Rapport package team

Outlier tests

2011-04-26 20:25 CET

Contents

Description

This template will check if provided variable has any outliers.

Introduction

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. There are several ways to detect the outliers of our data. However, we cannot say one of them is the perfect method for that, thus it could be useful to take different methods into consideration. We present here four of them, one by a chart (a Box Plot based on IQR) and three by statistical descriptions (Lund Test, Grubb's test, Dixon's test).

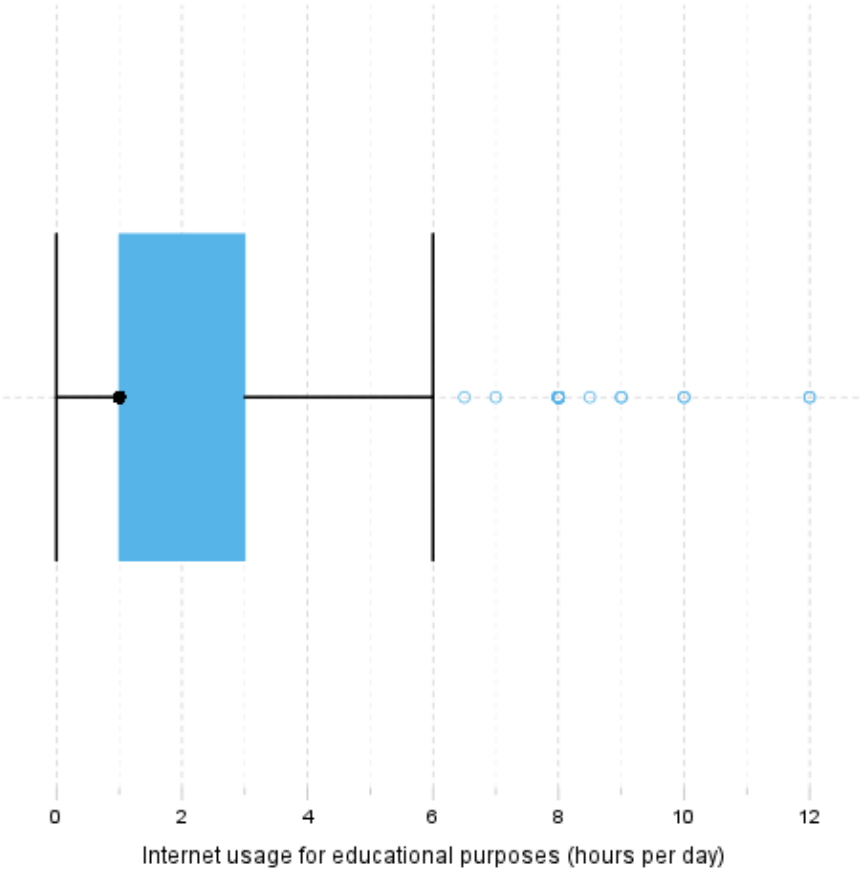
References

- Grubbs, F. E.: 1969, Procedures for detecting outlying observations in samples. Technometrics 11, pp. 1-21.

Charts

Among the graphical displays the Box plots are quite widespread, because of their several advantages. For example, one can easily get approximately punctual first impression from the data and one can visually see the positions of the (possible) outliers, with the help of them.

The Box Plot we used here is based on IQR (Interquartile Range), which is the difference between the higher and the lower quartiles. On the chart the blue box shows the "middle-half" of the data, the so-called whiskers shows the border where from the possible values can be called outliers. The lower whisker is placed 1.5 times below the first quartile, similarly the higher whisker 1.5 times above the third quartile.



References

- Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), Graphical Methods for Data Analysis, Wadsworth.
- Upton, Graham; Cook, Ian (1996). Understanding Statistics. Oxford University Press. p. 55.

Lund test

It seems that *4* extreme values can be found in "Internet usage for educational purposes (hours per day)". These are: *10*, *0.5*, *1.5* and *0.5*.

Explanation The above test for outliers was based on *lm(edu ~ 1)*:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.048	0.07797	26.27	7.939e-105

Table 1: Linear model: edu ~ 1

References

- Lund, R. E. 1975, "Tables for An Approximate Test for Outliers in Linear Models", Technometrics, vol. 17, no. 4, pp. 473-476.
- Prescott, P. 1975, "An Approximate Test for Outliers in Linear Models", Technometrics, vol. 17, no. 1, pp. 129-132.

Grubb's test

Grubbs test for one outlier shows that highest value 12 is an outlier (p=*0.0001964*).

References

- Grubbs, F.E. (1950). Sample Criteria for testing outlying observations. Ann. Math. Stat. 21, 1, 27-58.

Dixon's test

chi-squared test for outlier shows that highest value 12 is an outlier (p=*7.441e-07*).

References

- Dixon, W.J. (1950). Analysis of extreme values. Ann. Math. Stat. 21, 4, 488-506.

Description

This template will check if provided variable has any outliers.

Introduction

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. There are several ways to detect the outliers of our data. However, we cannot say one of them is the perfect method for that, thus it could be useful to take different methods into consideration. We present here four of them, one by a chart (a Box Plot based on IQR) and three by statistical descriptions (Lund Test, Grubb's test, Dixon's test).

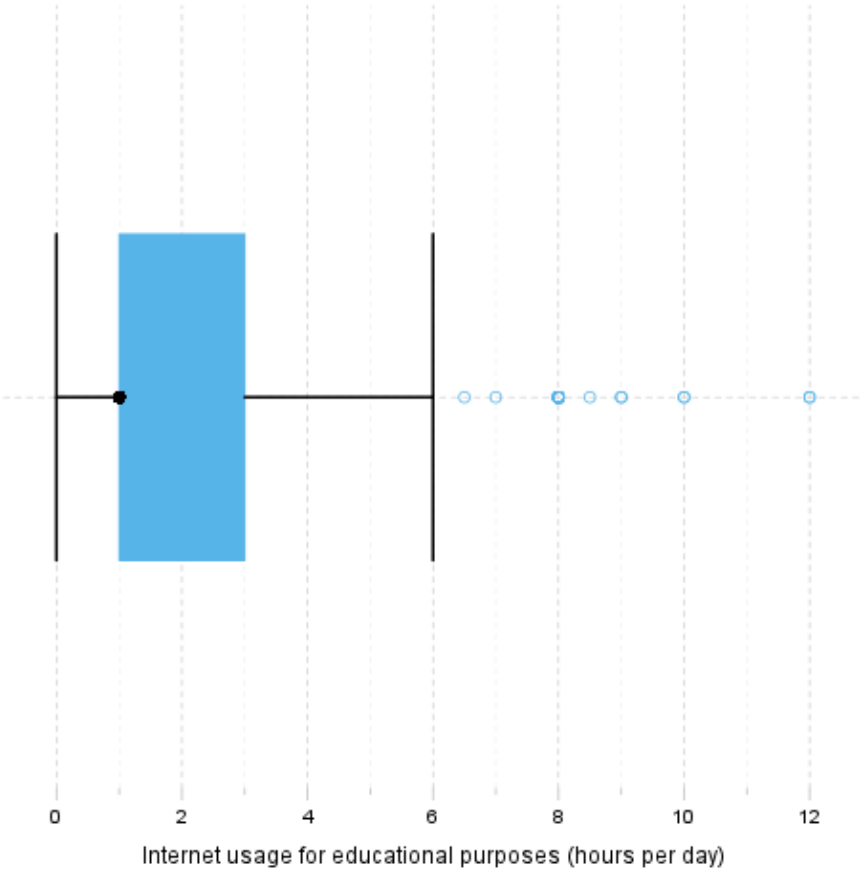
References

- Grubbs, F. E.: 1969, Procedures for detecting outlying observations in samples. Technometrics 11, pp. 1-21.

Charts

Among the graphical displays the Box plots are quite widespread, because of their several advantages. For example, one can easily get approximately punctual first impression from the data and one can visually see the positions of the (possible) outliers, with the help of them.

The Box Plot we used here is based on IQR (Interquartile Range), which is the difference between the higher and the lower quartiles. On the chart the blue box shows the "middle-half" of the data, the so-called whiskers shows the border where from the possible values can be called outliers. The lower whisker is placed 1.5 times below the first quartile, similarly the higher whisker 1.5 times above the third quartile.



References

- Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), Graphical Methods for Data Analysis, Wadsworth.
- Upton, Graham; Cook, Ian (1996). Understanding Statistics. Oxford University Press. p. 55.

Lund test

It seems that *4* extreme values can be found in "Internet usage for educational purposes (hours per day)". These are: *10*, *0.5*, *1.5* and *0.5*.

Explanation The above test for outliers was based on *lm(edu ~ 1)*:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.048	0.07797	26.27	7.939e-105

Table 2: Linear model: edu ~ 1

References

- Lund, R. E. 1975, "Tables for An Approximate Test for Outliers in Linear Models", Technometrics, vol. 17, no. 4, pp. 473-476.
- Prescott, P. 1975, "An Approximate Test for Outliers in Linear Models", Technometrics, vol. 17, no. 1, pp. 129-132.

Grubb's test

Grubbs test for one outlier shows that highest value 12 is an outlier (p=*0.0001964*).

References

- Grubbs, F.E. (1950). Sample Criteria for testing outlying observations. Ann. Math. Stat. 21, 1, 27-58.

Dixon's test

chi-squared test for outlier shows that highest value 12 is an outlier (p=*7.441e-07*).

References

- Dixon, W.J. (1950). Analysis of extreme values. Ann. Math. Stat. 21, 4, 488-506.

Description

This template will check if provided variable has any outliers.

Introduction

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. There are several ways to detect the outliers of our data. However, we cannot say one of them is the perfect method for that, thus it could be useful to take different methods into consideration. We present here four of them, one by a chart (a Box Plot based on IQR) and three by statistical descriptions (Lund Test, Grubb's test, Dixon's test).

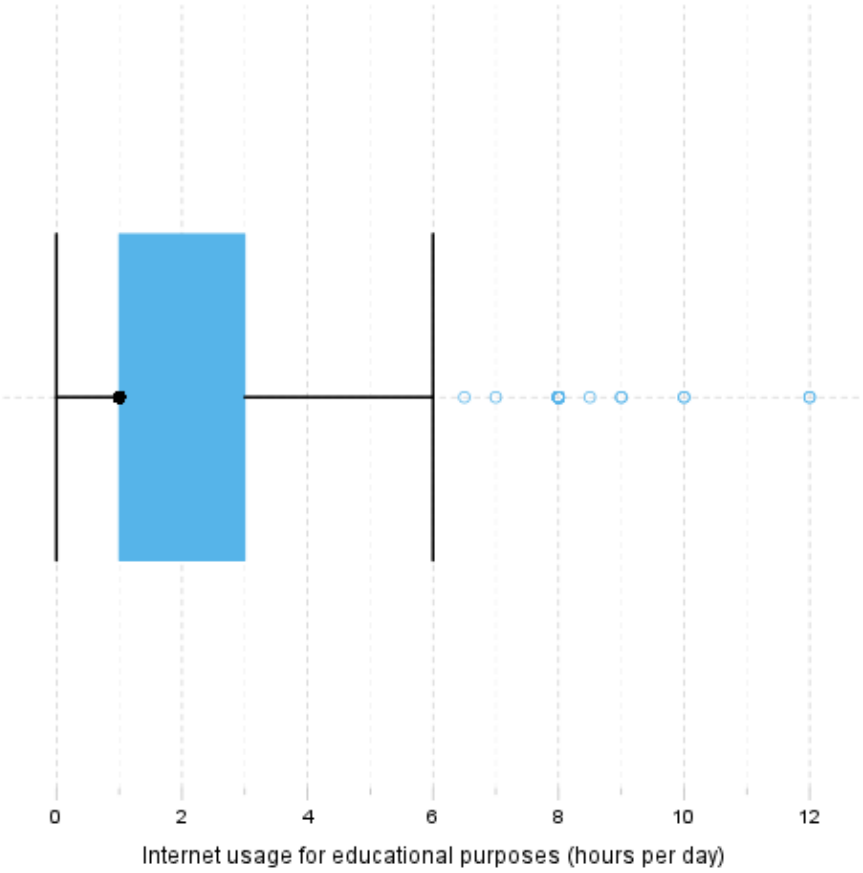
References

- Grubbs, F. E.: 1969, Procedures for detecting outlying observations in samples. Technometrics 11, pp. 1-21.

Charts

Among the graphical displays the Box plots are quite widespread, because of their several advantages. For example, one can easily get approximately punctual first impression from the data and one can visually see the positions of the (possible) outliers, with the help of them.

The Box Plot we used here is based on IQR (Interquartile Range), which is the difference between the higher and the lower quartiles. On the chart the blue box shows the "middle-half" of the data, the so-called whiskers shows the border where from the possible values can be called outliers. The lower whisker is placed 1.5 times below the first quartile, similarly the higher whisker 1.5 times above the third quartile.



References

- Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), Graphical Methods for Data Analysis, Wadsworth.
- Upton, Graham; Cook, Ian (1996). Understanding Statistics. Oxford University Press. p. 55.

Lund test

It seems that *4* extreme values can be found in "Internet usage for educational purposes (hours per day)". These are: *10*, *0.5*, *1.5* and *0.5*.

Explanation The above test for outliers was based on *lm(edu ~ 1)*:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.048	0.07797	26.27	7.939e-105

Table 3: Linear model: edu ~ 1

This report was generated with R (3.0.1) and rapport (0.51) in *1.082* sec on x86_64-unknown-linux-gnu platform.