Applied Biostatistics with R and rk.Teaching

Alfredo Sánchez Alberca (asalber@ceu.es)

Department of Applied Math and Statistics
CEU San Pablo

March 8, 2016

CEU
Universidad San Pablo

Applied Biostatistics with R

Alfredo Sánchez Alberca (asalber@ceu.es)

License terms

Contents

Practice 1

Frequency distributions and charts

1 Solved exercises

1. The number of children in a sample of 25 families is

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.$$

Do the following operations:

(a) Create a data set with the variable children and enter the data.

(b) Create the frequency table.

1. Select the menu Teaching > Frequency distribution > Frequency tabulation .
2. In the dialog displayed, select the variable children in the field Variable to tabulate and click the button Submit.

(c) Create the absolute frequency bar chart.

1. Select the menu Teaching > Charts > Bar chart.
2. In the dialog displayed, select the variable children in the field Variable and click the button Submit.

(d) Create also the relative frequency, cumulative absolute frequency and cumulative relative frequency bar charts, with their respective polygons.

Follow the steps above checking, in the Bar options tab, the box Relative frequencies for the relative frequencies bar chart, the box Cumulative frequencies for the cumulative absolute bar chart, and both of them for the cumulative relative frequency bar chart. Check the box Polygon, to plot the corresponding polygon.

2. The number of people treated in the emergency service of a hospital every day of November was

15	23	12	10	28	7	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

Do the following operations:

(a) Create a data set with the variable emergencies and enter the data.

(b) Create the box plot. Are there some outlier? In that case, remove the outliers and proceed with the next part.

1. Select the menu Teaching > charts > Box plot.
2. In the dialog displayed select the variable emergencies in the field Variables and click the button Submit.
3. In the output windows with the box plot identify the outliers.
4. In the data set edition tab, remove the rows with the outliers right-clicking the row header and selecting Delete this row.

(c) Create the frequency table grouping data into 5 classes.

1. Select the menu Teaching > Frequency distribution > Frequency tabulation.
2. In the dialog displayed select the variable emergencies.
3. In the Classes tab check the box Grouping intervals, check the option Number of intervals, enter the desired number of intervals in the field Suggested intervals and click the button Submit.

(d) Create the absolute frequency histogram.

1. Select the menu Teaching > charts > Histogram.
2. In the dialog displayed select the variable emergencies in the field Variable.
3. In the Classes tab, check the box Grouping intervals, check the box Number of intervals, enter the desired number of intervals in the field Suggested intervals and click the button Submit.

(e) Create also the relative frequency, cumulative absolute frequency and cumulative relative frequency histograms, with their respective polygons.

Follow the steps above checking, in the Histogram options, the box Relative frequencies for the relative frequency histogram, the box Cumulative frequencies for the cumulative absolute frequency histogram, and both of them for the cumulative relative frequency histogram. Check the box Polygon to plot the corresponding polygon.

3. The blood type of a sample of 30 persons are:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB,
A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

Do the following operations:

(a) Create a data set with the variable blood.type and enter the data.

(b) Create the frequency table.

1. Select the menu Teaching > Frequency distribution > Frequency tabulation.
2. In the dialog displayed, select the variable blood.type in the field Variable to tabulate and click the button Submit.

(c) Create the pie chart.

1. Select the menu Teaching > charts > Pie chart.
2. In the dialog displayed, select the variable blood.type in the field Variables and click the button Submit.

4. The age and the marital status of a sample of 28 persons are:

Marital status	Age								
Single	31	45	35	65	21	38	62	22	31
Married	72	39	62	59	25	44	54		
Widow(er)	80	68	65	40	78	69	75		
Divorced	31	65	59	58	50				

Do the following operations:

(a) Create a data set with the variables marital.status and age and enter the data.

(b) Create the frequency table of the variable age for every marital status.

1. Select the menu Teaching > Frequency distribution > Frequency tabulation.
2. In the dialog displayed, select the variable age in the field Variable to tabulate, check the box Tabulate by groups and select the variable marital.status in the field Grouping variable(s).
3. In the Classes tab, check the box Grouping intervals and click the button Submit.

(c) Create the box plots of age for every marital status. Are there outliers? Which group have more spread in ages?

1. Select the menu Teaching > charts > Box plot.
2. In the dialog displayed, select the variable age in the field Variable(s), check the box Plot by groups, select the variable marital.status in the field Grouping variable(s) and click the button Submit.

(d) Create the relative frequency histogram of age for every marital status. Compare the histograms.

1. Select the menu Teaching > charts > Histogram.
2. In the dialog displayed, select the variable age in the field Variable, check the box Plot by groups and select the variable marital.status in the field Grouping variable(s).
3. In the Classes tab, check the box Grouping intervals and click the button Submit.

2 Proposed exercises

1. The number of injuries suffered by the members of a soccer team in a league were

$$0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1$$

Do the following operations:

(a) Construct the frequency table.

(b) Create the relative frequency and the cumulative relative frequency bar charts.

(c) Create the box plot.

2. The heights (in cm) of 30 students are

$$179, 173, 181, 170, 158, 174, 172, 166, 194, 185,$$
$$162, 187, 198, 177, 178, 165, 154, 188, 166, 171,$$
$$175, 182, 167, 169, 172, 186, 172, 176, 168, 187.$$

Do the following operations:

(a) Create the absolute frequency histogram with classes of width 10 cm from 150 to 200 cm. Are there outliers?

3. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

(a) Construct the frequency table of the APGAR score at 1 minute. If an score less than or equal to 3 indicates that the neonate is depressed, what percentage of neonates is depressed in the sample?

(b) Construct the frequency table of the neonate weight, grouping into classes of width 0.5 Kg from 2 to 4.5 Kg. What intervals contains more neonates?

(c) Compare the frequency distribution of the APGAR score at 1 minute for mothers less than 20 years old and for mothers greater than 20 years old. What group has more depressed neonates?

(d) Compare the relative frequency distribution of the neonate weight according to whether the mother smoked or not during the pregnancy. If a weight under 2.5 Kg is considered a low weight, what group has a higher percentage of neonates with low weight?

(e) Compute the prevalence of neonates with low weight for smoking and non-smoking mothers during the pregnancy.

(f) Compute the relative risk of low weight of neonate when the mother smokes vs when then mother doesn't smoke during the pregnancy.

(g) Create the bar chart of the APAGAR score at 1 minute. What is the more common score?

(h) Construct the cumulative relative frequency bar chart of the APGAR score at 1 minute. Below what value is half of the neonates?

(i) Compare the relative frequency distribution bar charts of the APGAR score at 1 minute according to whether the mother smoked or not during the pregnancy. What conclusion can be drawn?

(j) Construct the histogram of the neonates weights with classes of width 0.5 Kg from 2 to 4.5 Kg. What class contains more neonates?

(k) Compare the relative frequency histograms of the neonates weights, with classes of width 0.5 Kg from 2 to 4.5, Kg, according to whether the mother smoked or not during the pregnancy. What group has neonates with less weight?

(l) Compare the relative frequency histograms of the neonates weights, with classes of width 0.5 Kg from 2 to 4.5, Kg, according to whether the mother smoked or not before the pregnancy. What conclusions can be drawn?

(m) Construct the box plot of the neonates weights. What range of weights can be considered normal in the sample? Are there outliers in the sample?

(n) Compare the box plots of the neonates weights according to whether the mother smoked or not during the pregnancy and whether the mother was less than 20 or greater than 20 years old. What group has more central spread? What group has neonates with less weight?

(o) Compare the box plots of the APGAR scores at 1 minute and at 5 minutes. What variable has more central spread?

Practice 2

Linear regression

1 Solved exercises

1. The values of two variables X and Y measured in a sample of 10 individuals are:

X	0	1	2	3	4	5	6	7	8	9
Y	2	5	8	11	14	17	20	23	26	29

Do the following operations:

(a) Create a data set with the variables X and Y and enter the data.

(b) Construct the scatter plot of X and Y.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, select the variable Y in the field Y variable, and the variable X in the field X variable, and click the button Submit.

According to the point cloud, what type of regression model explains better the relation between X and Y?

(c) Compute the regression line of Y on X.

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed, insert the variable Y in the field Dependent variable and the variable X in the field Independent variable, and click the button Submit.

(d) Plot the regression line on the scatter plot.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, insert the variable Y in the field Y variable and the variable X in the field X variable.
3. In the Fitted line tab, check the box Linear and click the button Submit.

(e) Compute la regression line of X on Y and plot it on the scatter plot.

Repeat the steps of the previous part but inserting the variable X in the field Dependent variable and the variable Y in the field Independent variable.

(f) How are the residuals? Comment the results.

2. A study pretends to determine the relation between the daily hours of study and the number of failed subjects in a course. The values of these variables in a sample of 30 students were:

Study hours	Failed subjects	Study hours	Failed subjects	Study hours	Failed subjects
3.5	1	2.2	2	1.3	4
0.6	5	3.3	0	3.1	0
2.8	1	1.7	3	2.3	2
2.5	3	1.1	3	3.2	2
2.6	1	2.0	3	0.9	4
3.9	0	3.5	0	1.7	2
1.5	3	2.1	2	0.2	5
0.7	3	1.8	2	2.9	1
3.6	1	1.1	4	1.0	3
3.7	1	0.7	4	2.3	2

Do the following operations:

(a) Create a data set with the variables Study.hours and Failed.subjects and enter the data of the sample.

(b) Construct the two-dimensional frequency table of the variables Study.hours and Failed.subjects.

1. Select the menu Teaching » Frequency distribution » Two-dimensional frequency table.
2. In the dialog displayed insert the variable Study.hours in the field Variable tabulate in rows, the variable Failed.subjects in the field Variable to tabulate in columns, and click the button Submit.

(c) Compute la regression line of Failed.subjects on Study.hours and plot it.

To compute the regression line:

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable Failed.subjects in the field Dependent variable and the variable Study.hours in the field Independent variable, check the box Save the model, enter the name linear.model.failed.subjects.on.study.hours for the regression model and click the button Submit.

To plot he regression line:

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable Failed.subjects in the field Y variable and the variable Study.hours in the field X variable.
3. In the Fitted line tab, check the box Linear and click the button Submit.

(d) What is the regression coefficient of the failed subjects on the daily hours of study? Interpret it.

The regression coefficient is the slope of the regression line.

(e) The linear relation is stronger or weaker than in the previous exercise? Answer the question comparing the residuals in both linear models.

(f) Compute the linear coefficient of determination and the correlation coefficient. Is the linear model a good model to explain the relation between the failed subjects and the daily hours of study? What percentage of the variability of the failed subjects is explained by the linear model?

The coefficient of determination is showed as R^2 in output window, and the correlation coefficient is the square root.

(g) Use the linear model to predict the expected number of failed subjects for a student that studies 3 hours a day. Is this prediction reliable?

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.model.failed.subjects.on.study.hours in the field Regression model, enter the value 3 in the in field Predictions for and click the button Submit.

(h) According to the linear model, how many hours of study are required at least to pass all the subjects?

To compute the regression line:

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable Study.hours in the field Dependent variable and the variable Failed.subjects in the field Independent variable, check the box Save the model, enter the name linear.model.study.hours.on.failed.subjects for the regression model and click the button Submit.

To make the prediction:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.model.study.hours.on.failed.subjects in the field Regression model, enter the value 0 in the in field Predictions for and click the button Submit.

3. To determine how an organism metabolizes the alcohol, an experiment was conducted where we measured the alcohol in blood every half an hour after drinking a liter of wine. The data of the experiment are below.

Time (min)	30	60	90	120	150	180	210
Alcohol (gr/l)	1.6	1.7	1.5	1.1	0.7	0.2	2.1

Do the following operations:

(a) Create a data set with the variables time and alcohol and enter the data of the sample.

(b) Compute the linear correlation coefficient of the alcohol and the time and interpret it. Is the linear model a good model to explain the metabolization of alcohol?

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable alcohol in the field Dependent variable, the variable time in the field Independent variable, and click the button Submit.

(c) Plot the regression line of alcohol on time. Are there some point with a big residual? In such a case, remove the point from the sample and compute again the linear correlation coefficient. Has the model improved?

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable alcohol in the field Y variable, the variable time in the field X variable.
3. In the Fitted line tab click the box Linear and click the button Submit.

It is observed that the point $(210, 2.1)$ has a huge residual compared to the others, what means that it's an outlier. To remove the outlier in the data edition windows, right-click the header row corresponding to the point and select Delete this row.

(d) If, according to the law, the maximum concentration of alcohol in blood to drive is 0.3 g/l, how much time must wait this person to drive after drinking a liter of wine? Is this prediction reliable?

To compute the regression line:

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable time in the field Dependent variable and the variable alcohol in the field Independent variable.
3. Check the box Save the model, enter the name linear.model.time.on.alcohol for the linear model and click the button Submit.

To make the prediction:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.model.time.on.alcohol in the field Regression model, enter the value 0.3 in the field Predictions for and click the button Submit.

4. The data set age.height of the package rk.Teaching contains the age and the height of 30 persons. Do the following operations:

(a) Load the data set age.height from the package rk.Teaching.

(b) Compute la regression line of the height on the age. Is the linear model a good model to explain the relation between the height and the age?

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable height in the field Dependent variable, the variable age in the field Independent variable, and click the button Submit.

(c) Create the scatter plot of the height on the age. Around which age changes the tendency?

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable height in the field Y variable, the variable age in the field X variable and click the button Submit.

(d) Recode the variable age into the categorical variable age.group with two categories for younger and older than 20 years.

1. Select the menu Teaching » Data » Variable recoding.
2. In the dialog displayed insert the variable age in the field Variable to recode.
3. In the field Recoding rules enter the following rules:

```
lo:20 = "younger"
20:hi = "older"
```

4. In the field Save new variable click the button Change.
5. In the dialog displayed select as parent object the data set age.height and click the button OK.
6. Enter the name age.group for the new variable and click the button Submit.

(e) Compute the regression line of the height on the age for every age group. In which group the linear model explains better the relation between the height and the age? Justify the answer.

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable height in the field Dependent variable and la variable age in the field Independent variable.
3. Check the box Regression by groups and insert the variable age.group in the field Grouping variable(s).
4. Check the box Save the model, enter the name linear.model.height.on.age for the linear model and click the button Submit.

(f) Plot the regression lines of the previous part.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable height in the field Y variable and the variable age in the field X variable.
3. Check the box Plot by groups and insert the variable age.group in the field Grouping variable(s).
4. In the Fitted line tab, check the box Linear and click the button Submit.

(g) According to the linear model, what is the expected height for a 14 years old person? And for a 38 years old person?

To predict the height of the 14 years old person:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.regression.height.on.age.younger in the field Regression model, enter the value 14 in the field Predictions for and click the button Submit.

To predict the height of the 38 years old person:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.regression.height.on.age.older in the field Regression model, enter the value 38 in the field Predictions for and click the button Submit.

2 Proposed exercises

1. A research study has been conducted to determine the loss of activity of a drug. The table below shows the results of the experiment.

Time (in years)	1	2	3	4	5
Activity (%)	96	84	70	58	52

Do the following operations:

(a) Compute the regression line of the drug activity on time.

(b) What percentage decreases the drug activity every year?

(c) How much time must pass for the drug to have an activity of 80Are these predictions reliable?

2.

3. In an study about the effect of different doses of a medicament, 2 patients got 2 mg and took 5 days to cure, 4 patients got 2 mg and took 6 days to cure, 2 patients got 3 mg ant took 3 days to cure, 4 patients got 3 mg and took 5 days to cure, 1 patient got 3 mg and took 6 days to cure, 5 patients got 4 mg and took 3 days to cure and 2 patients got 4 mg and took 5 days to cure. Do the following operations:
   (a) Compute the regression line of the days to cure on the dose.
   (b) Compute the regression coefficient of the days to cure on the dose and interpret it.
   (c) Compute the correlation coefficient and interpret it.
   (d) Determine the expected time required to cure with a 5 mg dose. Is this prediction reliable? Justify the answer.
   (e) What dose must be applied to last 4 days to cure? Is this prediction reliable? Justify the answer.
4. The data set heights.weights.students of the package rk.Teaching, contains the height, the weight and the gender of a sample of students. Do the following operations:
   (a) Load the data set heights.weights.students from the package rk.Teaching.
   (b) Compute the regression line of weight on height and plot it.
   (c) Compute the regression lines of weight on height for males and females and plot them.
   (d) Compute the coefficients of determination for both models. Which model explains better the relation between weight and height, the males or the females one? Justify the answer.
   (e) What is the expected weight for a man 170 cm tall? And for a women of the same height?
5. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:
   (a) Construct the two-dimensional frequency table of the APGAR score at 1 minute and whether the mother smoked or not during the pregnancy. What conclusions can you draw?
   (b) Construct the two-dimensional frequency table of te weight and the age of the mother. What conclusions can you draw?
   (c) Compute the regression line of the weight on the daily number of cigarettes smoked by the mother during the pregnancy. Is there a strong linear relation between the variables?
   (d) Plot the regression line of the previous part. Why the regression line doesn't fit well the point cloud?
   (e) Compute the regression line of the weight on the daily cigarettes smoked by the mother during the pregnancy in the group of smoking mothers. Is this regression model better or worse than the previous one? According to this model, how much decreases the weight of newborns for every daily cigarette smoked by the mother?
   (f) According to the previous linear model, what will be the expected weight of a nenonate with a mother that smokes 5 daily cigarettes during the pregnancy? And for a mother that smokes 30 daily cigarettes? Are these predictions reliable?
   (g) Are there the same linear relation between the weight and the daily cigarettes somoked by the mother for mothers younger than 20 years and mothers older than 20 years? What conclusions can you draw?