Quantitative variables

Data tameRs

 Press `A` to change slides into text Press `T` to display table of contents.

What is this episode about?

When we analyze data we come across various types of variables. This episode will be devoted to the most popular groups of variables in the classical statistics ?quantitative variables.

In this episode you will learn:

- Which variables/ characteristics are referred to as quantitative characteristics?
- What basic operations can be performed on the quantitative characteristics?
- How to summarize/ describe the quantitative characteristics?

We will use two data sets to illustrate these issues. The first set, `koty_ptaki` is quite small, while the other one, `auta2012` is a much more extensive data set. Both data sets are available in the package `PogromcyDanych`.

What does it mean: a quantitative variable

In the next four episodes I will present five basic types of data: quantitative data, qualitative data, texts, logical values and dates. Although they do not exhaust the list of all possible types of data (there are also censored data and other more specialized types), we will focus on them because you will meet them in 90% of all the analyses.

The word ?type? means the same as the word ?kind? in this context. We can talk about types and kinds of data. However, data analysis jargon commonly uses the word ?type? and we will also use that word in our discussion.

Quantitative variables are characteristics which describe quantities ?height, length, weight, speed, surface area, age etc. They are also sometimes called the numerical variables as they are expressed in numbers. However, not every variable expressed in numbers is a quantitative characteristic. The National Personal Identification Number PESEL or a postal code are expressed in numbers but they do not represent measurable quantities such as amount or size.

Loading data

We will start our presentation with an example of `koty_ptaki` from the package `PogromcyDanych`. In order to load this data you only need to activate the package. You can check how to do that in the episode 2.

Load the package and use the function `head()` to display the first six rows. Which of the displayed variables are quantitative variables?

```
library(PogromcyDanych)
head(koty_ptaki)
```

```
##   gatunek waga dlugosc predkosc habitat zywotnosc druzyn
## 1  Tygrys  300     2.5        60    Azja        25     Ko
## 2     Lew  200     2.0        80  Afryka        29     Ko
## 3  Jaguar  100     1.7        90 Ameryka        15     Ko
## 4    Puma   80     1.7        70 Ameryka        13     Ko
## 5 Leopard   70     1.4        85    Azja        21     Ko
## 6  Gepard   60     1.4       115  Afryka        12     Ko
```

Descriptive statistics

The data set `koty_ptaki` consists of 13 rows, so that each quantitative variable includes 13 numbers.

It is good to characterize each variable with one or two indices, such as mean or median (median is a value which appears in the middle of the row of figures placed in the ascending order) in order to facilitate description of the variables.

How can we check value of the mean and the median for weight? We can use the function `mean()` to calculate the mean and the function `median()` to calculate the median.

```
mean(koty_ptaki$waga)
```

```
## [1] 78.59615
```

```
median(koty_ptaki$waga)
```

```
## [1] 60
```

Loading data

The data set `koty_ptaki` consists of 13 rows. All of them can be displayed on the computer screen. For such small sets we do not need special descriptive statistics to understand their content.

For this reason from now on we will practice working with quantitative values on a much bigger data set containing over 200 thousand values called `auta_2012`, which is also available in the package PogromcyDanych.

That data set is described in details in the episode `https://rawgit.com/pbiecek/MOOC/master/0_dane/0_dane.html`

Let us load that data set and look at its first three rows.

```
library(PogromcyDanych)
head(auta2012, 3)
```

```
##    Cena Waluta Cena.w.PLN Brutto.netto  KM  kW       Mark
## 1 49900    PLN      49900        brutto 140 103         Ki
## 2 88000    PLN      88000        brutto 156 115 Mitsubish
```

Descriptive statistics

Let us focus on characteristics such as `_Cena.w.PLN_` or `_Przebieg.w.km_`. These characteristics are expressed in numbers and they describe amounts.

First of the characteristics describes amount of money for which the seller wants to sell the car, while the second one describes the number of kilometers already covered by the car offered for sale.

We can perform a few operations on the quantitative variables. We can calculate the mean price of the cars offered for sale with the function `mean()` and we can calculate the median with the function `median()`, just like in case of the previous data set.

```
mean(auta2012$Cena.w.PLN)
```

```
## [1] 35755.11
```

```
median(auta2012$Cena.w.PLN)
```

```
## [1] 19900
```

Descriptive statistics

When we load data from a data set which was created by some other person we often want to find out what are the extreme values of specific variables. Extreme mean the smallest and the biggest.

We can check what is the smallest and the biggest price in the data set using the functions `min()` and `max()`.

```
min(auta2012$Cena.w.PLN)
```

```
## [1] 400
```

```
max(auta2012$Cena.w.PLN)
```

```
## [1] 11111111
```

The smallest price is PLN 400 what I find quite plausible, especially so because some cars offered for sale were damaged.

However, the biggest price is PLN 11 million 111 thousand 111. Maybe somebody suggested this price as a joke rather than a real

Descriptive statistics

The function `summary()` allows us to determine five basic characteristics: minimum, first quartile, median (second quartile), mean, third quartile and maximum.

Five of these characteristics, that is, minimum, 1st, 2nd, 3rd quartiles and maximum, are the so-called Tukey's five number summary which divide the values into four equinumerous ranges.

- Price of one fourth of all the cars is lower than the first quartile.
- Price of one fourth of all the cars is higher than the first quartile but lower than the median (second quartile).
- Price of one fourth of all the cars is higher than the median (second quartile) but lower than then third quartile.
- Price of one fourth of all the cars is higher than the third quartile.

Description of the quantitative variable provided by these six figures gives us a lot of information. Let us take a closer look at these characteristics on the example of prices of cars.

Descriptive statistics

Let us describe results of the function summary() using the following example:

```
summary(auta2012$Cena.w.PLN)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##       400     10900     19900     35760     37470 11110000
```

- The minimum price is PLN 400.
- One fourth of all the offers for sale offered cars for less than PLN 10 900 (as there are over 200 thousand cars we can quickly calculate that over 50 thousand offers is beyond that range).
- Half of all the offers for sale offered cars for less than PLN 19 900.
- Three fourth of all the offers for sale offered cars for less than 37 470 and, at the same time, one fourth of all the offers for sale offered cars for a higher price.

Descriptive statistics

Until now we have discussed the descriptive statistics for only one variable. However, sometimes we would be interested in the relationship between two variables.

The most typical descriptive statistics for a pair of quantitative variables is correlation. You can determine it using the function `cor()`. Let us check if there is a correlation between the year of manufacture and price.

```
cor(auta2012$Rok.produkcji, auta2012$Cena.w.PLN)
```

```
## [1] 0.3419877
```

There is a positive correlation what means that the later the year of manufacture of the car (and the younger the car), the more expensive it is. Correlation assumes values from -1 to 1. Value of an average correlation is 0,34. Most of us probably expected some higher value. In this case, however, we deal with a list containing very many different makes and models. The function `cor()`

Exercises

- There are 7 quantitative characteristics in the data set `auta2012`. Name them.
- One of the quantitative characteristics is `Rok.produkcji`. What is the median year of manufacture of the cars offered for sale? All the offers were submitted in 2012 ?what was the median age of the cars offered for sale?

You may find sample answers at `https://rawgit.com/pbiecek/MOOC/master/0_dane/9_zadania.html`

Missing values

It might happen that some offers for sale do not contain all the information about the car, for example, there is no information about mileage. Such gaps in data should be marked somehow. We cannot put e.g. 0 instead of information about mileage because we say two completely different things when we state that mileage is 0km and when we state that there is no information about mileage.

For this reason we need some special value which would stand for the missing data. In R that value is *NA* (short for *not available*).

How can we know that there is some data missing in the data set? We can use the function `summary()` which displays number of gaps. The variable `Przebieg.w.km` includes almost 40 thousand gaps. It means that many offers did not give information about mileage of cars at all.

Other characteristics, such as minimum, maximum and median, are calculated for the remaining 160 thousand cars for which the mileage was given.

```
summary(auta2012$Przebieg.w.km)
```

Missing values

What are the consequences of gaps in the data?

The most important consequence is the fact that some statistics simply cannot be calculated.

For example, the mean of 1 and 3 is 2. But what is the mean of 1 and *not avaliable*?

This is why the value of some statistics calculated for vectors including gaps is also `NA`.

```
mean(auta2012$Przebieg.w.km)
```

```
## [1] NA
```

```
min(auta2012$Przebieg.w.km)
```

```
## [1] NA
```

```
max(auta2012$Przebieg.w.km)
```

Exercises

- Which characteristic is most incomplete? Which characteristic contains the greatest number of gaps?
- What is the value of median size of the engine (`Pojemnosc.skokowa`)?

You may find sample answers at `https://rawgit.com/pbiecek/MOOC/master/0_dane/9_zadania.html`

Graphic descriptive statistic ?a bar chart

The descriptive statistics presented above can be displayed in the graphic form. A great advantage of the graphic presentations is that fact that a well-trained eye quickly reads a lot of information from them.

We can use many manners of presentation when we deal with a quantitative variable. We only need to take into consideration the amount of information that we want to present.

Let us assume that we want to present nothing more than the mean. One number. We will display it using a bar chart using the function `barplot()`.

Notice that if the result of the attribution is put in brackets, it will be displayed on the screen as well

```
(srednia <- mean(auta2012$Przebieg.w.km, na.rm=TRUE))
```

```
## [1] 147167.4
```

```
barplot(srednia)
```

Graphic descriptive statistic ?a bar chart

Let us check average prices calculated for specific groups. We can divide the cars in groups taking into consideration for example the type of fuel that they use.

We will use the function `tapply()` to calculate the mean for groups. Its first argument is the quantitative variable; its second argument is information about groups and its third argument is name of the function which is to be performed for each group (and then all the additional arguments).

The result will be a vector with means. Now we can use the function `barplot()`. The argument `las=2` will ensure that the labels on the OX axis will be displayed vertically ?they will be easier to read.

You will learn much more about the art of preparing good-looking charts in the season 2.

```
(srednie <- tapply(auta2012$Przebieg.w.km,
                   auta2012$Rodzaj.paliwa,
                   mean, na.rm=TRUE))
```

Graphic descriptive statistic ?a box plot

One bar looks strange and it conveys little information. It presents only one number and it is difficult to read it as there is nothing to compare it with.

A chart presenting more information would be much more interesting. For the descriptive statistics we used the function `summary()` which presents six characteristic pieces of information concerning distribution (minimum, maximum, quartiles, mean). Five of these numbers, except for mean, are included in the box plot which can be prepared in R with the function `boxplot()`.

The example below presents these five numbers calculated for the variable `Przebieg.w.km`. Thanks to the argument `horizontal=TRUE` the whole chart is drawn horizontally, while the argument `range=0` ensures that the outliers are not included on the plot (as we have not discussed them yet, let us not draw them ?we will return to this topic later on).

```
summary(auta2012$Przebieg.w.km)
```

Graphic descriptive statistic ?a box plot

What a strange diagram! Is there something wrong? Look at the axis ?its range is up to 10^9 = 1 000 000 000 km. Yet when we look at the result of the function `summary()` we notice that indeed one of the rows contains such enormous value. Such big values are most probably mistakes. It is most unlikely that mileage of any car in the world was even close to that value (if any car was driving non-stop at the speed of 100km/h, after 100 years its mileage would be only 8 000 000km).

Data is often polluted and the graphic presentation allows us to notice that fact very easily.

That is also true in this case. Let us clean up the data by removing all the rows for which the value of the characteristic `Przebieg.w.km` is over 1 000 000km. We will use a logical condition to index only the rows with plausible mileage. You can find more information about indexing in the episode 7.

```
auta2012wybrane <- auta2012[auta2012$Przebieg.w.km < 10000
summary(auta2012wybrane$Przebieg.w.km)
```

Graphic descriptive statistic ?a histogram

Each element of this diagram (left whisker, left part of the box, right part of the box, right whisker) presents 1/4 of all data. This plot shows us clearly that 3/4 of all data concerns cars with mileage less than 200 thousand km. Only 1/4 of data contains higher values and for this group of cars the range of mileage is very wide.

If we want to present more detailed data on mileage we can use a histogram. This type of diagram presents number of observations in certain intervals. The intervals represent equal ranges of values. There is a special algorithm calculating the number of intervals depending on the variability of the characteristic (there are usually 6 to 10 intervals).

You can use the function `hist()` to create a histogram in R. Its first arguments specifiy the data which are to be presented on the diagram. The bars of the sample diagram below were painted grey to make them more visible.

```
hist(auta2012wybrane$Przebieg.w.km, col="grey")
```

Graphic descriptive statistic ?a histogram

The argument `breaks` allows us to specify the number of intervals and even the boundaries of the intervals. If we have many observations, we usually create more detailed and informative diagrams by drawing many intervals. Each interval of the histogram presents the number of observations. The higher the bar, the more cars with a certain mileage there are in the data set.

You can notice two ?hills? on the histogram below. One of them is near 0 and the other around 180 thousand km. It means that the data set includes many cars with very small mileage and only few cars with the mileage close to 50 thousand km. There are also many cars with the mileage close to 150 thousand km and just several cars with the mileage of 300 thousand km.

As you could see, quantitative values can be described with various level of detail. The more detailed is our description, the more elements there are on the chart. The more elements, the more information can be read from the chart, but also the more difficult it is to read it at all.

Graphic descriptive statistic ?a scatter plot

Can we present two quantitative variables and the relationship between them at the same time?

Yes, we can and a scatter plot is the best solution in that case. If the first two arguments of the function `plot()` are quantitative characteristics, we will receive a scatter plot. Additional arguments such as `xlab` and `ylab` specify the names of the axes OX and OY.

The example below presents the relationship between weight and speed of the species described in the data set `koty_ptaki`. As you can see, high weight does not correspond with high speed.

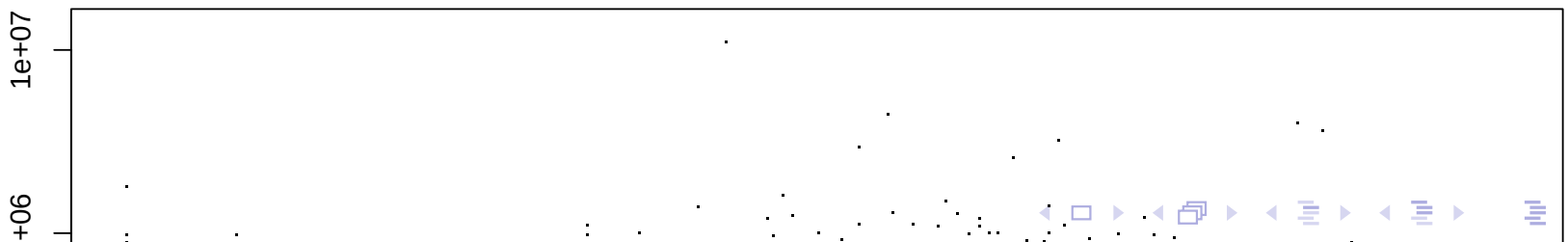
```
plot(koty_ptaki$waga, koty_ptaki$predkosc, ylab="Speed", xl
```

Graphic descriptive statistic ?a scatter plot

Let us use the function `plot()` to show the relationship between mileage and price on the basis of the data on cars. There are so many offers that it turns out that it is better to mark specific offers as dots, not big circles. We can do that by adding the argument `pch=` which allows us to modify the signs used to mark points representing data on the plot. Individual values of both mileage and price are very high. Instead of using a standard scale it is better to use a logarithmic scale. You can do that when you add the argument `log=?xy?`. The value ?xy? means that both axes must be logarithmed.

More information on scatter plots and other ways of displaying quantitative characteristics will be presented in the season 3.

```
plot(auta2012wybrane$Przebieg.w.km, auta2012wybrane$Cena.w
```



Exercises

- Present graphically the distribution of the characteristic `Cena.w.PLN` as a box-and-whisker plot and a histogram.
- Notice that several cars offered for very high prices make the whole diagram a little illegible. Clean up the data and leave only the cars which cost less than PLN 100 thousand. Then present the distribution of prices of cars for the group of cars which cost up to PLN 100 thousand.
- What is the most frequent price of cars in the set of gathered offers?

You can find sample answers at `https://rawgit.com/pbiecek/MOOC/master/0_dane/9_zadania.html`