Chapter 15

A Bernoulli Response

15.1 Student Learning Objectives

Chapters 13 and 14 introduced statistical inference that involves a response and an explanatory variable that may affect the distribution of the response. In both chapters the response was numeric. The two chapters differed in the data type of the explanatory variable. In Chapter 13 the explanatory variable was a factor with two levels that splits the sample into two sub-samples. In Chapter 14 the explanatory variable was numeric and produced, together with the response, a linear trend. The aim in this chapter is to consider the case where the response is a Bernoulli variable. Such a variable may emerge as the indicator of the occurrence of an event associated with the response or as a factor with two levels. The explanatory variable is a factor with two levels in one case or a numerical variable in the other case.

Specifically, when the explanatory variable is a factor with two levels then we may use the function "`prop.test`". This function was used in Chapter 12 for the analysis of the probability of an event in a single sample. Here we use it in order to compare between two sub-samples. This is similar to the way the function "`t.test`" was used for a numeric response for both a single sample and for the comparison between sub-samples. For the case where the explanatory variable is numeric we may use the function "`glm`", acronym for *Generalized Linear Model*, in order to fit an appropriate regression model to the data.

By the end of this chapter, the student should be able to:

- Produce mosaic plots of the response and the explanatory variable.

- Apply the function "`prop.test`" in order to compare the probability of an event between two sub-populations

- Define the logistic regression model that relates the probability of an event in the response to a numeric explanatory variable.

- Fit the logistic regression model to data using the function "`glm`" and produce statistical inference on the fitted model.

15.2 Comparing Sample Proportions

In this chapter we deal with a Bernoulli response. Such a response has two levels, "`TRUE`" or "`FALSE`"[1], and may emerges as the indicator of an event. Else, it may be associated with a factor with two levels and correspond to the indication of one of the two levels. Such response was considered in Chapters 11 and 12 where confidence intervals and tests for the probability of an event where discussed in the context of a single sample. In this chapter we discuss the investigation of relations between a response of this form and an explanatory variable.

We start with the case where the explanatory variable is a factor that has two levels. These levels correspond to two sub-populations (or two settings). The aim of the analysis is to compare between the two sub-populations (or between the two settings) the probability of the even.

The discussion in this section is parallel to the discussion in Section 13.3. That section considered the comparison of the expectation of a numerical response between two sub-populations. We denoted these sub-populations a and b with expectations $\mathrm{E}(X_a)$ and $\mathrm{E}(X_b)$, respectively. The inference used the average $\bar{X}_a$, which was based on a sub-sample of size n_a, and the average $\bar{X}_b$, which was based on the other sub-sample of size n_b. The sub-samples variances S^2_a and S^2_b participated in the inference as well. The application of a test for the equality of the expectations and a confidence interval where produced by the application of the function "`t.test`" to the data.

The inference problem, which is considered in this chapter, involves an event. This event is being examined in two different settings that correspond to two different sub-population a and b. Denote the probabilities of the event in each of the sub-populations by p_a and p_b. Our concern is the statistical inference associated with the comparison of these two probabilities to each other.

Natural estimators of the probabilities are $\hat{P}_a$ and $\hat{P}_b$, the sub-samples proportions of occurrence of the event. These estimators are used in order to carry out the inference. Specifically, we consider here the construction of a confidence interval for the difference $p_a - p_b$ and a test of the hypothesis that the probabilities are equal.

The methods for producing the confidence intervals for the difference and for testing the null hypothesis that the difference is equal to zero are similar is principle to the methods that were described in Section 13.3 for making parallel inferences regarding the relations between expectations. However, the derivations of the tools that are used in the current situation are not identical to the derivations of the tools that were used there. The main differences between the two cases is the replacement of the sub-sample averages by the sub-sample proportions, a difference in the way the standard deviation of the statistics are estimated, and the application of a continuity correction. We do not discuss in this chapter the theoretical details associated with the derivations. Instead, we demonstrate the application of the inference in an example.

The variable "`num.of.doors`" in the data frame "`cars`" describes the number of doors a car has. This variable is a factor with two levels, "`two`" and "`four`". We treat this variable as a response and investigate its relation to explanatory variables. In this section the explanatory variable is a factor with two levels and in the next section it is a numeric variable. Specifically, in this

[1] The levels are frequently coded as 1 or 0, "success" or "failure", or any other pair of levels.

section we use the factor "fuel.type" as the explanatory variable. Recall that this variable identified the type of fuel, diesel or gas, that the car uses. The aim of the analysis is to compare the proportion of cars with four doors between cars that run on diesel and cars that run on gas.

Let us first summarize the data in a 2 × 2 frequency table. The function "table" may be used in order to produce such a table:

```
> cars <- read.csv("cars.csv")
> table(cars$fuel.type,cars$num.of.doors)

         four two
  diesel   16   3
  gas      98  86
```

When the function "table" is applied to a combination of two factors then the output is a table of joint frequencies. Each entry in the table contains the frequency in the sample of the combination of levels, one from each variable, that is associated with the entry. For example, there are 16 cars in the data set that have the level "four" for the variable "num.of.doors" and the level "diesel" for the variable "fuel.type". Likewise, there are 3 cars that are associated with the combination "two" and "diesel". The total number of entries to the table is $16 + 3 + 98 + 86 = 203$, which is the number of cars in the data set, minus the two missing values in the variable "num.of.doors".

A graphical representation of the relation between the two factors can be obtained using a mosaic plot. This plot is produced when the input to the function "plot" is a formula where both the response and the explanatory variables are factors:

```
> plot(num.of.doors ~ fuel.type,data=cars)
```

The resulting mosaic plot is presented in Figure 15.1.

The box plot describes the distribution of the explanatory variable and the distribution of the response for each level of the explanatory variable. In the current example the explanatory variable is the factor "fuel" that has 2 levels. The two levels of this variable, "diesel" and "gas", are given at the x-axis. A vertical rectangle is associated with each level. These 2 rectangles split the total area of the square. The total area of the square represents the total relative frequency (which is equal to 1). Consequently, the area of each rectangle represents the relative frequency of the associated level of the explanatory factor.

A rectangle associated with a given level of the explanatory variable is further divided into horizontal sub-rectangles that are associated with the response factor. In the current example each darker rectangle is associated with the level "four" of the response "num.of.door" and each brighter rectangle is associated with the level "two". The relative area of the horizontal rectangles within each vertical rectangle represent the relative frequency of the levels of the response within each subset associated with the level of the explanatory variable.

Looking at the plot one may appreciate the fact that diesel cars are less frequent than cars that run on gas. The graph also displays the fact that the relative frequency of cars with four doors among diesel cars is larger than the relative frequency of four doors cars among cars that run on gas.

The function "prop.test" may be used in order test the hypothesis that, at the population level, the probability of the level "four" of the response within the

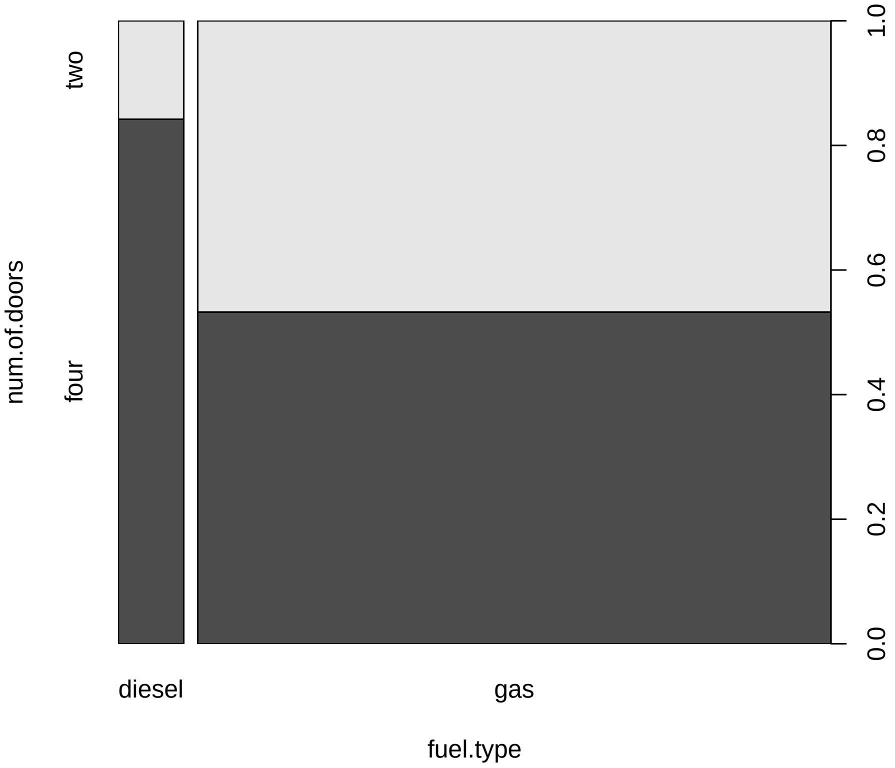

Figure 15.1: Number of Doors versus Fuel Type

sub-population of diesel cars (the height of the leftmost darker rectangle in the theoretic mosaic plot that is produced for the entire population) is equal to the probability of the same level of the response with in the sub-population of cars that run on gas (the height of the rightmost darker rectangle in that theoretic mosaic plot). Specifically, let us test the hypothesis that the two probabilities of the level "four", one for diesel cars and one for cars that run on gas, are equal to each other.

The output of the function "`table`" may serve as the input to the function "`prop.test`"[2]. The Bernoulli response variable should be the second variable in the input to the table whereas the explanatory factor is the first variable in the table. When we apply the test to the data we get the report:

```
> prop.test(table(cars$fuel.type,cars$num.of.doors))
```

[2] The function "`prop.test`" was applied in Section 12.4 in order to test that the probability of an event is equal to a given value ("`p = 0.5`" by default). The input to the function was a pair of numbers: the total number of successes and the sample size. In the current application the input is a 2 × 2 table. When applied to such input the function carries out a test of the equality of the probability of the first column between the rows of the table.

```
2-sample test for equality of proportions with continuity correction

data:  table(cars$fuel.type, cars$num.of.doors)
X-squared = 5.5021, df = 1, p-value = 0.01899
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1013542 0.5176389
sample estimates:
   prop 1    prop 2
0.8421053 0.5326087
```

The two sample proportions of cars with four doors among diesel and gas cars are presented at the bottom of the report and serve as estimates of the sub-populations probabilities. Indeed, the relative frequency of cars with four doors among diesel cars is equal to $\hat{p}_a = 16/(16+3) = 16/19 = 0.8421053$. Likewise, the relative frequency of cars with four doors among cars that ran on gas is equal to $\hat{p}_b = 98/(98+86) = 98/184 = 0.5326087$. The confidence interval for the difference in the probability of a car with four doors between the two sub-populations, $p_a - p_b$, is reported under the title "`95 percent confidence interval`" and is given as $[0.1013542, 0.5176389]$.

The null hypothesis, which is the subject of this test, is $H_0 : p_a = p_b$. This hypothesis is tested against the two-sided alternative hypothesis $H_1 : p_a \neq p_b$. The test itself is based on a test statistic that obtains the value `X-squared = 5.5021`. This test statistic corresponds essentially to the deviation between the estimated value of the parameter (the difference in sub-sample proportions of the event) and the theoretical value of the parameter ($p_a - p_b = 0$). This deviation is divided by the estimated standard deviation and the ratio is squared. The statistic itself is produced via a continuity correction that makes its null distribution closer to the limiting chi-square distribution on one degree of freedom. The p-value is computed based on this limiting chi-square distribution.

Notice that the computed p-value is equal to `p-value = 0.01899`. This value is smaller than 0.05. Consequently, the null hypothesis is rejected at the 5% significance level in favor of the alternative hypothesis. This alternative hypothesis states that the sub-populations probabilities are different from each other.

15.3 Logistic Regression

In the previous section we considered a Bernoulli response and a factor with two levels as an explanatory variable. In this section we use a numeric variable as the explanatory variable. The discussion in this section is parallel to the discussion in Chapter 14 that presented the topic of linear regression. However, since the response is not of the same form, it is the indicator of a level of a factor and not a regular numeric response, then the tools the are used are different. Instead of using linear regression we use another type of regression that is called *Logistic Regression.*

Recall that linear regression involved fitting a straight line to the scatter plot of data points. This line corresponds to the expectation of the response as a function of the explanatory variable. The estimated coefficients of this line are computed from the data and used for inference.

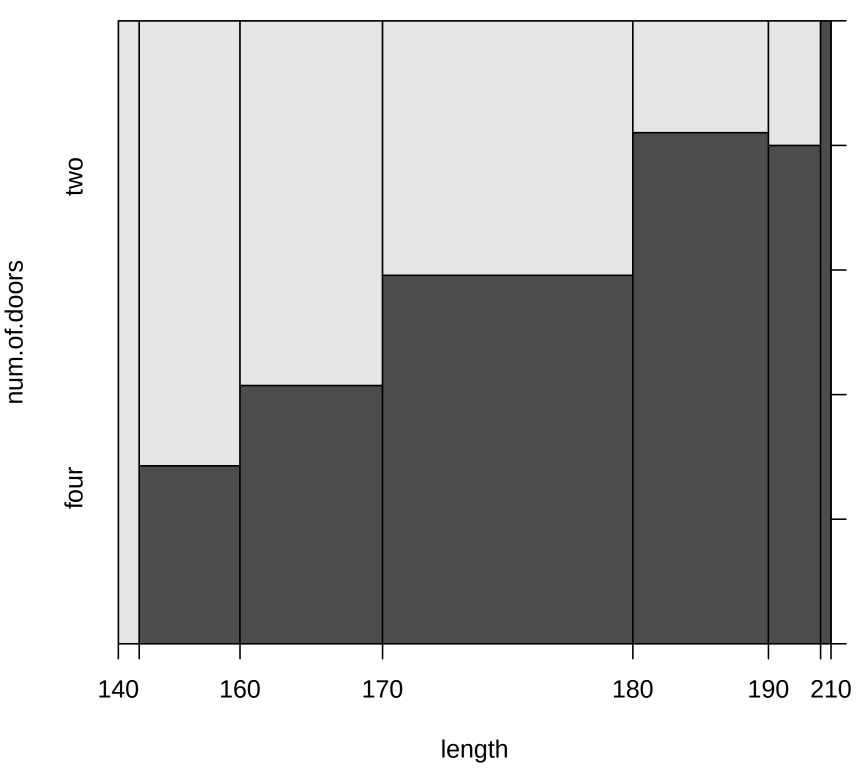

Figure 15.2: Number of Doors Versus Fuel Type

In logistic regression, instead of the consideration of the expectation of a numerical response, one considers the probability of an event associated with the response. This probability is treated as a function of the explanatory variable. Parameters that determine this function are estimated from the data and are used for inference regarding the relation between the explanatory variable and the response. Again, we do not discuss the theoretical details involved in the derivation of logistic regression. Instead, we apply the method to an example.

We consider the factor "`num.of.doors`" as the response and the probability of a car with four doors as the probability of the response. The length of the car will serve as the explanatory variable. Measurements of lengths of the cars are stored in the variable "`length`" in the data frame "`cars`".

First, let us plot the relation between the response and the explanatory variable:

```
> plot(num.of.doors ~ length,data=cars)
```

The plot that is produced by the given expression is displayed in Figure 15.2. It is a type of a mosaic plot and it is produced when the input to the function "`plot`" is a formula with a factor as a response and a numeric variable as the

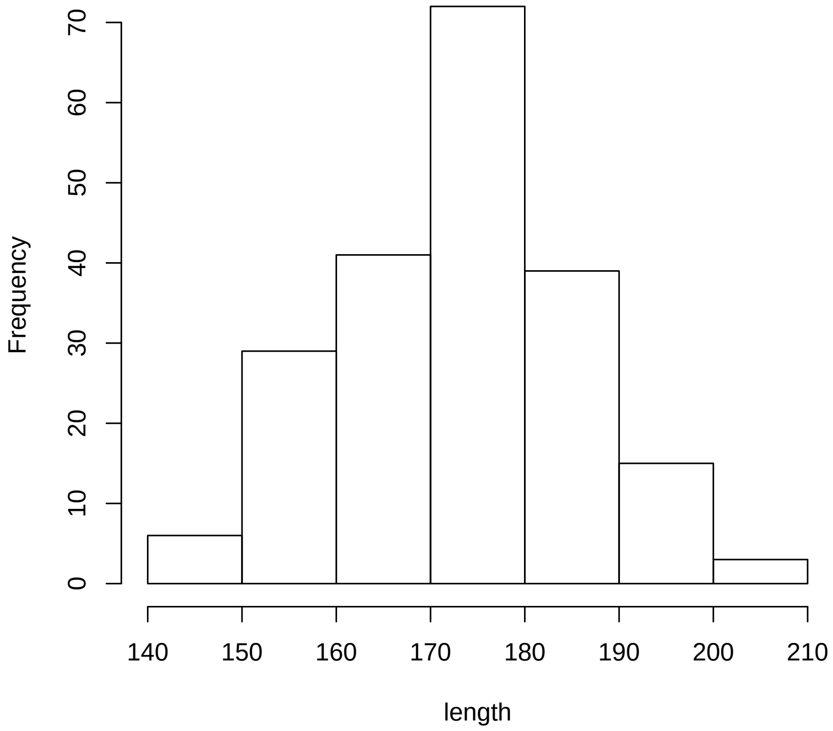

Figure 15.3: Histogram of the Length of Cars

explanatory variable. The plot presents, for interval levels of the explanatory variable, the relative frequencies of each interval. It also presents the relative frequency of the levels of the response within each interval level of the explanatory variable.

In order to get a better understanding of the meaning of the given mosaic plot one may consider the histogram of the explanatory variable. This histogram is presented in Figure 15.3. The histogram involves the partition of the range of variable length into intervals. These interval are the basis for rectangles. The height of the rectangles represent the frequency of cars with lengths that fall in the given interval.

The mosaic plot in Figure 15.2 is constructed on the basis of this histogram. The x-axis in this plot corresponds to the explanatory variable "`length`". The total area of the square in the plot is divided between 7 vertical rectangles. These vertical rectangles correspond to the 7 rectangles in the histogram of Figure 15.3, turn on their sides. Hence, the width of each rectangle in Figure 15.2 correspond to the hight of the parallel rectangle in the histogram. Consequently, the area of the vertical rectangles in the mosaic plot represents the relative frequency of the associated interval of values of the explanatory variable.

The rectangle that is associated with each interval of values of the explanatory variable is further divided into horizontal sub-rectangles that are associated with the response factor. In the current example each darker rectangle is associated with the level "`four`" of the response "`num.of.door`" and each brighter rectangle is associated with the level "`two`". The relative area of the horizontal rectangles within each vertical rectangle represent the relative frequency of the levels of the response within each interval of values of the explanatory variable.

From the examination of the mosaic plot one may identify relations between the explanatory variable and the relative frequency of an identified level of the response. In the current example one may observe that the relative frequency of the cars with four doors is, overall, increasing with the increase in the length of cars.

Logistic regression is a method for the investigation of relations between the probability of an event and explanatory variables. Specifically, we use it here for making inference on the number of doors as a response and the length of the car as the explanatory variable.

Statistical inference requires a statistical model. The statistical model in logistic regression relates the probability p_i, the probability of the event for observation i, to x_i, the value of the response for that observation. The relation between the two in given by the formula:

$$p_i = \frac{e^{a+b\cdot x_i}}{1 + e^{a+b\cdot x_i}} ,$$

where a and b are coefficients common to all observations. Equivalently, one may write the same relation in the form:

$$\log(p_i/[1-p_i]) = a + b \cdot x_i ,$$

that states that the relation between a (function of) the probability of the event and the explanatory variable is a linear trend.

One may fit the logistic regression to the data and test the null hypothesis by the use of the function "`glm`":

```
> fit.doors <- glm(num.of.doors=="four"~length,
+ family=binomial,data=cars)
> summary(fit.doors)

Call:
glm(formula = num.of.doors == "four" ~ length, family = binomial,
    data = cars)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1646  -1.1292   0.5688   1.0240   1.6673

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.14767    2.58693  -5.082 3.73e-07 ***
length        0.07726    0.01495   5.168 2.37e-07 ***
---
```

```
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 278.33  on 202  degrees of freedom
Residual deviance: 243.96  on 201  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 247.96

Number of Fisher Scoring iterations: 3
```

Generally, the function "glm" can be used in order to fit regression models in cases where the distribution of the response has special forms. Specifically, when the argument "family=binomial" is used then the model that is being used in the model of logistic regression. The formula that is used in the function involves a response and an explanatory variable. The response may be a sequence with logical "TRUE" or "FALSE" values as in the example[3]. Alternatively, it may be a sequence with "1" or "0" values, "1" corresponding to the event occurring to the subject and "0" corresponding to the event not occurring. The argument "data=cars" is used in order to inform the function that the variables are located in the given data frame.

The "glm" function is applied to the data and the fitted model is stored in the object "fit.doors".

A report is produced when the function "summary" is applied to the fitted model. Notice the similarities and the differences between the report presented here and the reports for linear regression that are presented in Chapter 14. Both reports contain estimates of the coefficients a and b and tests for the equality of these coefficients to zero. When the coefficient b, the coefficient that represents the slope, is equal to 0 then the probability of the event and the explanatory variable are unrelated. In the current case we may note that the null hypothesis $H_0 : b = 0$, the hypothesis that claims that there is no relation between the explanatory variable and the response, is clearly rejected (p-value 2.37×10^{-7}).

The estimated values of the coefficients are -13.14767 for the intercept a and 0.07726 for the slope b. One may produce confidence intervals for these coefficients by the application of the function "confint" to the fitted model:

```
> confint(fit.doors)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -18.50384373 -8.3180877
length        0.04938358  0.1082429
```

15.4 Solved Exercises

Question 15.1. This exercise deals with a comparison between Mediterranean diet and low-fat diet recommended by the American Heart Association in the

[3] The response is the output of the expression "num.of.doors=="four"". This expression produces logical values. "TRUE" when the car has 4 doors and "FALSE" when it has 2 doors.

context of risks for illness or death among patients that survived a heart attack[4]. 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 "Mediterranean Diet and Health" that is presented in that site.

The subjects, 605 survivors of a heart attack, were randomly assigned follow either (1) a diet close to the "prudent diet step 1" of the American Heart Association (AHA) or (2) a Mediterranean-type diet consisting of more bread and cereals, more fresh fruit and vegetables, more grains, more fish, fewer delicatessen food, less meat.

The subjects' diet and health condition were monitored over a period of four-year. Information regarding deaths, development of cancer or the development of non-fatal illnesses was collected. The information from this study is stored in the file "`diet.csv`". The file "`diet.csv`" contains two factors: "`health`" that describes the condition of the subject, either healthy, suffering from a non-fatal illness, suffering from cancer, or dead; and the "`type`" that describes the type of diet, either Mediterranean or the diet recommended by the AHA. The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/diet.csv`. Answer the following questions based on the data in the file:

1. Produce a frequency table of the two variable. Read off from the table the number of healthy subjects that are using the Mediterranean diet and the number of healthy subjects that are using the diet recommended by the AHA.

2. Test the null hypothesis that the probability of keeping healthy following an heart attack is the same for those that use the Mediterranean diet and for those that use the diet recommended by the AHA. Use a two-sided alternative and a 5% significance level.

3. Compute a 95% confidence interval for the difference between the two probabilities of keeping healthy.

Solution (to Question 15.1.1): First we save the file "`diet.csv`" in the working directory of `R` and read it's content. Then we apply the function "`table`" to the two variables in the file in order to obtain the requested frequency table:

```
> diet <- read.csv("diet.csv")
> table(diet$health,diet$type)

          aha med
  cancer   15   7
  death    24  14
  healthy 239 273
  illness  25   8
```

The resulting table has two columns and 4 rows. The third row corresponds to healthy subjects. Of these, 239 subjects used the AHA recommended diet and 273 used the Mediterranean diet. We may also plot this data using a mosaic plot:

[4] De Lorgeril, M., Salen, P., Martin, J., Monjaud, I., Boucher, P., Mamelle, N. (1998). Mediterranean Dietary pattern in a Randomized Trial. Archives of Internal Medicine, 158, 1181-1187.

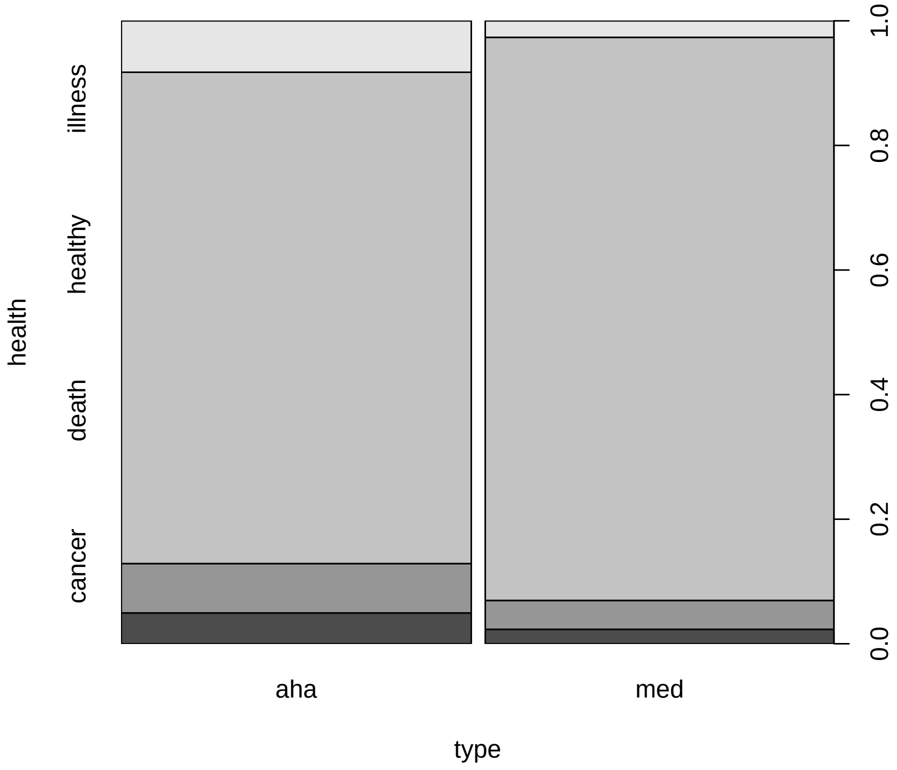

Figure 15.4: Health Condition Versus Type of Diet

```
> plot(health~type,data=diet)
```

The mosaic plot produced by the function “`plot`” is presented in Figure 15.4. Examining this plot one may appreciate the fact that the vast majority of the subjects were healthy and the relative proportion of healthy subjects among users of the Mediterranean diet is higher than the relative proportion among users of the AHA recommended diet.

Solution (to Question 15.1.2): In order to test the hypothesis that the probability of keeping healthy following an heart attack is the same for those that use the Mediterranean diet and for those that use the diet recommended by the AHA we create a 2 × 2. This table compares the response of being healthy or not to the type of diet as an explanatory variable. A sequence with logical components, “`TRUE`” for healthy and “`FALSE`” for not, is used as the response. Such a sequence is produced via the expression “`diet$health=="healthy"`”. The table may serve as input to the function “`prop.test`”:

```
> table(diet$health=="healthy",diet$type)
```

```
        aha med
  FALSE  64  29
  TRUE  239 273

> prop.test(table(diet$health=="healthy",diet$type))

2-sample test for equality of proportions with continuity correction

data:  table(diet$health == "healthy", diet$type)
X-squared = 14.5554, df = 1, p-value = 0.0001361
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1114300 0.3313203
sample estimates:
   prop 1    prop 2
0.6881720 0.4667969
```

The function “`prop.test`” conducts the test that compares between the probabilities of keeping healthy. In particular, the computed p-value for the test is 0.0001361, which is less than 0.05. Therefore, we reject the null hypothesis that both diets have the same effect on the chances of remaining healthy following an heart attack.

Solution (to Question 15.1.3): The confidence interval for the difference in probabilities is equal to $[0.1114300, 0.3313203]$. The point estimation of the difference between the probabilities is $\hat{p}_a - \hat{p}_b = 0.6881720 - 0.4667969 \approx 0.22$ in favor of a Mediterranean diet. The confidence interval proposes that a difference as low as 0.11 or as high as 0.33 are not excluded by the data.

Question 15.2. Cushing’s syndrome disorder results from a tumor (adenoma) in the pituitary gland that causes the production of high levels of cortisol. The symptoms of the syndrome are the consequence of the elevated levels of this steroid hormone in the blood. The syndrome was first described by Harvey Cushing in 1932.

The file “`coshings.csv`" contains information on 27 patients that suffer from Cushing’s syndrome[5]. The three variables in the file are “`tetra`”, “`pregn`”, and “`type`”. The factor “`type`” describes the underlying type of syndrome, coded as “`a`”, (adenoma), ”`b`” (bilateral hyperplasia), “`c`” (carcinoma) or “`u`” for unknown. The variable “`tetra`” describe the level of urinary excretion rate (mg/24hr) of Tetrahydrocortisone, a type of steroid, and the variable “`pregn`” describes urinary excretion rate (mg/24hr) of Pregnanetriol, another type of steroid. The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/coshings.csv`. Answer the following questions based on the information in this file:

1. Plot the histogram of the variable “`tetra`” and the mosaic plot that describes the relation between the variable “`type`” as a response and the variable “`tetra`”. What is the information that is conveyed by the second vertical triangle from the right (the third from the left) in the mosaic plot.

[5] The source of the data is the data file “`Cushings`” from the package “`MASS`” in `R`.

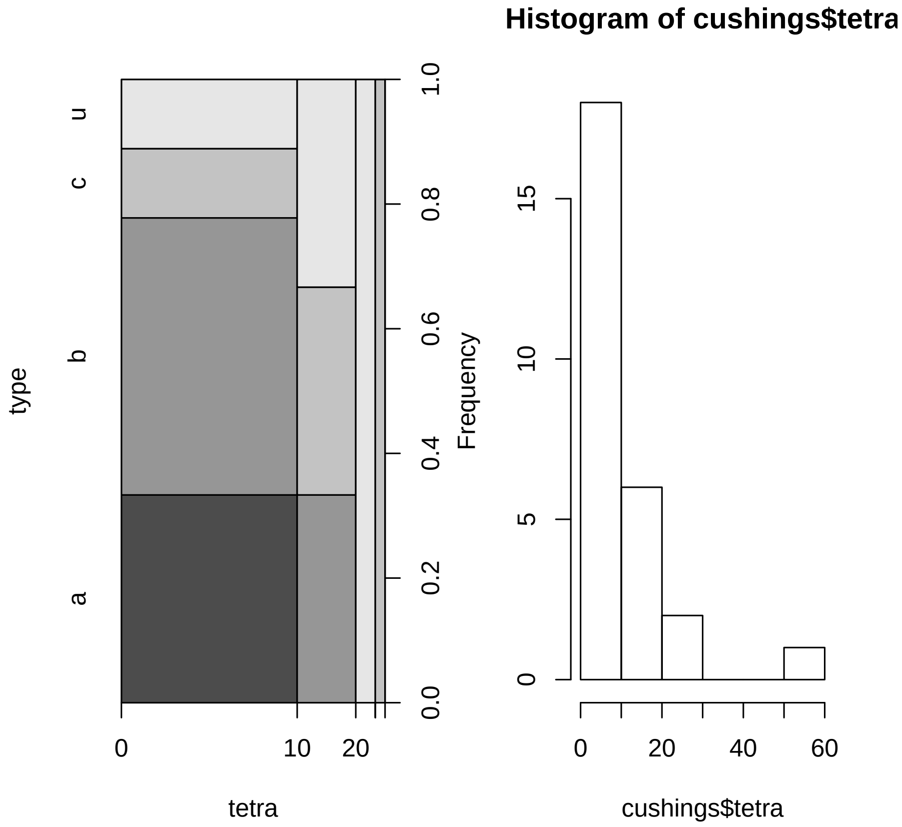

Figure 15.5: Health Condition Versus Type of Diet

2. Test the null hypothesis that there is no relation between the variable "`tetra`" as an explanatory variable and the indicator of the type being equal to "`b`" as a response. Compute a confidence interval for the parameter that describes the relation.

3. Repeat the analysis from 2 using only the observations for which the type is known. (Hint: you may fit the model to the required subset by the inclusion of the argument "`subset=(type!="u")`" in the function that fits the model.) Which of the analysis do you think is more appropriate?

Solution (to Question 15.2.1): We save the data of the file in a data frame by the name "`cushings`", produce a mosaic plot with the function "`plot`" and an histogram with the function "`hist`":

```
> cushings <- read.csv("cushings.csv")
> plot(type~tetra,data=cushings)
> hist(cushings$tetra)
```

The mosaic plot describes the distribution of the 4 levels of the response within the different intervals of values of the explanatory variable. The intervals coincide with the intervals that are used in the construction of the histogram. In particular, the third vertical rectangle from the left in the mosaic is associated with the third interval from the left in the histogram[6]. This interval is associated with the range of values between 20 and 30. The height of the given interval in the histogram is 2, which is the number of patients with "`terta`" levels that belong to the interval.

There are 4 shades of *grey* in the first vertical rectangle from the left. Each shade is associated with a different level of the response. The lightest shade of grey, the upmost one, is associated with the level "`u`". Notice that this is also the shade of grey of the entire third vertical rectangle from the left. The conclusion is that the 2 patients that are associated with this rectangle have Tetrahydrocortisone levels between 2 and 30 and have an unknown type of syndrome.

Solution (to Question 15.2.2): We fit the logistic regression to the entire data in the data frame "`cushings`" using the function "`glm`", with the "`family=binomial`" argument. The response is the indicator that the type is equal to "`b`". The fitted model is saved in an object called "`cushings.fit.all`". The application of the function "`summary`" to the fitted model produces a report that includes the test of the hypothesis of interest:

```
> cushings.fit.all <- glm((type=="b")~tetra,family=binomial,
+ data=cushings)
> summary(cushings.fit.all)

Call:
glm(formula = (type == "b") ~ tetra, family = binomial,
    data = cushings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0924  -1.0461  -0.8652   1.3427   1.5182

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12304    0.61330  -0.201    0.841
tetra       -0.04220    0.05213  -0.809    0.418

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.594  on 26  degrees of freedom
Residual deviance: 34.739  on 25  degrees of freedom
AIC: 38.739

Number of Fisher Scoring iterations: 4
```

The test of interest examines the coefficient that is associated wit the explanatory variable "`tetra`". The estimated value of this parameter is −0.04220. The

[6] This is also the third interval from the left in the histogram. However, since the second and third intervals, counting from the right, in the histogram are empty, it turns out that the given interval is the second rectangle from the right in the mosaic plot.

p-value for testing that the coefficient is 0 is equal to 0.418. Consequently, since the p-value is larger than 0.05, we do not reject the null hypothesis that states that the response and the explanatory variable are statistically unrelated.

Confidence intervals may be computed by applying the function "`confint`" to the fitted model:

```
> confint(cushings.fit.all)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -1.2955624 1.18118256
tetra       -0.1776113 0.04016772
```

Specifically, the confidence interval for the coefficient that is associated with the explanatory variable is equal to $[-0.1776113, 0.04016772]$

Solution (to Question 15.2.3): If we want to fit the logistic model to a partial subset of the data, say all the observations with values of the response other that "`u`", we may apply the argument "`subset`"[7]. Specifically, adding the expression "`subset=(type!="u")`" would do the job[8]. We repeat the same analysis as before. The only difference is the addition of the given expression to the function that fits the model to the data. The fitted model is saved in an object we call "`cushings.fit.known`":

```
> cushings.fit.known <- glm((type=="b")~tetra,family=binomial,
+ data=cushings,subset=(type!="u"))
> summary(cushings.fit.known)

Call:
glm(formula = (type == "b") ~ tetra, family = binomial,
    data = cushings, subset = (type != "u"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2078  -1.1865  -0.7548   1.2033   1.2791

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.11457    0.59947   0.191    0.848
tetra       -0.02276    0.04586  -0.496    0.620

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29.065  on 20  degrees of freedom
Residual deviance: 28.789  on 19  degrees of freedom
AIC: 32.789
```

[7] This argument may be used in other functions. For example, it may be used in the function "`lm`" that fits the linear regression.

[8] The value of the argument "`subset`" is a sequence with logical components that indicate which of the observations to include in the analysis. This sequence is formed with the aid of "`!=`", which corresponds to the relation "not equal to". The expression "`type!="u"`" indicates all observations with a "`type`" value not equal to "`u`".

```
Number of Fisher Scoring iterations: 4
```

The estimated value of the coefficient when considering only subject with a known type of the syndrome is slightly changed to -0.02276. The new p-value, which is equal to 0.620, is larger than 0.05. Hence, yet again, we do not reject the null hypothesis.

```
> confint(cushings.fit.known)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -1.0519135 1.40515473
tetra       -0.1537617 0.06279923
```

For the modified confidence interval we apply the function "`confint`". We get now $[-0.1537617, 0.06279923]$ as a confidence interval for the coefficient of the explanatory variable.

We started with the fitting the model to all the observations. Here we use only the observations for which the type of the syndrome is known. The practical implication of using all observations in the fit is equivalent to announcing that the type of the syndrome for observations of an unknown type is not type "`b`". This is not appropriate and may introduce bias, since the type may well be "`b`". It is more appropriate to treat the observations associated with the level "`u`" as missing observations and to delete them from the analysis. This approach is the approach that was used in the second analysis.

Glossary

Mosaic Plot: A plot that describes the relation between a response factor and an explanatory variable. Vertical rectangles represent the distribution of the explanatory variable. Horizontal rectangles within the vertical ones represent the distribution of the response.

Logistic Regression: A type of regression that relates between an explanatory variable and a response of the form of an indicator of an event.

Discuss in the forum

In the description of the statistical models that relate one variable to the other we used terms that suggest a causality relation. One variable was called the "explanatory variable" and the other was called the "response". One may get the impression that the explanatory variable is the cause for the statistical behavior of the response. In negation to this interpretation, some say that all that statistics does is to examine the joint distribution of the variables, but casuality cannot be inferred from the fact that two variables are statistically related. What do you think? Can statistical reasoning be used in the determination of casuality?

As part of your answer in may be useful to consider a specific situation where the determination of casuality is required. Can any of the tools that were discussed in the book be used in a meaningful way to aid in the process of such determination?

Notice that the last 3 chapters dealt with statistical models that related an explanatory variable to a response. We considered tools that can be used when both variable are factors and when both are numeric. Other tools may be used when one of the variables is a factor and the other is numeric. An analysis that involves one variable as the response and the other as explanatory variable can be reversed, possibly using a different statistical tool, with the roles of the variables exchanged. Usually, a significant statistical finding will be still significant when the roles of a response and an explanatory variable are reversed.

Formulas:

- Logistic Regression, (Probability): $p_i = \frac{e^{a+b\cdot x_i}}{1+e^{a+b\cdot x_i}}$.
- Logistic Regression, (Predictor): $\log(p_i/[1-p_i]) = a + b\cdot x_i$.

Chapter 16

Case Studies

16.1 Student Learning Objective

This chapter concludes this book. We start with a short review of the topics that were discussed in the second part of the book, the part that dealt with statistical inference. The main part of the chapter involves the statistical analysis of 2 case studies. The tools that will be used for the analysis are those that were discussed in the book. We close this chapter and this book with some concluding remarks. By the end of this chapter, the student should be able to:

- Review the concepts and methods for statistical inference that were presented in the second part of the book.

- Apply these methods to requirements of the analysis of real data.

- Develop a resolve to learn more statistics.

16.2 A Review

The second part of the book dealt with statistical inference; the science of making general statement on an entire population on the basis of data from a sample. The basis for the statements are theoretical models that produce the sampling distribution. Procedures for making the inference are evaluated based on their properties in the context of this sampling distribution. Procedures with desirable properties are applied to the data. One may attach to the output of this application summaries that describe these theoretical properties.

In particular, we dealt with two forms of making inference. One form was estimation and the other was hypothesis testing. The goal in estimation is to determine the value of a parameter in the population. Point estimates or confidence intervals may be used in order to fulfill this goal. The properties of point estimators may be assessed using the mean square error (MSE) and the properties of the confidence interval may be assessed using the confidence level.

The target in hypotheses testing is to decide between two competing hypothesis. These hypotheses are formulated in terms of population parameters. The decision rule is called a statistical test and is constructed with the aid of a test statistic and a rejection region. The default hypothesis among the two, is rejected if the test statistic falls in the rejection region. The major property a test must possess is a bound on the probability of a Type I error, the probability of erroneously rejecting the null hypothesis. This restriction is called the significance level of the test. A test may also be assessed in terms of it's statistical power, the probability of rightfully rejecting the null hypothesis.

Estimation and testing were applied in the context of single measurements and for the investigation of the relations between a pair of measurements. For single measurements we considered both numeric variables and factors. For numeric variables one may attempt to conduct inference on the expectation and/or the variance. For factors we considered the estimation of the probability of obtaining a level, or, more generally, the probability of the occurrence of an event.

We introduced statistical models that may be used to describe the relations between variables. One of the variables was designated as the response. The other variable, the explanatory variable, is identified as a variable which may affect the distribution of the response. Specifically, we considered numeric variables and factors that have two levels. If the explanatory variable is a factor with two levels then the analysis reduces to the comparison of two sub-populations, each one associated with a level. If the explanatory variable is numeric then a regression model may be applied, either linear or logistic regression, depending on the type of the response.

The foundations of statistical inference are the assumption that we make in the form of statistical models. These models attempt to reflect reality. However, one is advised to apply healthy skepticism when using the models. First, one should be aware what the assumptions are. Then one should ask oneself how reasonable are these assumption in the context of the specific analysis. Finally, one should check as much as one can the validity of the assumptions in light of the information at hand. It is useful to plot the data and compare the plot to the assumptions of the model.

16.3 Case Studies

Let us apply the methods that were introduced throughout the book to two examples of data analysis. Both examples are taken from the case studies of the Rice Virtual Lab in Statistics can be found in their Case Studies section. The analysis of these case studies may involve any of the tools that were described in the second part of the book (and some from the first part). It may be useful to read again Chapters 9–15 before reading the case studies.

16.3.1 Physicians' Reactions to the Size of a Patient

Overweight and obesity is common in many of the developed contrives. In some cultures, obese individuals face discrimination in employment, education, and relationship contexts. The current research, conducted by Mikki Hebl and Jingping Xu[1], examines physicians' attitude toward overweight and obese patients in comparison to their attitude toward patients who are not overweight.

[1] Hebl, M. and Xu, J. (2001). Weighing the care: Physicians' reactions to the size of a patient. International Journal of Obesity, 25, 1246-1252.

The experiment included a total of 122 primary care physicians affiliated with one of three major hospitals in the Texas Medical Center of Houston. These physicians were sent a packet containing a medical chart similar to the one they view upon seeing a patient. This chart portrayed a patient who was displaying symptoms of a migraine headache but was otherwise healthy. Two variables (the gender and the weight of the patient) were manipulated across six different versions of the medical charts. The weight of the patient, described in terms of Body Mass Index (BMI), was average (BMI = 23), overweight (BMI = 30), or obese (BMI = 36). Physicians were randomly assigned to receive one of the six charts, and were asked to look over the chart carefully and complete two medical forms. The first form asked physicians which of 42 tests they would recommend giving to the patient. The second form asked physicians to indicate how much time they believed they would spend with the patient, and to describe the reactions that they would have toward this patient.

In this presentation, only the question on how much time the physicians believed they would spend with the patient is analyzed. Although three patient weight conditions were used in the study (average, overweight, and obese) only the average and overweight conditions will be analyzed. Therefore, there are two levels of patient weight (average and overweight) and one dependent variable (time spent).

The data for the given collection of responses from 72 primary care physicians is stored in the file "`discriminate.csv`"[2]. We start by reading the content of the file into a data frame by the name "`patient`" and presenting the summary of the variables:

```
> patient <- read.csv("discriminate.csv")
> summary(patient)
    weight        time
 BMI=23:33   Min.   : 5.00
 BMI=30:38   1st Qu.:20.00
             Median :30.00
             Mean   :27.82
             3rd Qu.:30.00
             Max.   :60.00
```

Observe that of the 72 "patients", 38 are overweight and 33 have an average weight. The time spend with the patient, as predicted by physicians, is distributed between 5 minutes and 1 hour, with a average of 27.82 minutes and a median of 30 minutes.

It is a good practice to have a look at the data before doing the analysis. In this examination on should see that the numbers make sense and one should identify special features of the data. Even in this very simple example we may want to have a look at the histogram of the variable "`time`":

```
> hist(patient$time)
```

The histogram produced by the given expression is presented in Figure 16.1. A feature in this plot that catches attention is the fact that there is a high concventration of values in the interval between 25 and 30. Together with the

[2] The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/discriminate.csv`.

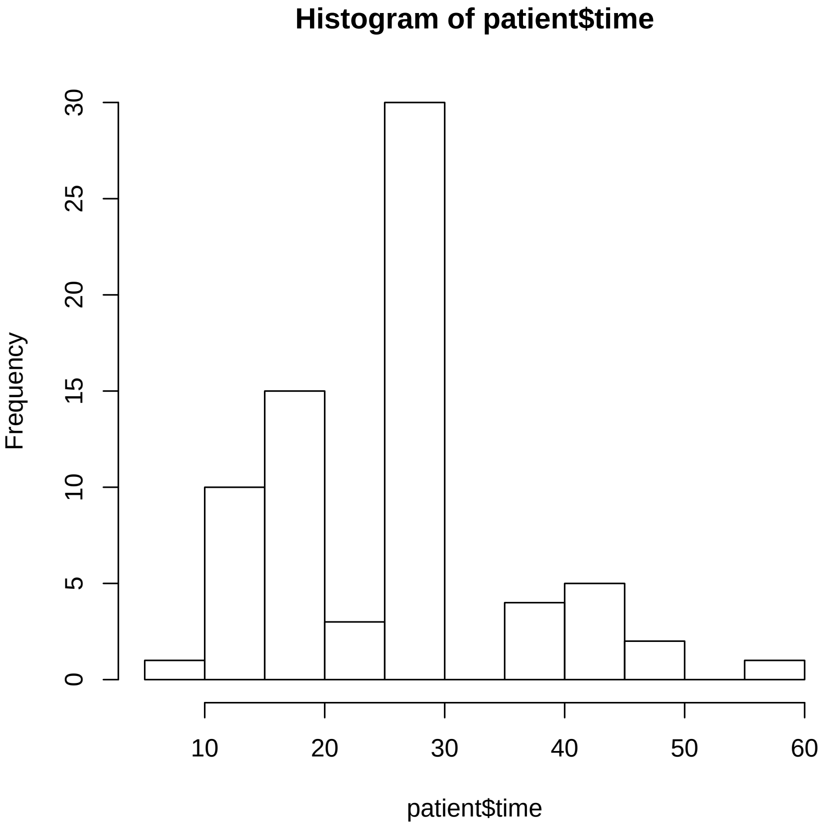

Figure 16.1: Histogram of "`time`"

fact that the median is equal to 30, one may suspect that, as a matter of fact, a large numeber of the values are actually equal to 30. Indeed, let us produce a table of the response:

```
> table(patient$time)

 5 15 20 25 30 40 45 50 60
 1 10 15  3 30  4  5  2  1
```

Notice that 30 of the 72 physicians marked "`30`" as the time they expect to spend with the patient. This is the middle value in the range, and may just be the default value one marks if one just needs to complete a form and do not really place much importance to the question that was asked.

The goal of the analysis is to examine the relation between overweigh and the Doctor's response. The explanatory variable is a factor with two levels. The response is numeric. A natural tool to use in order to test this hypothesis is the t-test, which is implemented with the function "`t.test`".

First we plot the relation between the response and the explanatory variable and then we apply the test:

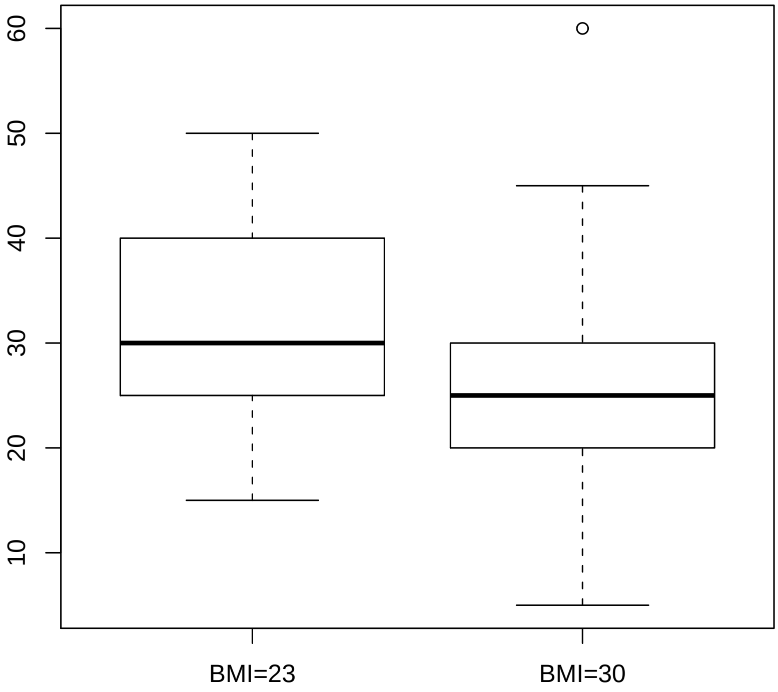

Figure 16.2: Time Versus Weight Group

```
> boxplot(time~weight,data=patient)
> t.test(time~weight,data=patient)

        Welch Two Sample t-test

data:  time by weight
t = 2.8516, df = 67.174, p-value = 0.005774
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.988532 11.265056
sample estimates:
mean in group BMI=23 mean in group BMI=30
            31.36364             24.73684
```

The box plots that describe the distribution of the response for each level of the explanatory variable are presented in Figure 16.2. Nothing seems problematic in this plot. The two distributions, as they are reflected in the box plots, look fairly symmetric.

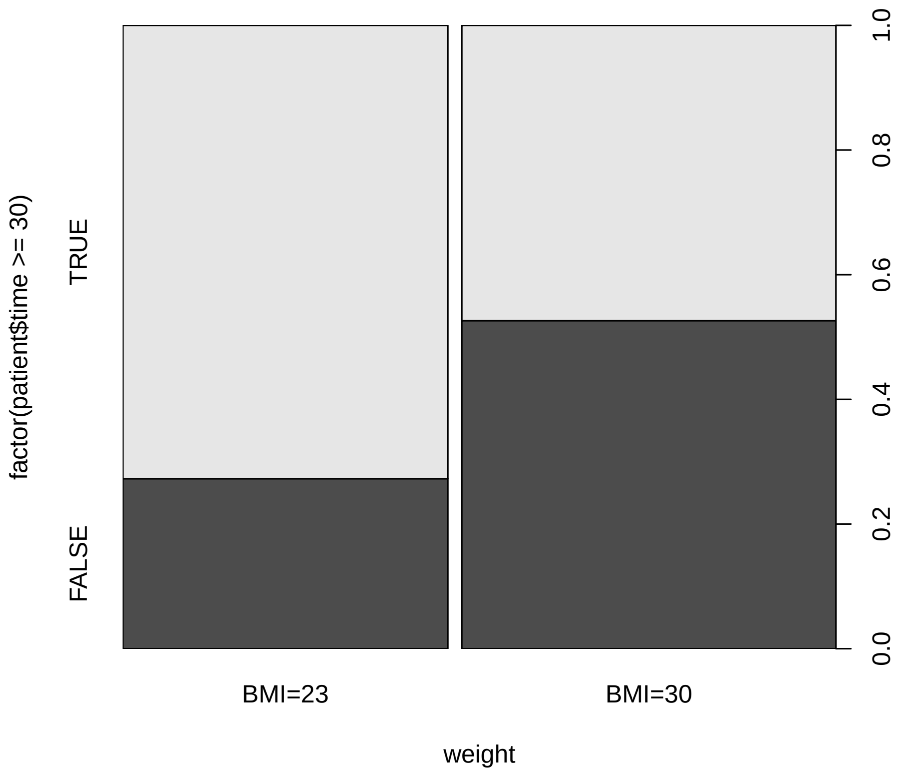

Figure 16.3: At Least 30 Minutes Versus Weight Group

When we consider the report that produced by the function “`t.test`” we may observe that the p-value is equal to 0.005774. This p-value is computed in testing the null hypothesis that the expectation of the response for both types of patients are equal against the two sided alternative. Since the p-value is less than 0.05 we do reject the null hypothesis.

The estimated value of the difference between the expectation of the response for a patient with BMI=23 and a patient with BMI=30 is $31.36364 - 24.73684 \approx 6.63$ minutes. The confidence interval is (approximately) equal to $[1.99, 11.27]$. Hence, it looks as if the physicians expect to spend more time with the average weight patients.

After analyzing the effect of the explanatory variable on the expectation of the response one may want to examine the presence, or lack thereof, of such effect on the variance of the response. Towards that end, one may use the function ”`var.test`”:

```
> var.test(time~weight,data=patient)

        F test to compare two variances
```

```
data:  time by weight
F = 1.0443, num df = 32, denom df = 37, p-value = 0.893
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5333405 2.0797269
sample estimates:
ratio of variances
          1.044316
```

In this test we do not reject the null hypothesis that the two variances of the response are equal since the p-value is larger than 0.05. The sample variances are almost equal to each other (their ratio is 1.044316), with a confidence interval for the ration that essentially ranges between 1/2 and 2.

The production of p-values and confidence intervals is just one aspect in the analysis of data. Another aspect, which typically is much more time consuming and requires experience and healthy skepticism is the examination of the assumptions that are used in order to produce the p-values and the confidence intervals. A clear violation of the assumptions may warn the statistician that perhaps the computed nominal quantities do not represent the actual statistical properties of the tools that were applied.

In this case, we have noticed the high concentration of the response at the value "30". What is the situation when we split the sample between the two levels of the explanatory variable? Let us apply the function "table" once more, this time with the explanatory variable included:

```
> table(patient$time,patient$weight)

     BMI=23 BMI=30
  5       0      1
  15      2      8
  20      6      9
  25      1      2
  30     14     16
  40      4      0
  45      4      1
  50      2      0
  60      0      1
```

Not surprisingly, there is still high concentration at that level "30". But one can see that only 2 of the responses of the "BMI=30" group are above that value in comparison to a much more symmetric distribution of responses for the other group.

The simulations of the significance level of the one-sample t-test for an Exponential response that were conducted in Question 12.2 may cast some doubt on how trustworthy are nominal p-values of the t-test when the measurements are skewed. The skewness of the response for the group "BMI=30" is a reason to be worry.

We may consider a different test, which is more robust, in order to validate the significance of our findings. For example, we may turn the response into a factor by setting a level for values larger or equal to "30" and a different level for values less than "30". The relation between the new response and the explanatory variable can be examined with the function "prop.test". We first plot and then test:

```
> plot(factor(patient$time>=30)~weight,data=patient)
> prop.test(table(patient$time>=30,patient$weight))

2-sample test for equality of proportions with continuity correction

data:  table(patient$time >= 30, patient$weight)
X-squared = 3.7098, df = 1, p-value = 0.05409
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.515508798 -0.006658689
sample estimates:
   prop 1    prop 2
0.3103448 0.5714286
```

The mosaic plot that presents the relation between the explanatory variable and the new factor is given in Figure 16.3. The level "TRUE" is associated with a value of the predicted time spent with the patient being 30 minutes or more. The level "FALSE" is associated with a prediction of less than 30 minutes.

The computed p-value is equal to 0.05409, that almost reaches the significance level of 5%[3]. Notice that the probabilities that are being estimated by the function are the probabilities of the level "FALSE". Overall, one may see the outcome of this test as supporting evidence for the conclusion of the t-test. However, the p-value provided by the t-test may over emphasize the evidence in the data for a significant difference in the physician attitude towards overweight patients.

16.3.2 Physical Strength and Job Performance

The next case study involves an attempt to develop a measure of physical ability that is easy and quick to administer, does not risk injury, and is related to how well a person performs the actual job. The current example is based on study by Blakely et al. [4], published in the journal Personnel Psychology.

There are a number of very important jobs that require, in addition to cognitive skills, a significant amount of strength to be able to perform at a high level. Construction worker, electrician and auto mechanic, all require strength in order to carry out critical components of their job. An interesting applied problem is how to select the best candidates from amongst a group of applicants for physically demanding jobs in a safe and a cost effective way.

The data presented in this case study, and may be used for the development of a method for selection among candidates, were collected from 147 individuals

[3] One may propose splinting the response into two groups, with one group being associated with values of "time" strictly *larger* than 30 minutes and the other with values less or equal to 30. The resulting p-value from the expression "prop.test(table(patient$time>30,patient$weight))" is 0.01276. However, the number of subjects in one of the cells of the table is equal only to 2, which is problematic in the context of the Normal approximation that is used by this test.

[4] Blakley, B.A., Quiñones, M.A., Crawford, M.S., and Jago, I.A. (1994). The validity of isometric strength tests. Personnel Psychology, 47, 247-274.

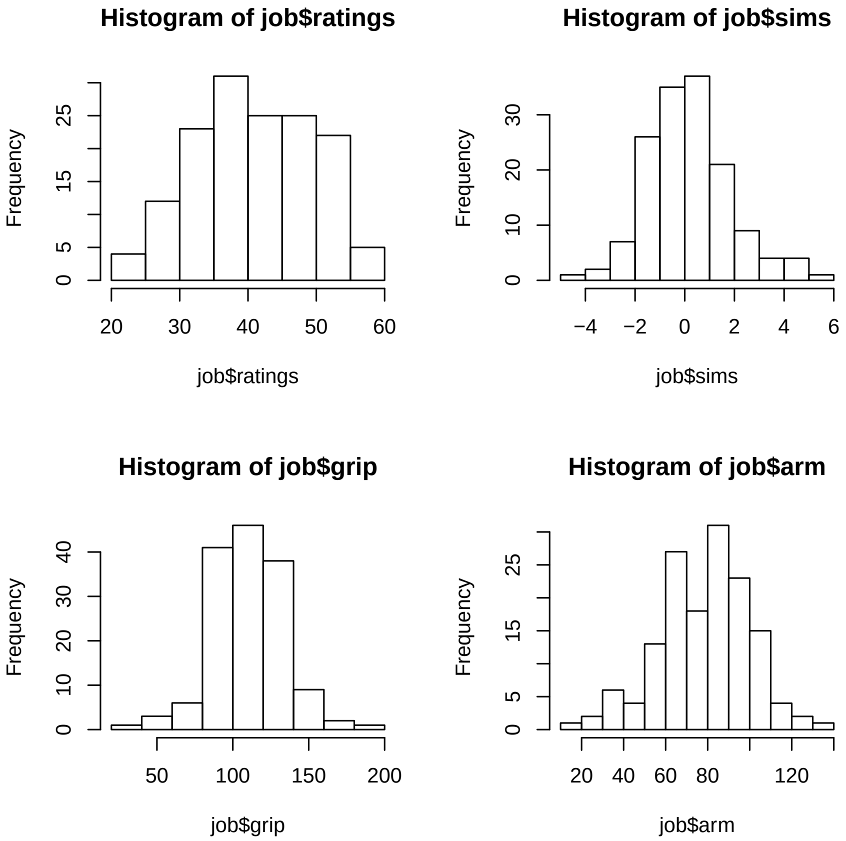

Figure 16.4: Histograms of Variables

working in physically demanding jobs. Two measures of strength were gathered from each participant. These included grip and arm strength. A piece of equipment known as the Jackson Evaluation System (JES) was used to collect the strength data. The JES can be configured to measure the strength of a number of muscle groups. In this study, grip strength and arm strength were measured. The outcomes of these measurements were summarized in two scores of physical strength called "`grip`" and "`arm`".

Two separate measures of job performance are presented in this case study. First, the supervisors for each of the participants were asked to rate how well their employee(s) perform on the physical aspects of their jobs. This measure is summarizes in the variable "`ratings`". Second, simulations of physically demanding work tasks were developed. The summary score of these simulations are given in the variable "`sims`". Higher values of either measures of performance indicates better performance.

The data for the 4 variables and 147 observations is stored in "`job.csv`"[5].

[5] The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/job.csv`.

We start by reading the content of the file into a data frame by the name "`job`", presenting a summary of the variables, and their histograms:

```
> job <- read.csv("job.csv")
> summary(job)
      grip             arm              ratings            sims        
 Min.   : 29.0   Min.   : 19.00   Min.   :21.60   Min.   :-4.1700  
 1st Qu.: 94.0   1st Qu.: 64.50   1st Qu.:34.80   1st Qu.:-0.9650  
 Median :111.0   Median : 81.50   Median :41.30   Median : 0.1600  
 Mean   :110.2   Mean   : 78.75   Mean   :41.01   Mean   : 0.2018  
 3rd Qu.:124.5   3rd Qu.: 94.00   3rd Qu.:47.70   3rd Qu.: 1.0700  
 Max.   :189.0   Max.   :132.00   Max.   :57.20   Max.   : 5.1700  
> hist(job$grip)
> hist(job$arm)
> hist(job$ratings)
> hist(job$sims)
```

All variables are numeric. Their histograms are presented in Figure 16.5. Examination of the 4 summaries and histograms does not produce interest findings. All variables are, more or less, symmetric with the distribution of the variable "`ratings`" tending perhaps to be more uniform then the other three.

The main analyses of interest are attempts to relate the two measures of physical strength "`grip`" and "`arm`" with the two measures of job performance, "`ratings`" and "`sims`". A natural tool to consider in this context is a linear regression analysis that relates a measure of physical strength as an explanatory variable to a measure of job performance as a response.

Let us consider the variable "`sims`" as a response. The first step is to plot a scatter plot of the response and explanatory variable, for both explanatory variables. To the scatter plot we add the line of regression. In order to add the regression line we fit the regression model with the function "`lm`" and then apply the function "`abline`" to the fitted model. The plot for the relation between the response and the variable "`grip`" is produced by the code:

```
> plot(sims~grip,data=job)
> sims.grip <- lm(sims~grip,data=job)
> abline(sims.grip)
```

The plot that is produced by this code is presented on the upper-left panel of Figure 16.5.

The plot for the relation between the response and the variable "`arm`" is produced by this code:

```
> plot(sims~arm,data=job)
> sims.arm <- lm(sims~arm,data=job)
> abline(sims.arm)
```

The plot that is produced by the last code is presented on the upper-right panel of Figure 16.5.

Both plots show similar characteristics. There is an overall linear trend in the relation between the explanatory variable and the response. The value of the response increases with the increase in the value of the explanatory variable

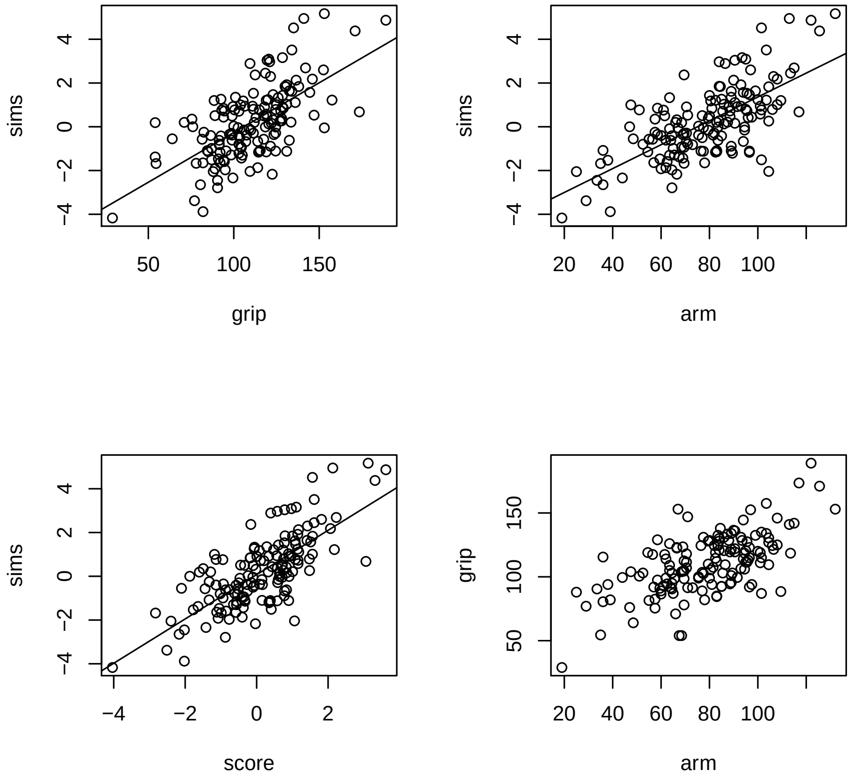

Figure 16.5: Scatter Plots and Regression Lines

(a positive slope). The regression line seems to follow, more or less, the trend that is demonstrated by the scatter plot.

A more detailed analysis of the regression model is possible by the application of the function "`summary`" to the fitted model. First the case where the explanatory variable is "`grip`":

```
> summary(sims.grip)

Call:
lm(formula = sims ~ grip, data = job)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9295 -0.8708 -0.1219  0.8039  3.3494

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.809675   0.511141   -9.41   <2e-16 ***
```

```
grip         0.045463   0.004535   10.03   <2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 1.295 on 145 degrees of freedom
Multiple R-squared: 0.4094,     Adjusted R-squared: 0.4053
F-statistic: 100.5 on 1 and 145 DF,  p-value: < 2.2e-16
```

Examination of the report reviles a clear statistical significance for the effect of the explanatory variable on the distribution of response. The value of R-squared, the ration of the variance of the response explained by the regression is 0.4094. The square root of this quantity, $\sqrt{0.4094} \approx 0.64$, is the proportion of the standard deviation of the response that is explained by the explanatory variable. Hence, about 64% of the variability in the response can be attributed to the measure of the strength of the grip.

For the variable "`arm`" we get:

```
> summary(sims.arm)

Call:
lm(formula = sims ~ arm, data = job)

Residuals:
     Min       1Q   Median       3Q      Max
-3.64667 -0.75022 -0.02852  0.68754  3.07702

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.095160   0.391745  -10.45   <2e-16 ***
arm          0.054563   0.004806   11.35   <2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 1.226 on 145 degrees of freedom
Multiple R-squared: 0.4706,     Adjusted R-squared: 0.467
F-statistic: 128.9 on 1 and 145 DF,  p-value: < 2.2e-16
```

This variable is also statistically significant. The value of R-squared is 0.4706. The proportion of the standard deviation that is explained by the strength of the are is $\sqrt{0.4706} \approx 0.69$, which is slightly higher than the proportion explained by the grip.

Overall, the explanatory variables do a fine job in the reduction of the variability of the response "`sims`" and may be used as substitutes of the response in order to select among candidates. A better prediction of the response based on the values of the explanatory variables can be obtained by combining the information in both variables. The production of such combination is not discussed in this book, though it is similar in principle to the methods of linear regression that are presented in Chapter 14. The produced score[6] takes the form:

$$\texttt{score} = -5.434 + 0.024 \cdot \texttt{grip} + 0.037 \cdot \texttt{arm}\ .$$

[6] The score is produced by the application of the function "`lm`" to *both* variables as explanatory variables. The code expression that can be used is "`lm(sims ~ grip + arm, data=job)`".

We use this combined score as an explanatory variable. First we form the score and plot the relation between it and the response:

```
> score <- -5.434 + 0.024*job$grip+ 0.037*job$arm
> plot(sims~score,data=job)
> sims.score <- lm(sims~score,data=job)
> abline(sims.score)
```

The scatter plot that includes the regression line can be found at the lower-left panel of Figure 16.5. Indeed, the linear trend is more pronounced for this scatter plot and the regression line a better description of the relation between the response and the explanatory variable. A summary of the regression model produces the report:

```
> summary(sims.score)

Call:
lm(formula = sims ~ score, data = job)

Residuals:
     Min       1Q   Median       3Q      Max
-3.18897 -0.73905 -0.06983  0.74114  2.86356

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07479    0.09452   0.791     0.43
score        1.01291    0.07730  13.104   <2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 1.14 on 145 degrees of freedom
Multiple R-squared: 0.5422,     Adjusted R-squared: 0.539
F-statistic: 171.7 on 1 and 145 DF,  p-value: < 2.2e-16
```

Indeed, the score is highly significant. More important, the R-squared coefficient that is associated with the score is 0.5422, which corresponds to a ratio of the standard deviation that is explained by the model of $\sqrt{0.5422} \approx 0.74$. Thus, almost 3/4 of the variability is accounted for by the score, so the score is a reasonable mean of guessing what the results of the simulations will be. This guess is based only on the results of the simple tests of strength that is conducted with the JES device.

Before putting the final seal on the results let us examine the assumptions of the statistical model. First, with respect to the two explanatory variables. Does each of them really measure a different property or do they actually measure the same phenomena? In order to examine this question let us look at the scatter plot that describes the relation between the two explanatory variables. This plot is produced using the code:

```
> plot(grip~arm,data=job)
```

It is presented in the lower-right panel of Figure 16.5. Indeed, one may see that the two measurements of strength are not independent of each other but tend

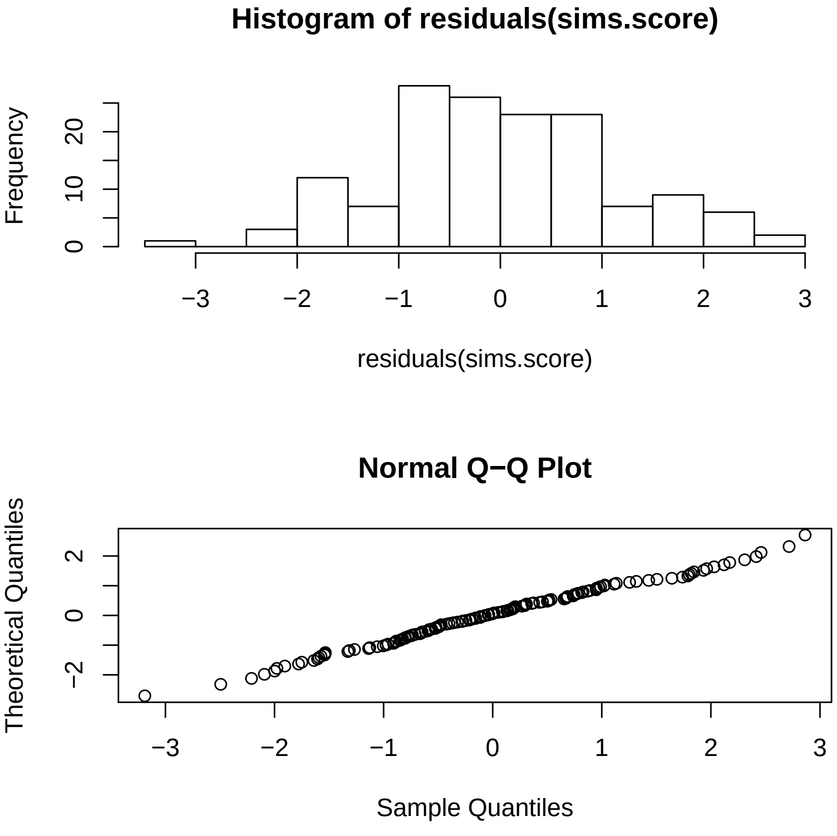

Figure 16.6: An Histogram and a QQ-Plot of Residuals

to produce an increasing linear trend. Hence, it should not be surprising that the relation of each of them with the response produces essentially the same goodness of fit. The computed score gives a slightly improved fit, but still, it basically reflects either of the original explanatory variables.

In light of this observation, one may want to consider other measures of strength that represents features of the strength not captures by these two variable. Namely, measures that show less joint trend than the two considered.

Another element that should be examined are the probabilistic assumptions that underly the regression model. We described the regression model only in terms of the functional relation between the explanatory variable and the expectation of the response. In the case of linear regression, for example, this relation was given in terms of a linear equation. However, another part of the model corresponds to the distribution of the measurements about the line of regression. The assumption that led to the computation of the reported p-values is that this distribution is Normal.

A method that can be used in order to investigate the validity of the Normal assumption is to analyze the residuals from the regression line. Recall that these residuals are computed as the difference between the observed value of the response and its estimated expectation, namely the fitted regression line. The residuals can be computed via the application of the function "`residuals`" to the fitted regression model.

Specifically, let us look at the residuals from the regression line that uses the score that is combined from the grip and arm measurements of strength. One may plot a histogram of the residuals:

```
> hist(residuals(sims.score))
```

The produced histogram is represented on the upper panel of Figure 16.6. The histogram portrays a symmetric distribution that my result from Normally distributed observations. A better method to compare the distribution of the residuals to the Normal distribution is to use the *Quantile-Quantile plot*. This plot can be found on the lower panel of Figure 16.6. We do not discuss here the method by which this plot is produced[7]. However, we do say that any deviation of the points from a straight line is indication of violation of the assumption of Normality. In the current case, the points seem to be on a single line, which is consistent with the assumptions of the regression model.

The next task should be an analysis of the relations between the explanatory variables and the other response "`ratings`". In principle one may use the same steps that were presented for the investigation of the relations between the explanatory variables and the response "`sims`". But of course, the conclusion may differ. We leave this part of the investigation as an exercise to the students.

16.4 Summary

16.4.1 Concluding Remarks

The book included a description of some elements of statistics, element that we thought are simple enough to be explained as part of an introductory course to statistics and are the minimum that is required for any person that is involved in academic activities of any field in which the analysis of data is required. Now, as you finish the book, it is as good time as any to say some words regarding the elements of statistics that are missing from this book.

One element is more of the same. The statistical models that were presented are as simple as a model can get. A typical application will required more complex models. Each of these models may require specific methods for estimation and testing. The characteristics of inference, e.g. significance or confidence levels, rely on assumptions that the models are assumed to possess. The user should be familiar with computational tools that can be used for the analysis of these more complex models. Familiarity with the probabilistic assumptions is required in order to be able to interpret the computer output, to diagnose possible divergence from the assumptions and to assess the severity of the possible effect of such divergence on the validity of the findings.

Statistical tools can be used for tasks other than estimation and hypothesis testing. For example, one may use statistics for prediction. In many applications it is important to assess what the values of future observations may be

[7] Generally speaking, the plot is composed of the empirical percentiles of the residuals, plotted against the theoretical percentiles of the standard Normal distribution. The current plot is produced by the expression "`qqnorm(residuals(sims.score))`".

and in what range of values are they likely to occur. Statistical tools such as regression are natural in this context. However, the required task is not testing or estimation the values of parameters, but the prediction of future values of the response.

A different role of statistics in the design stage. We hinted in that direction when we talked about in Chapter 11 about the selection of a sample size in order to assure a confidence interval with a given accuracy. In most applications, the selection of the sample size emerges in the context of hypothesis testing and the criteria for selection is the minimal power of the test, a minimal probability to detect a true finding. Yet, statistical design is much more than the determination of the sample size. Statistics may have a crucial input in the decision of how to collect the data. With an eye on the requirements for the final analysis, an experienced statistician can make sure that data that is collected is indeed appropriate for that final analysis. Too often is the case where researcher steps into the statistician's office with data that he or she collected and asks, when it is already too late, for help in the analysis of data that cannot provide a satisfactory answer to the research question the researcher tried to address. It may be said, with some exaggeration, that good statisticians are required for the final analysis only in the case where the initial planning was poor.

Last, but not least, is the theoretical mathematical theory of statistics. We tried to introduce as little as possible of the relevant mathematics in this course. However, if one seriously intends to learn and understand statistics then one must become familiar with the relevant mathematical theory. Clearly, deep knowledge in the mathematical theory of probability is required. But apart from that, there is a rich and rapidly growing body of research that deals with the mathematical aspects of data analysis. One cannot be a good statistician unless one becomes familiar with the important aspects of this theory.

I should have started the book with the famous quotation: "Lies, damned lies, and statistics". Instead, I am using it to end the book. Statistics can be used and can be misused. Learning statistics can give you the tools to tell the difference between the two. My goal in writing the book is achieved if reading it will mark for you the beginning of the process of learning statistics and not the end of the process.

16.4.2 Discussion in the Forum

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

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

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