M.Sc. in ‚Transportation Systems'

Applied Statistics in Transport

Exercises: Descriptive analysis and graphics with R

Description of the data set:

- The dataset consists of information on 2053 dwellings.

Variables:

- nr net-rent in EUR
- nrsqm net-rent per m2 in EUR
- ndwa net dwelling area in m2
- rooms number of rooms in the dwelling
- yc year of construction
- n neighbourhood
- agood good address? (y=1, n=0)
- abest best address? (y=1, n=0)
- hw hot water supply? (y=0, n=1)
- ch central heating available? (y=0, n=1)
- tb tiled bathroom? (y=0, n=1)
- bathextra special extra equipment in the bathroom? (y=1, n=0)
- kextra special extra equipment in the kitchen? (y=1, n=0)

Exercises – for solutions see also Ex_AppliedStatisticsDescriptiveR20101018.r:

You can solve the exercises with or without R, but please solve them without using the R-functions.

The following table shows the first 10 observations of the first 7 variables.

No. Obs.	nr	nrsqm	ndwa	rooms	yc	n	agood
1	741.39	10.9	68	2	1918	2	1
2	715.82	11.01	65	2	1995	2	1
3	528.25	8.38	63	3	1918	2	1
4	553.99	8.52	65	3	1983	16	0
5	698.21	6.98	100	4	1995	16	1
6	935.65	11.55	81	4	1980	16	0
7	204.85	3.72	55	2	1924	6	0
8	426.93	5.4	79	3	1924	6	0
9	446.33	8.58	52	1	1957	6	0
10	381.45	4.95	77	3	1948	6	0

Reading your data:

- setwd("##")
- getwd()
- rent.data<-read.table("rent.asc", header=T, sep="\t")
- ls()

- class(rent.data)
- str(rent.data)
- head(rent.data)
- summary(rent.data)
- dim(rent.data)
- names(rent.data)
- rent.data [1:8,]
- rent.data [1:5,4:7]
- rent.data [3,4]
- rent.data $rooms[3:8]
- fix(rent.data)
- attach(rent.data)

1. Which scale do the variables belong to? Choose suitable ways to visualise the variables nr, rooms, n. Classify the data for net rent if suitable.

Solution:

- net rent, nrsqm, ndwa continuous, quantitative, ratio; rooms discrete, ordinal or ratio, neighbourhood discrete, qualitative, nominal; yc, agood, abest discrete, qualitative, ordinal/nominal

some examples (see also your handout):

- table(rooms)
- table(n)/2053
- nr.group <- cut(nr, breaks = c(seq(0,1800,by=100)))
- table(nr.group)
- table(nr.group)/n
- par(mfrow = c(2,2)) #par can be used to set or query graphical parameters
- barplot(table(rooms), main = "rooms")
- barplot(table(n), main = "n")
- barplot(table(nm.group), main = "nr")
- hist(nr)
- hist(nr, breaks = 10, main = "breaks=10")

2. Which measures of central tendency are suitable for the variables? Compute the arithmetic mean, the median and the mode for the variables nr, rooms, n – so than each of the measures is used once.

Solution:

- cumsum(table(rooms)/2053)
- par(mfrow = c(2,1))
- plot(ecdf(rooms)) #Compute or plot an empirical cumulative distribution function
- plot(ecdf(nr))
- mean(nr)

- median(nr)

#computing the mode in R, alternative 1:

- table_room<-table(rent.data$room)
- subset(table_room, table_room==max(table_room))

#computing the mode in R, alternative 2:

- (my_mode = as.numeric(names(table(rent.data$room))[which.max(table(rent.data$room))]))

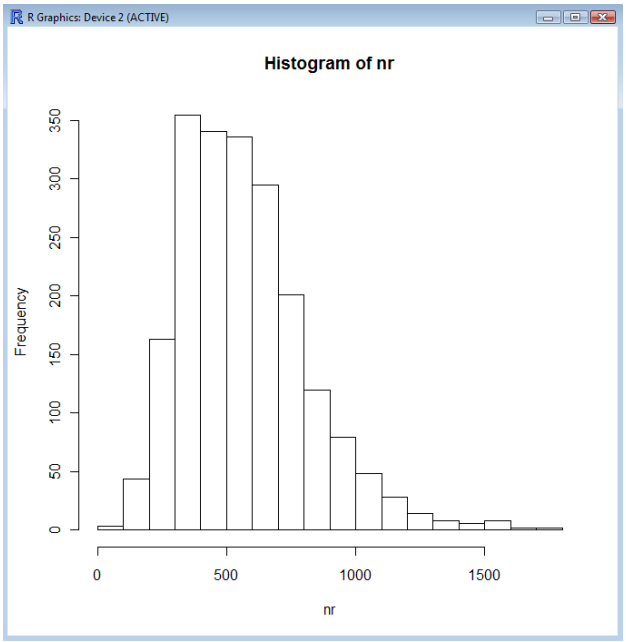
#computing the mode in R, alternative 3:

- statmod <- function(x) {
- z <- table(as.vector(x))
- names(z)[z == max(z)]
- }
- statmod(rent.data$room)

3. Compute the arithmetic mean and the median for the variable nr. Which advantage has the median compared to the arithmetic mean? What can you say about the degree of skew of the variable nr in the sub-sample? Compare this result with the graphs in exercise 1.

Solution:

- mean(nr)
- median(nr) #Median is more robust
- library(fBasics)
- skewness(nr) #[1] 1.1
- kurtosis(nr) #[1] 1.8
- e.g. hist(nr)



4. For which variables can you plot the empirical cumulative distribution function (ecdf)? Plot the ecdf for the variable rooms.

Solution:

- e.g. rooms, yc, n

- par(mfrow = c(1,1))
- plot(ecdf(rooms))

5. For which variables boxplots are suitable? Create and interpret a boxplot for the variable nr. Compute the quartiles for the variable nr.

Solution:

- quantile(rent.data$nr, probs = seq(0, 1, 0.2))
- quantile(rent.data$nr, probs = seq(0, 1, 0.25))
- boxplot(nr, main = "nr")
- IQR(rent.data$nr)

6. Which measures of dispersion are suitable for the variables? Compute the variance for the variable nr. Indicate the unit for the variance and the standard deviation of the variable nr.

Solution:

- Variance: Average of the squared distances of the elements of the sample from the mean. The unit of variance is the square of the unit of the original variable [e.g. Euro^2]. The positive square root of the variance is the standard deviation [e.g. Euro].
- sd(nr)
- apply(cbind(nr, nrsqm, ndwa), 2, sd)
- table(rooms, n)
- par(mfrow = c(2, 1))
- plot(nr, ndwa, xlab = "ndwa", ylab = "nr")
- plot(nrsqm, ndwa, xlab = "ndwa", ylab = " nrsqm ")
- cor(nrsqm, ndwa, use = "complete.obs")
- cor(nr, ndwa, use = "complete.obs")
- par(mfrow = c(1, 1))
- plot(nr ~ ndwa, data = rent.data, pch = rooms, col = rooms)
- legend(150, 600, legend = paste("rooms", c(1, 2, 4, 5,6)), col = c(1, 2, 4, 5,6), pch = c(1, 2, 4, 5,6))
- par(mfrow = c(1, 1))
- plot(nrsqm ~ ndwa, data = rent.data, pch = rooms, col = rooms)
- legend(150, 21, legend = paste("rooms", c(1, 2, 4, 5,6)), col = c(1, 2, 4, 5,6), pch = c(1, 2, 4, 5,6))
- par(mfrow = c(1, 1))

- boxplot(nrsqm ~ n, main = " nrsqm ")
- boxplot(nrsqm ~ n, main = " nrsqm ", xlab = "n", ylab = " nrsqm ")

7. Get an overview of the variables with the help of the function “summary”.

Solution:

- summary(rent.data)

8. Compute the absolute and relative frequencies of the variable rooms. Group the variable nr and compute the relative frequencies for the groups.

Solution:

- table(rooms)
- Number of groups: sqrt(n)=sqrt(2053)=45
- nr.group <- cut(nr, breaks = c(seq(0,1800,by=40)))
- table(nr.group)/2053 #sum(table(nr.group)/2053) gives 1
- length(rent) #gives 13, we can get the number of observations with n<-dim(rent), n[1]
- hist(n)
- hist(rooms)
- barplot(table(nr.group), main = "nr")

9. Create a contingency table for the variables agood and rooms.

Solution:

- table(agood, rooms)
- summary(nrsqm)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  1.5 6.8 8.5 8.4 10.1 20.1
- summary(ndwa)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  17 53 67 70 83 185
- nrsqm.group <- cut(nrsqm, breaks = c(seq(0,21,by=5)))
- table(nrsqm.group)
- ndwa.group <- cut(ndwa, breaks = c(seq(17,185,by=20)))
- table(ndwa.group)
- table(nrsqm.group, ndwa.group)

10. Visualize the variable nrsqm, grouped by the number of rooms (rooms). Interpret the graphic.

Solution:

nrsqm4.group <- cut(nrsqm, breaks = c(seq(0,21,by=5)))

```
nrsqm4.group
nrsqm.rooms <- table(nrsqm.group, rooms)
nrsqm.rooms
barplot(nrsqm.rooms, main = " nrsqm", xlab="rooms", ylab=" nrsqm")
barplot(nrsqm.rooms, main = " nrsqm", xlab="rooms", ylab=" nrsqm", beside = TRUE)
nrsqmRelative.rooms <- table(nmqm4.group, rooms)/2053
barplot(nrsqmRelative.rooms, main = "nrsqm", xlab="rooms", ylab="nrsqm")
```