence interval is bounded by

$$\texttt{qnorm(0.995)} \cdot \{\hat{P}(1-\hat{P})/n\}^{1/2} \leq 2.575829 \cdot \sqrt{0.25/n} = 1.287915/\sqrt{n}\,,$$

since $\texttt{qnorm(0.995)} = 2.575829$ and $\hat{P}(1-\hat{P}) \leq 0.25$. Consequently, the sample size n should satisfy the inequality:

$$\begin{aligned} 1.287915/\sqrt{n} \leq 0.03 \quad &\Longrightarrow \quad \sqrt{n} \geq 1.287915/0.03 = 42.9305 \\ &\Longrightarrow \quad n \geq (42.9305)^2 = 1843.028\,. \end{aligned}$$

The smallest integer larger than the lower bound is $n = 1844$.

Solution (to Question 11.3.1): The 80% confidence interval for the probability is computed by the formula $\hat{p} \pm$ `qnorm(0.90)` $\cdot \sqrt{\hat{p}(1-\hat{p})/n}$:

```
> n <- 400
> p.hat <- 320/400
> p.hat - qnorm(0.90)*sqrt(p.hat*(1-p.hat)/n)
[1] 0.774369
> p.hat + qnorm(0.90)*sqrt(p.hat*(1-p.hat)/n)
[1] 0.825631
```

We obtain a confidence interval of the form $[0.774369, 0.825631]$.

11.6 Summary

Glossary

Confidence Interval: An interval that is most likely to contain the population parameter.

Confidence Level: The sampling probability that random confidence intervals contain the parameter value. The confidence level of an observed interval indicates that it was constructed using a formula that produces, when applied to random samples, such random intervals.

t-Distribution: A bell-shaped distribution that resembles the standard Normal distribution but has wider tails. The distribution is characterized by a positive parameter called *degrees of freedom*.

Chi-Square Distribution: A distribution associated with the sum of squares of Normal random variable. The distribution obtains only positive values and it is not symmetric. The distribution is characterized by a positive parameter called *degrees of freedom*.

Discuss in the forum

When large samples are at hand one may make fewer a-priori assumptions regarding the exact form of the distribution of the measurement. General limit theorems, such as the Central Limit Theorem, may be used in order to establish the validity of the inference under general conditions. On the other hand, for small sample sizes one must make strong assumptions with respect to the distribution of the observations in order to justify the validity of the procedure.

It may be claimed that making statistical inferences when the sample size is small is worthless. How can one trust conclusions that depend on assumptions regarding the distribution of the observations, assumptions that cannot be verified? What is your opinion?

For illustration consider the construction of a confidence interval. Confidence interval for the expectation is implemented with a specific formula. The significance level of the interval is provable when the sample size is large or when the sample size is small but the observations have a Normal distribution. If the

sample size is small and the observations have a distribution different from the Normal then the nominal significance level may not coincide with the actual significance level.

Formulas for Confidence Intervals, 95% Confidence Level:

- Expectation: $\bar{x} \pm \texttt{qnorm(0.975)} \cdot s/\sqrt{n}$.
- Probability: $\bar{p} \pm \texttt{qnorm(0.975)} \cdot \hat{p}(1-\hat{p})/\sqrt{n}$.
- Normal Expectation: $\bar{x} \pm \texttt{qt(0.975,n-1)} \cdot s/\sqrt{n}$.
- Normal Expectation: $\Big[\frac{n-1}{\texttt{qchisq(0.975,n-1)}} s^2,\ \frac{n-1}{\texttt{qchisq(0.025,n-1)}} s^2\Big]$.

Chapter 12

Testing Hypothesis

12.1 Student Learning Objectives

Hypothesis testing emerges as a crucial component in decision making where one of two competing options needs to be selected. Statistical hypothesis testing provides formal guidelines for making such a selection. This chapter deals with the formulation of statistical hypothesis testing and describes the associated decision rules. Specifically, we consider hypothesis testing in the context of the expectation of a measurement and in the context of the probability of an event. In subsequent chapters we deal with hypothesis testing in the context of other parameters as well. By the end of this chapter, the student should be able to:

- Formulate statistical hypothesis for testing.
- Test, based on a sample, hypotheses regarding the expectation of the measurement and the probability of an event.
- Identify the limitations of statistical hypothesis testing and the danger of misinterpretation of the test's conclusions.

12.2 The Theory of Hypothesis Testing

Statistical inference is used in order to detect and characterize meaningful phenomena that may be hidden in an environment contaminated by random noise. Hypothesis testing is an important step, typically the first, in the process of making inferences. In this step one tries to answer the question: "Is there a phenomena at all?". The basic approach is to determine whether the observed data can or cannot be reasonably explained by a model of randomness that does not involve the phenomena.

In this section we introduce the structure and characteristics of statistical hypothesis testing. We start with an informal application of a statistical test and proceed with formal definitions. In the next section we discuss in more detail the testing of hypotheses on the expectation of a measurement and the testing of hypotheses on the probability of an event. More examples are considered in subsequent chapters.

12.2.1 An Example of Hypothesis Testing

The variable "price" in the file "cars.csv" contains data on the prices of different types of cars that were sold in the United States during 1985. The average price of a car back then — the average of the variable "price" — was \$13,207. One may be interested in the question: Do Americans pay today for cars a different price than what they used to pay in the 80's? Has the price of cars changed significantly since 1985?

The average price of a car in the United States in 2009 was \$27,958[1]. Clearly, this figure is higher than \$13,207. However, in order to produce a fair answer to the question we have to take into account that, due to inflation, the prices of all products went up during these years. A more meaningful comparison will involve the current prices of cars in terms of 1985 Dollars. Indeed, if we take into account inflation then we get that, on the average, the cost of today's cars corresponds to an average price of \$13,662 in 1985 values[2]. This price is still higher than the prices in the 1985 but not as much. The question we are asking is: "Is the difference between \$13,207 and \$13,662 significant or is it not so?".

In order to give a statistical answer to this question we carry out a statistical test. The specific test is conducted with the aid of the function "t.test". Later we will discuss in more details some of the arguments that may be used in this function. Currently, we simply apply it to the data stored in the variable "price" to test that the expected price is different than the \$13,662, the average price of a car in 2009, adjusted for inflation:

```
> cars <- read.csv("cars.csv")
> t.test(cars$price,mu=13662)

        One Sample t-test

data:  cars$price
t = -0.8115, df = 200, p-value = 0.4181
alternative hypothesis: true mean is not equal to 13662
95 percent confidence interval:
 12101.80 14312.46
sample estimates:
mean of x
 13207.13
```

The data in the file "cars.csv" is read into a data frame that is given the name "cars". Afterwards, the data on prices of car types in 1985 is entered as the first argument to the function "t.test". The other argument is the expected value that we want to test, the current average price of cars, given in terms of 1985 Dollar value. The output of the function is reported under the title: "One Sample t-test".

Let us read the report from the bottom up. The bottom part of the report describes the confidence interval and the point estimate of the expected price of a car in 1985, based on the given data. Indeed, the last line reports the

[1]Source: "http://wiki.answers.com/Q/Average_price_of_a_car_in_2009".

[2]Source: "http://www.westegg.com/inflation/". The interpretation of adjusting prices to inflation is that our comparison will correspond to changes in the price of cars, relative to other items that enter into the computation of the Consumer Price Index.

sample average of the price, which is equal to 13,207.13. This number, the average of the 201 non-missing values of the variable "`price`", serves as the estimate of the expected price of a car in 1985. The 95% confidence interval of the expectation, the interval [12101.80, 14312.46], is presented on the 4th line from the bottom. This is the confidence interval for the expectation that was computed in Subsection 11.2.1[3].

The information relevant to conducting the statistical test itself is given in the upper part of the report. Specifically, it is reported that the data in "`cars$price`" is used in order to carry out the test. Based on this data a test statistic is computed and obtains the value of "`t = -0.8115`". This statistic is associated with the t-distribution with "`df = 200`" degrees of freedom. The last quantity that is being reported is denoted the *p-value* and it obtains the value "`p-value = 0.4181`". The test may be carried out with the aid of the value of the t statistic or, more directly, using the p-value. Currently we will use the p-value.

The test itself examines the hypothesis that the expected price of a car in 1985 was equal to \$13,662, the average price of a car in 2009, given in 1985 values. This hypothesis is called the null hypothesis. The alternative hypothesis is that the expected price of a car in 1985 was not equal to that figure. The specification of the alternative hypothesis is reported on the third line of the output of the function "`t.test`".

One may decide between the two hypothesis on the basis of the size of the p-value. The rule of thumb is to reject the null hypothesis, and thus accept the alternative hypothesis, if the p-value is less than 0.05. In the current example the p-value is equal 0.4181 and is larger than 0.05. Consequently, we may conclude that the expected price of a car in 1985 was not significantly different than the current price of a car.

In the rest of this section we give a more rigorous explanation of the theory and practice of statistical hypothesis testing.

12.2.2 The Structure of a Statistical Test of Hypotheses

The initial step in statistical inference in general, and in statistical hypothesis testing in particular, is the formulation of the statistical model and the identification of the parameter/s that should be investigated. In the current situation the statistical model may correspond to the assumption that the data in the variable "`price`" are an instance of a *random* sample (of size $n = 201$). The parameter that we want to investigate is the expectation of the measurement that produced the sample. The variance of the measurement is also relevant for the investigation.

After the statistical model has been set, one may split the process of testing a statistical hypothesis into three steps: (i) formulation of the hypotheses, (ii) specification of the test, and (iii) reaching the final conclusion. The first two steps are carried out on the basis of the probabilistic characteristics of the

[3] As a matter of fact, the confidence interval computed in Subsection 11.2.1 is [12108.47, 14305.79], which is not identical to the confidence that appears in the report. The reason for the discrepancy is that we used the 0.975-percentile of the Normal distribution, 1.96, whereas the confidence interval computed here uses the 0.975-percentile of the t-distribution on 201-1=200 degrees of freedom. The latter is equal to 1.971896. Nonetheless, for all practical purposes, the two confidence intervals are the same.

statistical model and in the context of the sampling distribution. In principal, the first two steps may be conducted in the planning stage prior to the collection of the observations. Only the third step involves the actual data. In the example that was considered in the previous subsection the third step was applied to the data in the variable "`price`" using the function "`t.test`".

(i) Formulating the hypotheses: A statistical model involves a parameter that is the target of the investigation. In principle, this parameter may obtain any value within a range of possible values. The formulation of the hypothesis corresponds to splitting the range of values into two sub-collections: a sub-collection that emerges in response to the presence of the phenomena and a sub-collection that emerges in response to the situation when the phenomena is absent. The sub-collection of parameter values where the phenomena is absent is called the *null hypothesis* and is marked as "H_0". The other sub-collection, the one reflecting the presence of the phenomena, is denoted the *alternative hypothesis* and is marked "H_1".

For example, consider the price of cars. Assume that the phenomena one wishes to investigate is the change in the relative price of a car in the 80's as compared to prices today. The parameter of interest is the expected price of cars back then, which we denote by $\mathrm{E}(X)$. The formulation of the statement that the expected price of cars has changed is "$\mathrm{E}(X) \neq 13,662$". This statement corresponds to the *presence* of a phenomena, to a change, and is customarily defined as the alternative hypothesis. On the other hand, the situation "$\mathrm{E}(X) = 13,662$" corresponds to not having any change in the price of cars. Hence, this situation corresponds to the *absence* of the phenomena and is denoted the null hypothesis. In summary, in order to investigate the change in the relative price of cars we my consider the null hypothesis "$H_0 : \mathrm{E}(X) = 13,662$" and test it against the alternative hypothesis "$H_1 : \mathrm{E}(X) \neq 13,662$".

A variation in the formulation of the phenomena can change the definition of the null and alternative hypotheses. For example, if the intention is to investigate the *rise* in the price of cars then the phenomena will correspond to the expected price in 1985 being less than \$13,662. Accordingly, the alternative hypothesis should be defined as $H_1 : \mathrm{E}(X) < 13,662$, with the null hypothesis defined as $H_0 : \mathrm{E}(X) \geq 13,662$. Observe that in this case an expected price larger than \$13,662 relates to the phenomena of rising (relative) prices *not* taking place.

On the other hand, if one would wants to investigate a *decrease* in the price then one should define the alternative hypothesis to be $H_1 : \mathrm{E}(X) > 13,662$, with the null hypothesis being $H_0 : \mathrm{E}(X) \leq 13,662$.

The type of alternative that was considered in the example, $H_1 : \mathrm{E}(X) \neq 13,622$ is called a *two-sided* alternative. The other two types of alternative hypotheses that were considered thereafter, $H_1 : \mathrm{E}(X) < 13,662$ and $H_1 : \mathrm{E}(X) > 13,662$, are both called *one-sided* alternatives.

In summary, the formulation of the hypothesis is a reflection of the phenomena one wishes to examine. The setting associated with the presence of the phenomena is denoted the alternative hypothesis and the complimentary setting, the setting where the phenomena is absent, is denoted the null hypothesis.

(ii) Specifying the test: The second step in hypothesis testing involves the selection of the decision rule, i.e. the statistical test, to be used in order to decide

between the two hypotheses. The decision rule is composed of a statistic and a subset of values of the statistic that correspond to the rejection of the null hypothesis. The statistic is called the *test statistic* and the subset of values is called the *rejection region*. The decision is to reject the null hypothesis (and consequently choose the alternative hypothesis) if the test statistic falls in the rejection region. Otherwise, if the test statistic does not fall in the rejection region then the null hypothesis is selected.

Return to the example in which we test between $H_0 : \mathrm{E}(X) = 13,662$ and $H_1 : \mathrm{E}(X) \neq 13,662$. One may compute the statistic:

$$T = \frac{\bar{X} - 13,662}{S/\sqrt{n}} ,$$

where $\bar{X}$ is the sample average (of the variable "`price`"), S is the sample standard deviation, and n is the sample size ($n = 201$ in the current example).

The sample average $\bar{X}$ is an estimator of a expected price of the car. In principle, the statistic T measures the discrepancy between the estimated value of the expectation ($\bar{X}$) and the expected value under the null hypothesis ($\mathrm{E}(X) = 13,662$). This discrepancy is measured in units of the (estimated) standard deviation of the sample average[4].

If the null hypothesis $H_0 : \mathrm{E}(X) = 13,662$ is true then the sampling distribution of the sample average $\bar{X}$ should be concentrated about the value 13,662. Values of the sample average much larger or much smaller than this value may serve as evidence against the null hypothesis.

In reflection, if the null hypothesis holds true then the values of the sampling distribution of the statistic T should tend to be in the vicinity of 0. Values with a relative small absolute value are consistent with the null hypothesis. On the other hand, extremely positive or extremely negative values of the statistic indicate that the null hypothesis is probably false.

It is natural to set a value c and to reject the null hypothesis whenever the absolute value of the statistic T is larger than c. The resulting rejection region is of the form $\{|T| > c\}$. The rule of thumb, again, is to take threshold c to be equal the 0.975-percentile of the t-distribution on $n-1$ degrees of freedom, where n is the sample size. In the current example, the sample size is $n = 201$ and the percentile of the t-distribution is `qt(0.975,200)` = 1.971896. Consequently, the subset $\{|T| > 1.971896\}$ is the rejection region of the test.

A change in the hypotheses that are being tested may lead to a change in the test statistic and/or the rejection region. For example, for testing $H_0 : \mathrm{E}(X) \geq 13,662$ versus $H_1 : \mathrm{E}(X) < 13,662$ one may still use the same test statistic T as before. However, only very negative values of the statistic are inconsistent with the null hypothesis. It turns out that the rejection region in this case is of the form $\{T < -1.652508\}$, where `qt(0.05,200)` = -1.652508 is the 0.05-percentile of the t-distribution on 200 degrees of freedom. On the other hand, the rejection region for testing between $H_0 : \mathrm{E}(X) \leq 13,662$ and $H_1 : \mathrm{E}(X) > 13,662$ is $\{T > 1.652508\}$. In this case, `qt(0.95,200)` = 1.652508 is the 0.95-percentile of the t-distribution on 200 degrees of freedom.

[4]If the variance of the measurement $\mathrm{Var}(X)$ was known one could have use $Z = (\bar{X} - -13,662)/\sqrt{\mathrm{Var}X/n}$ as a test statistic. This statistic corresponds to the discrepancy of the sample average from the null expectation in units of its standard deviation, i.e. the z-value of the sample average. Since the variance of the observation is unknown, we use an estimator of the variance (S^2) instead.

Selecting the test statistic and deciding what rejection region to use specifies the statistical test and completes the second step.

(iii) Reaching a conclusion: After the stage is set, all that is left is to apply the test to the observed data. This is done by computing the observed value of the test statistic and checking whether or not the observed value belongs to the rejection region. If it does belong to the rejection region then the decision is to reject the null hypothesis. Otherwise, if the statistic does not belong to the rejection region, then the decision is to accept the null hypothesis.

Return to the example of testing the price of car types. The observed value of the T statistic is part of the output of the application of the function "`t.test`" to the data. The value is "`t = -0.8115`". As an exercise, let us recompute directly from the data the value of the T statistic:

```
> x.bar <- mean(cars$price,na.rm=TRUE)
> x.bar
[1] 13207.13
> s <- sd(cars$price,na.rm=TRUE)
> s
[1] 7947.066
```

The observed value of the sample average is $\bar{x} = 13207.13$ and the observed value of the sample standard deviation is $s = 7947.066$. The sample size (due to having 4 missing values) is $n = 201$. The formula for the computation of the test statistic in this example is $t = [\bar{x} - 13,662]/[s/\sqrt{n}]$. Plugging in this formula the sample size and the computed values of the sample average and standard deviation produces:

```
> (x.bar - 13662)/(s/sqrt(201))
[1] -0.8114824
```

This value, after rounding up, is equal to the value "`t = -0.8115`" that is reported in the output of the function "`t.test`".

The critical threshold for the absolute value of the T statistic on $201-1 = 200$ degrees of freedom is `qt(0.975,200)` $= 1.971896$. Since the absolute observed value ($|t| = 0.8114824$) is less then the threshold we get that the value of the statistic does not belong to the rejection region (which is composed of absolute values larger than the threshold). Consequently, we accept the null hypothesis. This null hypothesis declares that the expected price of a car was equal to the current expected price (after adjusting for the change in Consumer Price Index)[5].

12.2.3 Error Types and Error Probabilities

The T statistic was proposed for testing a change in the price of a car. This statistic measures the discrepancy between the sample average price of a car and

[5]Previously, we carried out the same test using the p-value. The computed p-value in this example is 0.4181. The null hypothesis was accepted since this value is larger than 0.05. As a matter of fact, the test that uses the T statistic as a test statistic and reject the null hypothesis for absolute values larger than `qt(0.975,n-1)` is equivalent to the test that uses the p-value and rejects the null hypothesis for p-values less than 0.05. Below we discuss the computation of the p-value.

the expected value of the sample average, where the expectation is computed under the null hypothesis. The structure of the rejection region of the test is $\{|T| > c\}$, where c is an appropriate threshold. In the current example the value of the threshold c was set to be equal to `qt(0.975,200)` $= 1.971896$. In general, the specification of the threshold c depends on the error probabilities that are associated with the test. In this section we describe these error probabilities.

The process of making decisions may involve errors. In the case of hypothesis testing one may specify two types of error. On the one hand, the case may be that the null hypothesis is correct (in the example, $\mathrm{E}(X) = 13,662$). However, the data is such that the null hypothesis is rejected (here, $|T| > 1.971896$). This error is called a *Type I* error.

A different type of error occurs when the alternative hypothesis holds ($\mathrm{E}(X) \neq 13,662$) but the null hypothesis is not rejected ($|T| \leq 1.971896$). This other type of error is called *Type II* error. A summary of the types of errors can be found in Table 12.1:

	$H_0 : \mathrm{E}(X) = 13,662$	$H_1 : \mathrm{E}(X) \neq 13,662$
Accept H_0: $\lvert T\rvert \leq 1.971896$	✓	Type II Error
Reject H_0: $\lvert T\rvert > 1.971896$	Type I Error	✓

Table 12.1: Error Types

In statistical testing of hypothesis the two types of error are not treated symmetrically. Rather, making a Type I error is considered more severe than making a Type II error. Consequently, the test's decision rule is designed so as to assure an acceptable probability of making a Type I error. Reducing the probability of a Type II error is desirable, but is of secondary importance.

Indeed, in the example that deals with the price of car types the threshold was set as high as `qt(0.975,200)` $= 1.971896$ in order to reject the null hypothesis. It is not sufficient that the sample average is not equal to 13,662 (corresponding to a threshold of 0), but it has to be significantly different from the expectation under the null hypothesis, the distance between the sample average and the null expectation should be relatively large, in order to exclude H_0 as an option.

The significance level of the evidence for rejecting the null hypothesis is based on the probability of the Type I error. The probabilities associated with the different types of error are presented in Table 12.2:

	$H_0 : \mathrm{E}(X) = 13,662$	$H_1 : \mathrm{E}(X) \neq 13,662$
$\mathrm{P}(\lvert T\rvert \leq c)$		Prob. of Type II Error
$\mathrm{P}(\lvert T\rvert > c)$	Significance Level	Statistical Power

Table 12.2: Error Probabilities

Observe that the probability of a Type I error is called the significance level. The significance level is set at some pre-specified level such as 5% or 1%, with 5% being the most widely used level. In particular, setting the threshold in the example to be equal to `qt(0.975,200)` $= 1.971896$ produces a test with a 5% significance level.

This lack of symmetry between the two hypothesis proposes another interpretation of the difference between the hypothesis. According to this interpretation

the null hypothesis is the one in which the cost of making an error is greater. Thus, when one separates the collection of parameter values into two subsets then the subset that is associated with a more severe error is designated as the null hypothesis and the other subset becomes the alternative.

For example, a new drug must pass a sequence of clinical trials before it is approved for distribution. In these trials one may want to test whether the new drug produces beneficial effect in comparison to the current treatment. Naturally, the null hypothesis in this case would be that the new drug is no better than the current treatment and the alternative hypothesis would be that it is better. Only if the clinical trials demonstrates a significant beneficiary effect of the new drug would it be released for marketing.

In scientific research, in general, the currently accepted theory, the conservative explanation, is designated as the null hypothesis. A claim for novelty in the form of an alternative explanation requires strong evidence in order for it to be accepted and be favored over the traditional explanation. Hence, the novel explanation is designated as the alternative hypothesis. It replaces the current theory only if the empirical data clearly supports its. The test statistic is a summary of the empirical data. The rejection region corresponds to values that are unlikely to be observed according to the current theory. Obtaining a value in the rejection region is an indication that the current theory is probably not adequate and should be replaced by an explanation that is more consistent with the empirical evidence.

The second type of error probability in Table 12.2 is the probability of a Type II error. Instead of dealing directly with this probability the tradition is to consider the complementary probability that corresponds to the probability of *not* making a Type II error. This complementary probability is called the *statistical power*:

$$\text{Statistical Power} = 1 - \text{Probability of Type II Error}$$

The statistical power is the probability of rejecting the null hypothesis when the state of nature is the alternative hypothesis. (In comparison, the significance level is the probability of rejecting the null hypothesis when the state of nature is the null hypothesis.) When comparing two decision rules for testing hypothesis, both having the same significance level, the one that possesses a higher statistical power should be favored.

12.2.4 p-Values

The p-value is another test statistic. It is associated with a specific test statistic and a structure of the rejection region. The p-value is equal to the significance level of the test in which the observed value of the statistic serves as the threshold. In the current example, where the T is the underlying test statistic and the structure of the rejection region is of the form $\{|T| > c\}$ then the p-value is equal to the probability of rejecting the null hypothesis in the case where the threshold c is equal to the observed absolute value of the T statistic. In other words:

$$p\text{-value} = \mathrm{P}(|T| > |t|) = \mathrm{P}(|T| > |-0.8114824|) = \mathrm{P}(|T| > 0.8114824)\ ,$$

where $t = -0.8114824$ is the observed value of the T statistic and the computation of the probability is conducted under the null hypothesis.

Specifically, under the null hypothesis $H_0 : \mathrm{E}(X) = 13,662$ we get that the distribution of the statistic $T = [\bar{X} - 13,662]/[S/\sqrt{n}]$ is the t-distribution on $n-1 = 200$ degrees of freedom. The probability of the event $\{|T| > 0.8114824\}$ corresponds to the sum of the probabilities of both tails of the distribution. By the symmetry of the t-distribution this equals twice the probability of the upper tail:

$$\mathrm{P}(|T| > 0.8114824) = 2 \cdot \mathrm{P}(T > 0.8114824) = 2 \cdot [1 - \mathrm{P}(|T| \leq 0.8114824)] \,.$$

When we compute this probability in `R` we get:

```
> 2*(1-pt(0.8114824,200))
[1] 0.4180534
```

This probability is equal, after rounding up, to the probability "`p-value = 0.4181`" that is reported in the output of the function "`t.test`".

The p-value is a function of the data. In the particular data set the computed value of the T statistic was -0.8114824. For a different data set the evaluation of the statistic would have produced a different value. As a result, the threshold that would have been used in the computation would have been different, thereby changing the numerical value of the p-value. Being a function of the data, we conclude that the p-value is a statistic.

The p-value is used as a test statistic by comparing its value to the predefined significance level. If the significance level is 1% then the null hypothesis is rejected for p-values less that 0.01. Likewise, if the significance level is set at the 5% level then the null hypothesis is rejected for p-values less than 0.05.

The statistical test that is based directly on the T statistic and the statistical test that is based on the p-value are equivalent to each other. The one rejects the null hypothesis if, and only if, the other does so. The advantage of using the p-value as the test statistic is that no further probabilistic computations are required. The p-value is compared directly to the significance level we seek. For the test that examines the T statistic we still need to identify the threshold associated with the given significance level.

In the next 2 sections we extend the discussion of the t-test and give further examples to the use of the function "`t.test`". We also deal with tests on probabilities of events and introduce the function "`prop.test`" for conducting such tests.

12.3 Testing Hypothesis on Expectation

Let us consider the variable "`dif.mpg`" that contains the difference in fuel consumption between highway and city conditions. This variable was considered in Chapter 11. Examine the distribution of this variable:

```
> dif.mpg <- cars$highway.mpg - cars$city.mpg
> summary(dif.mpg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   5.000   6.000   5.532   7.000  11.000
> plot(table(dif.mpg))
```

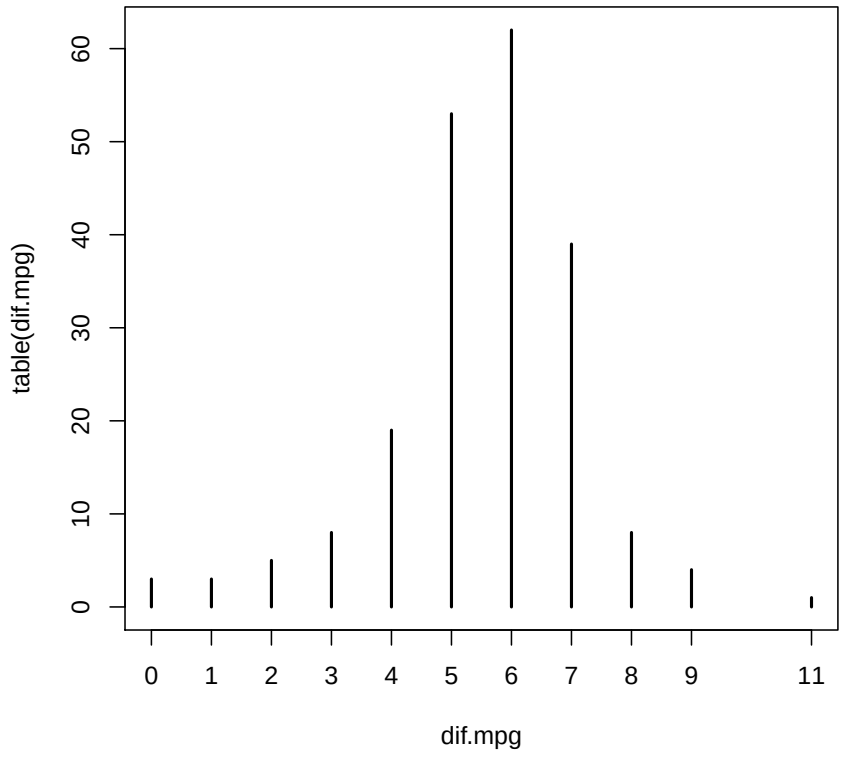

Figure 12.1: The Difference Between Highway and City MPG

In the first expression we created the variable "`dif.mpg`" that contains the difference in miles-per-gallon. The difference is computed for each car type between highway driving conditions and urban driving condition. The summary of this variable is produced in the second expression. Observe that the values of the variable range between 0 and 11, with 50% of the distribution concentrated between 5 and 7. The median is 6 and the mean is 5.532. The last expression produces the bar plot of the distribution. This bar plot is presented in Figure 12.1. It turns out that the variable "`dif.mpg`" obtains integer values.

In this section we test hypotheses regarding the expected difference in fuel consumption between highway and city conditions.

Energy is required in order to move cars. For heavier cars more energy is required. Consequently, one may conjecture that milage per gallon for heavier cars is less than the milage per gallon for lighter cars.

The relation between the weight of the car and the difference between the milage-per-gallon in highway and city driving conditions is less clear. On the one hand, urban traffic involves frequent changes in speed in comparison to highway conditions. One may presume that this change in speed is a cause for reduced efficiency in fuel consumption. If this is the case then one may predict that

heavier cars, which require more energy for acceleration, will be associated with a bigger difference between highway and city driving conditions in comparison to lighter cars.

One the other hand, heavier cars do less miles per gallon overall. The difference between two smaller numbers (the milage per gallon in highway and in city conditions for heavier cars) may tend to be smaller than the difference between two larger numbers (the milage per gallon in highway and in city conditions for lighter cars). If this is the case then one may predict that heavier cars will be associated with a smaller difference between highway and city driving conditions in comparison to lighter cars.

The average difference between highway and city conditions is approximately 5.53 for all cars. Divide the cars into to two groups of equal size: One group is composed of the heavier cars and the other group is composed of the lighter cars. We will examine the relation between the weight of the car and difference in miles per gallon between the two driving conditions by testing hypotheses separately for each weight group[6]. For each such group we start by testing the two-sided hypothesis $H_1 : \mathrm{E}(X) \neq 5.53$, where X is the difference between highway and city miles-per-gallon in cars that belong to the given weight group. After carrying the test for the two-sided alternative we will discuss results of the application of tests for one-sided alternatives.

We start by the definition of the weight groups. The variable "`curb.weight`" measures the weight of the cars in the data frame "`cars`". Let us examine the summary of the content of this variable:

```
> summary(cars$curb.weight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1488    2145    2414    2556    2935    4066
```

Half of the cars in the data frame weigh less than 2,414 lb and half the cars weigh more. The average weight of a car is 2,556 lb. Let us take 2,414 as a threshold and denote cars below this weight as "light" and cars above this threshold as "heavy":

```
> heavy <- cars$curb.weight > 2414
> table(heavy)
heavy
FALSE  TRUE
  103   102
```

The variable "`heavy`" indicates for each car type whether its weight is above or below the threshold weight of 2,414 lb. The variable is composed of a sequence with as many components as the number of observations in the data frame "`cars`" ($n = 205$). Each component is a logical value: "`TRUE`" if the car is heavier than the threshold and "`FALSE`" if it is not. When we apply the function "`table`" to this sequence we get that 102 of the cars are heavier than the threshold and 103 are not so.

[6] In the next chapters we will consider a more direct ways for comparing the effect of one variable (`curb.weight` in this example) on the distribution of another variable (`dif.mpg` in this example). Here, instead, we investigate the effect indirectly by the investigation of hypotheses on the expectation of the variable `dif.mpg` separately for heavier cars and for lighter cars.

We would like to apply the t-test first to the subset of all cars with weight above 2,414 lb (cars that are associated with the value "TRUE" in the variable "heavy"), and then to all cars with weights not exceeding the threshold (cars associated with value "FALSE"). In the past we showed that one may address components of a sequence using its position in the sequence[7]. Here we demonstrate an alternative approach for addressing specific locations by using a sequence with logical components.

In order to illustrate this second approach consider the two sequences:

```
> w <- c(5,3,4,6,2,9)
> d <- c(13,22,0,12,6,20)
```

Say we want to select the components of the sequence "d" in all the locations where the components of the sequence "w" obtain values larger than 5. Consider the code:

```
> w > 5
[1] FALSE FALSE FALSE  TRUE FALSE  TRUE
> d[w > 5]
[1] 12 20
```

The expression "w > 5" is a sequence of logical components, with the value "TRUE" at the positions where "w" is above the threshold and the value "FALSE" at the positions where "w" is below the threshold. We may use the sequence with logical components as an index to the sequence of the same length "d". The relevant expression is "d[w > 5]". The output of this expression is the sub-sequence of elements from "d" that are associated with the "TRUE" values of the logical sequence. Indeed, "TRUE" values are present at the 4th and the 6th positions of the logical sequence. Consequently, the output of the expression "d[w > 5]" contains the 4th and the 6th components of the sequence "d".

The operator "!", when applied to a logical value, reverses the value. A "TRUE" becomes "FALSE" and a "FALSE" becomes "TRUE". Consider the code:

```
> !(w > 5)
[1]  TRUE  TRUE  TRUE FALSE  TRUE FALSE
> d[!(w > 5)]
[1] 13 22  0  6
```

Observe that the sequence "!(w > 5)" obtains a value of "TRUE" at the positions where "w" is less or equal to 5. Consequently, the output of the expression "d[!(w > 5)]" are all the values of "d" that are associated with components of "w" that are less or equal to 5.

The variable "dif.mpg" contains data on the difference in miles-per-gallon between highway and city driving conditions for all the car types. The sequence "heavy" identifies the car types with curb weight above the threshold of 2,414 lb. The components of this sequence are logical with the value "TRUE" at positions associated with the heavier car types and the "FALSE" at positions associated with the lighter car types. Observe that the output of the expression "dif.mpg[heavy]" is the subsequence of differences in miles-per-gallon for

[7] For example, in Question 9.1 we referred to the first 29 observations of the sequence "change" using the expression "change[1:29]" and to the last 21 observations using the expression "change[30:50]".

the cars with curb weight above the given threshold. We apply the function "`t.test`" to this expression in order to conduct the t-test on the expectation of the variable "`dif.mpg`" for the heavier cars:

```
> t.test(dif.mpg[heavy],mu=5.53)

        One Sample t-test

data:  dif.mpg[heavy]
t = -1.5385, df = 101, p-value = 0.1270
alternative hypothesis: true mean is not equal to 5.53
95 percent confidence interval:
 4.900198 5.609606
sample estimates:
mean of x
 5.254902
```

The target population are the heavier car types. Notice that we test the null hypothesis that expected difference among he heavier cars is equal to 5.53 against the alternative hypothesis that the expected difference among heavier cars is not equal to 5.53. The null hypothesis is not rejected at the 5% significance level since the p-value, which is equal to 0.1735, is larger than 0.05. Consequently, based on the data at hand, we cannot conclude that the expected difference in miles-per-gallon for heavier cars is significantly different than the average difference for all cars.

Observe also that the estimate of the expectation, the sample mean, is equal to 5.254902, with a confidence interval of the form $[4.900198, 5.609606]$.

Next, let us apply the same test to the lighter cars. The expression "`dif.pmg[!heavy]`" produces the subsequence of differences in miles-per-gallon for the cars with curb weight below the given threshold. The application of the function "`t.test`" to this subsequence gives:

```
> t.test(dif.mpg[!heavy],mu=5.53)

        One Sample t-test

data:  dif.mpg[!heavy]
t = 1.9692, df = 102, p-value = 0.05164
alternative hypothesis: true mean is not equal to 5.53
95 percent confidence interval:
 5.528002 6.083649
sample estimates:
mean of x
 5.805825
```

Again, the null hypothesis is not rejected at the 5% significance level since a p-value of 0.05164 is still larger than 0.05. However, unlike the case for heavier cars where the p-value was undeniably larger than the threshold. In this example it is much closer to the threshold of 0.05. Consequently, we may almost conclude that the expected difference in miles-per-gallon for lighter cars is significantly different than the average difference for all car.

Why did we not reject the null hypothesis for the heavier cars but almost did so for the lighter cars? Both tests are based on the T statistic, which measures the ratio between the deviation of the sample average from its expectation under the null, divided by the estimate of the standard deviation of the sample average. The value of this statistic is "`t = -1.5385`" for heavier cars and it is "`t = 1.9692`" for lighter cars, an absolute value of about 25% higher.

The deviation of the sample average for the heavier cars and the expectation under the null is $5.254902 - 5.53 = -0.275098$. On the other hand, the deviation of the sample average for the lighter cars and the expectation under the null is $5.805825 - 5.53 = 0.275825$. The two deviations are practically equal to each other in the absolute value.

The estimator of the standard deviation of the sample average is $S/\sqrt{n}$, where S is the sample standard deviation and n is the sample size. The sample sizes, 103 for lighter cars and 102 for heavier cars, are almost equal. Therefore, the reason for the difference in the values of the T statistics for both weight groups must be differences in the sample standard deviations. Indeed, when we compute the sample standard deviation for lighter and heavier cars[8] we get that the standard deviation for lighter cars (1.421531) is much smaller than the standard deviation for heavier cars (1.805856):

```
> tapply(dif.mpg,heavy,sd)
   FALSE     TRUE
1.421531 1.805856
```

The important lesson to learn from this exercise is that simple minded notion of significance and statistical significance are not the same. A simple minded assessment of the discrepancy from the null hypothesis will put the evidence from the data on lighter cars and the evidence from the data on heavier cars on the same level. In both cases the estimated value of the expectation is the same distance away from the null value.

However, statistical assessment conducts the analysis in the context of the sampling distribution. The deviation of the sample average from the expectation is compared to the standard deviation of the sample average. Consequently, in statistical testing of hypothesis a smaller deviation of the sample average from the expectation under the null may be more significant than a larger one if the sampling variability of the former is much smaller than the sampling variability of the later.

Let us proceed with the demonstration of the application of the t-test by the testing of one-sided alternatives in the context of the lighter cars. One may test the one-sided alternative $H_1 : \mathrm{E}(X) > 5.53$ that the expected value of the difference in miles-per-gallon among cars with curb weight no more than 2,414 lb is *greater* than 5.53 by the application of the function "`t.test`" to the data on lighter cars. This data is the output of the expression "`dif.mpg[!heavy]`". As before, we specify the null value of the expectation by the introduction of the

[8] The function "`tapply`" applies the function that is given as its third argument (the function "`sd`" in this case) to each subset of values of the sequence that is given as its first argument (the sequence "`dif.mpg`" in the current application). The subsets are determined by the levels of the second arguments (the sequence "`heavy`" in this case). The output is the sample standard deviation of the variable "`dif.mpg`" for lighter cars (the level "`FALSE`") and for heavier cars (the level "`TRUE`").

expression "`mu=5.53`". The fact that we are interested in the testing of the specific alternative is specified by the introduction of a new argument of the form: "`alternative="greater"`". The default value of the argument "`alternative`" is "`"two.sided"`", which produces a test of a two-sided alternative. By changing the value of the argument to "`"greater"`" we produce a test for the appropriate one-sided alternative:

```
> t.test(dif.mpg[!heavy],mu=5.53,alternative="greater")

        One Sample t-test

data:  dif.mpg[!heavy]
t = 1.9692, df = 102, p-value = 0.02582
alternative hypothesis: true mean is greater than 5.53
95 percent confidence interval:
 5.573323      Inf
sample estimates:
mean of x
 5.805825
```

The value of the test statistic (`t = 1.9692`) is the same as for the test of the two-sided alternative and so is the number of degrees of freedom associated with the statistic (`df = 102`). However, the p-value is smaller (`p-value = 0.02582`), compared to the p-value in the test for the two-sided alternative (`p-value = 0.05164`). The p-value for the one-sided test is the probability under the sampling distribution that the test statistic obtains vales larger than the observed value of 1.9692. The p-value for the two-sided test is twice that figure since it involves also the probability of being less than the negative of the observes value.

The estimated value of the expectation, the sample average, is unchanged. However, instead of producing a confidence interval for the expectation the report produces a *one-sided* confidence interval of the form $[5.573323, \infty)$. Such an interval corresponds to the *smallest* value that the expectation may reasonably obtain on the basis of the observed data.

Finally, consider the test of the other one-sided alternative $H_1 : \mathrm{E}(X) < 5.53$:

```
> t.test(dif.mpg[!heavy],mu=5.53,alternative="less")

        One Sample t-test

data:  dif.mpg[!heavy]
t = 1.9692, df = 102, p-value = 0.9742
alternative hypothesis: true mean is less than 5.53
95 percent confidence interval:
     -Inf 6.038328
sample estimates:
mean of x
 5.805825
```

The alternative here is determined by the expression "`alternative="less"`". The p-value is equal to 0.9742, which is the probability that the test statistic

obtains values less than the observed value of 1.9692. Clearly, the null hypothesis is not rejected in this test.

12.4 Testing Hypothesis on Proportion

Consider the problem of testing hypothesis on the probability of an event. Recall that a probability p of some event can be estimated by the observed relative frequency of the event in the sample, denoted $\hat{P}$. The estimation is associated with the Bernoulli random variable X, that obtains the value 1 when the event occurs and the value 0 when it does not. The statistical model states that p is the expectation of X. The estimator $\hat{P}$ is the sample average of this measurement.

With this formulation we may relate the problem of testing hypotheses formulated in terms of p to the problem of tests associated to the expectation of a measurement. For the latter problem we applied the t-test. A similar, though not identical, test is used for the problem of testing hypothesis on proportions.

Assume that one in interested in testing the null hypothesis that the probability of the event is equal to some specific value, say one half, versus the alternative hypothesis that the probability is not equal to this value. These hypotheses are formulated as $H_0 : p = 0.5$ and $H_1 : p \neq 0.5$.

The sample proportion of the event $\hat{P}$ is the basis for the construction of the test statistic. Recall that the variance of the estimator $\hat{P}$ is given by $\mathrm{Var}(\hat{P}) = p(1-p)/n$. Under the null hypothesis we get that the variance is equal to $\mathrm{Var}(\hat{P}) = 0.5(1-0.5)/n$. A natural test statistic is the standardized sample proportion:

$$Z = \frac{\hat{P} - 0.5}{\sqrt{0.5(1-0.5)/n}} \ ,$$

that measures the ratio between the deviation of the estimator from its null expected value and the standard deviation of the estimator. The standard deviation of the sample proportion is used in the ratio.

If the null hypothesis that $p = 0.5$ holds true then one gets that the value 0 is the center of the sampling distribution of the test statistic Z. Values of the statistic that are much larger or much smaller than 0 indicate that the null hypothesis is unlikely. Consequently, one may consider a rejection region of the form $\{|Z| > c\}$, for some threshold value c. The threshold c is set at a high enough level to assure the required significance level, namely the probability under the null hypothesis of obtaining a value in the rejection region. Equivalently, the rejection region can be written in the form $\{Z^2 > c^2\}$.

As a result of the Central Limit Theorem one may conclude that the distribution of the test statistic is approximately Normal. Hence, Normal computations may be used in order to produce an approximate threshold or in order to compute an approximation for the p-value. Specifically, if Z has the standard Normal distribution then Z^2 has a chi-square distribution on one degree of freedom.

In order to illustrate the application of hypothesis testing for proportion consider the following problem: In the previous section we obtained the curb weight of 2,414 lb as the sample median. The weights of half the cars in the sample were above that level and the weights of half the cars were below this level. If this level was actually the population median then the probability that the weight of a random car is not exceeding this level would be equal to 0.5.

Let us test the hypothesis that the median weight of cars that run on diesel is also 2,414 lb. Recall that 20 out of the 205 car types in the sample have diesel engines. Let us use the weights of these cars in order to test the hypothesis.

The variable "`fuel.type`" is a factor with two levels "`diesel`" and "`gas`" that identify the fuel type of each car. The variable "`heavy`" identifies for each car whether its weight is above the level of 2414 or not. Let us produce a 2×2 table that summarizes the frequency of each combination of weight group and the fuel type:

```
> fuel <- cars$fuel.type
> table(fuel,heavy)
        heavy
fuel     FALSE TRUE
  diesel     6   14
  gas       97   88
```

Originally the function "`table`" was applied to a single factor and produced a sequence with the frequencies of each level of the factor. In the current application the input to the function are two factors[9]. The output is a table of frequencies. Each entry to the table corresponds to the frequency of a combination of levels, one from the first input factor and the other from the second input factor. In this example we obtain that 6 cars use diesel and their curb weight was below the threshold. There are 14 cars that use diesel and their curb weight is above the threshold. Likewise, there are 97 light cars that use gas and 88 heavy cars with gas engines.

The function "`prop.test`" produces statistical tests for proportions. The relevant information for the current application of the function is the fact that frequency of light diesel cars is 6 among a total number of 20 diesel cars. The first entry to the function is the frequency of the occurrence of the event, 6 in this case, and the second entry is the relevant sample size, the total number of diesel cars which is 20 in the current example:

```
> prop.test(6,20)

        1-sample proportions test with continuity correction

data:  6 out of 20, null probability 0.5
X-squared = 2.45, df = 1, p-value = 0.1175
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1283909 0.5433071
sample estimates:
  p
0.3
```

The function produces a report that is printed on the screen. The title identifies the test as a one-sample test of proportions. In later chapters we will apply the same function to more complex data structures and the title will

[9]To be more accurate, the variable "`heavy`" is not a factor but a sequence with logical components. Nonetheless, when the function "`table`" is applied to such a sequence it treats it as a factor with two levels, "`TRUE`" and "`FALSE`".

change accordingly. The title also identifies the fact that a continuity correction is used in the computation of the test statistic.

The line under the title indicates the frequency of the event in the sample and the sample size. (In the current example, 6 diesel cars with weights below the threshold among a total of 20 diesel cars.) The probability of the event, under the null hypothesis, is described. The default value of this probability is "`p = 0.5`", which is the proper value in the current example. This default value can be modified by replacing the value 0.5 by the appropriate probability.

The next line presents the information relevant for the test itself. The test statistic, which is essentially the square of the Z statistic described above[10], obtains the value 2.45. The sampling distribution of this statistic under the null hypothesis is, approximately, the chi-square distribution on 1 degree of freedom. The p-value, which is the probability that chi-square distribution on 1 degree of freedom obtains a value above 2.45, is equal to 0.1175. Consequently, the null hypothesis is not rejected at the 5% significance level.

The bottom part of the report provides the confidence interval and the point estimate for the probability of the event. The confidence interval for the given data is $[0.1283909, 0.5433071]$ and the point estimate is $\hat{p} = 6/20 = 0.3$.

It is interesting to note that although the deviation between the estimated proportion $\hat{p} = 0.3$ and the null value of the probability $p = 0.5$ is relatively large still the null hypothesis was not rejected. The reason for that is the smallness of the sample, $n = 20$, that was used in order to test the hypothesis. Indeed, as an exercise let us examine the application of the same test to a setting where $n = 200$ and the number of occurrences of the event is 60:

```
> prop.test(60,200)

        1-sample proportions test with continuity correction

data:  60 out of 200, null probability 0.5
X-squared = 31.205, df = 1, p-value = 2.322e-08
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2384423 0.3693892
sample estimates:
  p
0.3
```

The estimated value of the probability is the same as before since $\hat{p} = 60/200 = 0.3$. However, the p-value is 2.322×10^{-8}, which is way below the significance threshold of 0.05. In this scenario the null hypothesis is rejected with flying colors.

This last example is yet another demonstration of the basic characteristic of statistical hypothesis testing. The consideration is based not on the discrepancy of the estimator of the parameter from the value of the parameter under

[10] The test statistic that is computed by default is based on *Yates' correction for continuity*, which is very similar to the continuity correction that was used in Chapter 6 for the Normal approximation of the Binomial distribution. Specifically, the test statistic to the continuity correction for testing $H_0 : p = p_0$ takes the form $[|\hat{p} - p_0| - 0.5/n]^2/[p_0(1 - p_0)/n]$. Compare this statistic with the statistic proposed in the text that takes the form $[\hat{p} - p_0]^2/[p_0(1 - p_0)/n]$. The latter statistic is used if the argument "`correct = FALSE`" is added to the function.

the null. Instead, it is based on the *relative* discrepancy in comparison to the sampling variability of the estimator. When the sample size is larger the variability is smaller. Hence, the chances of rejecting the null hypothesis for the same discrepancy increases.

12.5 Solved Exercises

Question 12.1. Consider a medical condition that does not have a standard treatment. The recommended design of a clinical trial for a new treatment to such condition involves using a placebo treatment as a control. A placebo treatment is a treatment that externally looks identical to the actual treatment but, in reality, it does not have the active ingredients. The reason for using placebo for control is the "placebo effect". Patients tent to react to the fact that they are being treated regardless of the actual beneficial effect of the treatment.

As an example, consider the trial for testing magnets as a treatment for pain that was described in Question 9.1. The patients that where randomly assigned to the control (the last 21 observations in the file "`magnets.csv`") were treated with devises that looked like magnets but actually were not. The goal in this exercise is to test for the presence of a placebo effect in the case study "Magnets and Pain Relief" of Question 9.1 using the data in the file "`magnets.csv`".

1. Let X be the measurement of change, the difference between the score of pain before the treatment and the score after the treatment, for patients that were treated with the inactive placebo. Express, in terms of the expected value of X, the null hypothesis and the alternative hypothesis for a statistical test to determine the presence of a placebo effect. The null hypothesis should reflect the situation that the placebo effect is absent.

2. Identify the observations that can be used in order to test the hypotheses.

3. Carry out the test and report your conclusion. (Use a significance level of 5%.)

Solution (to Question 12.2.1): The null hypothesis of no placebo effect corresponds to the expectation of the change equal to 0 ($H_0 : \mathrm{E}(X) = 0$). The alternative hypothesis may be formulated as the expectation not being equal to 0 ($H_1 : \mathrm{E}(X) \neq 0$). This corresponds to a two sided alternative. Observe that a negative expectation of the change still corresponds to the placebo having an effect.

Solution (to Question 12.2.2): The observations that can be used in order to test the hypothesis are those associated with patients that were treated with the inactive placebo, i.e. the last 21 observations. We extracts these values from the data frame using the expression "`magnets$change[30:50]`".

Solution (to Question 12.2.3): In order to carry out the test we read the data from the file "`magnets.csv`" into the data frame "`magnets`". The function "`t.test`" is applied to the observations extracted from the data frame. Note that the default expectation value of tested by the function is "`mu = 0`":

```
> magnets <- read.csv("magnets.csv")
```

```
> t.test(magnets$change[30:50])

        One Sample t-test

data:  magnets$change[30:50]
t = 3.1804, df = 20, p-value = 0.004702
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.3768845 1.8135916
sample estimates:
mean of x
 1.095238
```

The computed p-value is 0.004702, which is below 0.05. Consequently, we reject the null hypothesis and conclude that a placebo effect seems to be present.

Question 12.2. It is assumed, when constructing the t-test, that the measurements are Normally distributed. In this exercise we examine the robustness of the test to divergence from the assumption. You are required to compute the significance level of a two-sided t-test of $H_0 : \mathrm{E}(X) = 4$ versus $H_1 : \mathrm{E}(X) \neq 4$. Assume there are $n = 20$ observations and use a t-test with a nominal 5% significance level.

1. Consider the case where $X \sim$ Exponential$(1/4)$.
2. Consider the case where $X \sim$ Uniform$(0, 8)$.

Solution (to Question 12.2.1): We simulate the sampling distribution of the sample average and standard deviation. The sample is composed of $n = 20$ observations from the given Exponential distribution:

```
> lam <- 1/4
> n <- 20
> X.bar <- rep(0,10^5)
> S <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- rexp(n,lam)
+    X.bar[i] <- mean(X)
+    S[i] <- sd(X)
+ }
> T <- (X.bar - 4)/(S/sqrt(n))
> mean(abs(T) > qt(0.975,n-1))
[1] 0.08047
```

We compute the test statistic "`T`" from the sample average "`X.bar`" and the sample standard deviation "`S`". In the last expression we compute the probability that the absolute value of the test statistic is larger than "`qt(0.975,19)`", which is the threshold that should be used in order to obtain a significance level of 5% for Normal measurements.

We obtain that the actual significance level of the test is 0.08047, which is substantially larger than the nominal significance level.

Solution (to Question 12.2.2): We repeat essentially the same simulations as before. We only change the distribution of the sample from the Exponential to the Uniform distribution:

```
> a <- 0
> b <- 8
> n <- 20
> X.bar <- rep(0,10^5)
> S <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X <- runif(n,a,b)
+    X.bar[i] <- mean(X)
+    S[i] <- sd(X)
+ }
> T <- (X.bar - 4)/(S/sqrt(n))
> mean(abs(T) > qt(0.975,n-1))
[1] 0.05163
```

The actual significance level of the test is 0.05163, much closer to the nominal significance level of 5%.

A possible explanation for the difference between the two cases is that the Uniform distribution is symmetric like the Normal distribution, whereas the Exponential is skewed. In any case, for larger sample sizes one may expect the Central Limit Theorem to kick in and produce more satisfactory results, even for the Exponential case.

Question 12.3. Assume that you are interested in testing $H_0 : \mathrm{E}(X) = 20$ versus $H_1 : \mathrm{E}(X) \neq 20$ with a significance level of 5% using the t-test. Let the sample average, of a sample of size $n = 55$, be equal to $\bar{x} = 22.7$ and the sample standard deviation be equal to $s = 5.4$.

1. Do you reject the null hypothesis?
2. Use the same information. Only now you are interested in a significance level of 1%. Do you reject the null hypothesis?
3. Use the information the presentation of the exercise. But now you are interested in testing $H_0 : \mathrm{E}(X) = 24$ versus $H_1 : \mathrm{E}(X) \neq 24$ (with a significance level of 5%). Do you reject the null hypothesis?

Solution (to Question 12.3.1): We input the data to `R` and then compute the test statistic and the appropriate percentile of the t-distribution:

```
> n <- 55
> x.bar <- 22.7
> s <- 5.4
> t <- (x.bar - 20)/(s/sqrt(n))
> abs(t)
[1] 3.708099
> qt(0.975,n-1)
[1] 2.004879
```

Observe that the absolute value of the statistic (3.708099) is larger than the threshold for rejection (2.004879). Consequently, we reject the null hypothesis.

Solution (to Question 12.3.2): We recompute the percentile of the t-distribution:

```
> qt(0.995,n-1)
[1] 2.669985
```

Again, the absolute value of the statistic (3.708099) is larger than the threshold for rejection (2.669985). Consequently, we reject the null hypothesis.

Solution (to Question 12.3.3): In this question we should recompute the test statistic:

```
> t <- (x.bar - 24)/(s/sqrt(n))
> abs(t)
[1] 1.785381
```

The absolute value of the new statistic (1.785381) is smaller than the threshold for rejection (2.004879). Consequently, we do not reject the null hypothesis.

12.6 Summary

Glossary

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.

Null Hypothesis (H_0): A sub-collection that emerges in response to the situation when the phenomena is absent. The established scientific theory that is being challenged. The hypothesis which is worse to erroneously reject.

Alternative Hypothesis (H_1): A sub-collection that emerges in response to the presence of the investigated phenomena. The new scientific theory that challenges the currently established theory.

Test Statistic: A statistic that summarizes the data in the sample in order to decide between the two alternative.

Rejection Region: A set of values that the test statistic may obtain. If the observed value of the test statistic belongs to the rejection region then the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected.

Type I Error The null hypothesis is correct but it is rejected by the test.

Type II Error The alternative hypothesis holds but the null hypothesis is not rejected by the test.

Significance Level: The probability of a Type I error. The probability, computed under the null hypothesis, of rejecting the null hypothesis. The test is constructed to have a given significance level. A commonly used significance level is 5%.

Statistical Power: The probability, computed under the alternative hypothesis, of rejecting the null hypothesis. The statistical power is equal to 1 minus the probability of a Type II error.

***p*-value:** A form of a test statistic. It is associated with a specific test statistic and a structure of the rejection region. The p-value is equal to the significance level of the test in which the observed value of the statistic serves as the threshold.

Discuss in the forum

In statistical thinking there is a tenancy towards conservatism. The investigators, enthusiastic to obtain positive results, may prefer favorable conclusions and may tend over-interpret the data. It is the statistician's role to add to the objectivity in the interpretation of the data and to advocate caution.

On the other hand, the investigators may say that conservatism and science are incompatible. If one is too cautious, if one is always protecting oneself against the worst-case scenario, then one will not be able to make bold new discoveries.

Which of the two approach do you prefer?

When you formulate your answer to this question it may be useful to recall cases in your past in which you where required to analyze data or you were exposed to other people's analysis. Could the analysis benefit or be harmed by either of the approaches?

For example, many scientific journal will tend to reject a research paper unless the main discoveries are statistically significant (p-value $< 5\%$). Should one not publish also results that show a significance level of 10%?

Formulas:

- Test Statistic for Expectation: $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$.
- Two-Sided Test: Reject H_0 if $\{|t| >$ `qt(0.975,n-1)`$\}$.
- Greater Than: Reject H_0 if $\{t >$ `qt(0.95,n-1)`$\}$.
- Less Than: Reject H_0 if $\{t <$ `qt(0.05,n-1)`$\}$.