Descriptive statistics

R is mainly used for data processing, analysis and visualisation. Subsequent parts of the present work are devoted to these three typical applications.

Before we discuss those complex applications, we will present some basic cases here. Variables in data analyses are usually characterised according to the classification by Stanley Stevens:

- Qualitative variables (also referred to as factors or categorical variables) are variables which can take on a limited number of values (usually non-numerical). They can be further divided into the following groups:
  - Binary variables (also known as dichotomous or binomial variables), such as gender (female/male).
  - Nominal variables (also known as unordered qualitative variables), such as car make: there is no specific order for car makes.
  - Ordinal variables (also known as ordered qualitative variables), such as education (primary/secondary/tertiary).
- Quantitative variables, which can be further divided into:
  - Count variables (count of occurrences of a given phenomenon expressed as a natural number), such as the number of education years.
  - Interval variables, measured on a scale where values can be subtracted, but not divided by each other, such as temperature in Celsius degrees, or A.D. year.
  - Ratio variables, measured on a scale where proportions are kept. This means that values can be divided by one another and there is a clear definition of 0.0. Examples include temperature in Kelvin degrees or height in centimetres.

In R, quantitative variables are represented with a numerical type called `numeric`. There are no separate types to describe numbers on a ratio scale or an interval scale.

Qualitative data in R are represented with a type called `factor. factor` variables can be additionally marked as ordered. In such cases, they have an additional class called `ordered`.

Binary variables can be represented with a logical type called `logical`.

Table @ref(tab:tab01) presents some functions which calculate the most popular descriptive statistics. We will practice calculating descriptive statistics on a data set called `socData` from the `Przewodnik` package.

```
library("Przewodnik")
socData <- read.csv("/home/krz/socData.csv"); head(socData, 3)
```

```
##   age  education      civil_status    sex          employment
## 1  70 vocational in a relationship   male student or employed
## 2  66 vocational in a relationship female student or employed
## 3  71 vocational            single female student or employed
##   systolic_pressure diastolic_pressure
## 1               143                 83
## 2               123                 80
## 3               167                 80
```

Table 1: Descriptive statistics for a vector or matrix

Function	Description
.	`base` package
`max()/min()`	Maximal/minimal value in the sample.
`mean()`	Arithmetic mean,$\bar{x} = \sum_i x_i/n$ `trim` is an optional argument. When it is different than 0, a trimmed mean is calculated. A trimmed mean is calculated just like the arithmetic mean after removing 200% $*$ `trim` of edge observations.
`length()`	Count of elements in the sample.
`range()`	Variability range of the sample, calculated as $[\min_i x_i, \max_i x_i]$.
.	`stats` package
`weighted.mean`	Weighted mean, calculated as $\frac{1}{n}\sum_i w_i x_i$. The weight vector w_i is the second argument.
`median()`	Median (middle value).
`quantile()`	Q-quantile. The second argument of `quantile()` is the vector of quantiles to find. This function implements 9 different algorithms to find quantiles, see the description of `type` argument for more information.
`IQR()`	Interquartile range, i.e. the difference between the upper and lower quartile, $IQR = q_{0.75} - q_{0.25}$.
`var()`	Variation in the sample. The unbiased estimator of variance is calculated as $S^2 = \frac{1}{n-1}\sum_i (x - \bar{x})^2$. For two vectors, the covariance of these two vectors will be calculated. For a matrix, the covariance matrix for its columns will be calculated instead.
`sd()`	Standard deviation, calculated as $\sqrt{S^2}$, where S^2 is the estimator of variance.
`cor(), cov()`	Correlation and covariance matrix. The arguments may be a pair of vectors, or a matrix.
`mad()`	Median absolute deviation, calculated as $1.4826 * median(\|x_i - median(x_i)\|)$.
.	other packages
`kurtosis()`	Kurtosis, measure of concentration, $\frac{n\sum_i (x_i-\bar{x})^4}{(\sum_i (x_i-\bar{x})^2)^2} - 3$. The normal distribution has a kurtosis of 0. This function comes from the `e1071` package.
`skewness()`	Skewness, measure of asymmetry, $\frac{\sqrt{n}\sum_i (x_i-\bar{x})^3}{(\sum_i (x_i-\bar{x})^2)^{3/2}}$. The symmetric distribution has a skewness of 0. This function comes from the `e1071` package.
`geometric.mean()`	Geometric mean, calculated as $(\prod_i x_i)^{1/n}$. This function comes from the `psych` package.
`harmonic.mean()`	Harmonic mean, calculated as $n/\sum_i x_i^{-1}$. This function comes from the `psych` package.
`moda()`	Mode, i.e. the most frequent value. This function comes from the `dprep` package. In Linux, we can also use the `mod()` function from `RVAideMemoire`.

Quantitative variables

Let us take a look at the values in the `age` column. We can refer to that column with `socData$age`.

Age is a quantitative ratio variable (ratios make sense in this case; for example, we can say that someone is twice as old as someone else).

Our first question is: what are the lowest and greatest values that the `age` variable can take on? It is always a good idea to check boundary values as they may help us identify errors in data.

```
range(socData$age)
```

```
## [1] 22 75
```

What is the mean age?

```
mean(socData$age)
```

```
## [1] 43.16176
```

And what is the trimmed mean calculated for the middle 60% of observations?

```
mean(socData$age, trim=0.2)
```

```
## [1] 42.58065
```

The median turns out to be close to the mean – that could mean there is no skewness.

```
median(socData$age)
```

```
## [1] 45
```

We can use the `summary()` function to quickly calculate the most important characteristics. In the case of quantitative variables, the result is given as a vector with the following values: the minimum, maximum, mean, median, first and third quartiles (also called lower and upper quartiles).

All of these values, apart from the mean, are always returned by the `fivenum()` function (the so-called five-number summary that divides the values observed into four equal parts). If there are missing observations in the variable, their count is also given.

```
summary(socData$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   30.00   45.00   43.16   53.00   75.00
```

Standard deviation:

```
sd(socData$age)
```

```
## [1] 13.8471
```

Kurtosis / measure of tailedness:

```
e1071::kurtosis(socData$age)
```

```
## [1] -0.9558479
```

Skewness:

```
e1071::skewness(socData$age)
```

```
## [1] 0.233151
```

Selected quantiles of the `age` variable:

```
quantile(socData$age, c(0.1, 0.25, 0.5, 0.75, 0.9))
```

```
##  10%  25%  50%  75%  90%
## 26.0 30.0 45.0 53.0 62.4
```

One statistic which is frequently computed for multiple variables is called correlation. We can use the `cor()` function to calculate it. A correlation matrix is given below for three selected columns:

```
cor(socData[,c(1,6,7)])
```

```
##                            age systolic_pressure diastolic_pressure
## age                 1.00000000       -0.02765239        -0.08313656
## systolic_pressure  -0.02765239        1.00000000         0.67852707
## diastolic_pressure -0.08313656        0.67852707         1.00000000
```

Qualitative variables

Let us now take a look at the `education` column. We can refer to it by typing `socData$education`.

Education is a qualitative variable. It can take on four different values and there is a natural order for them.

A contingency table is the most frequent statistic for qualitative variables. The example below uses the `table()` function:

```
table(socData$education)
```

```
##
##    primary  secondary   tertiary vocational
##         93         55         34         22
```

This function defines a contingency table for one, two or more count variables. Contingency tables can also be obtained with `xtabs()` and `ftable()`.

```
table(socData$education, socData$employment)
```

```
##
##               student or employed unemployed
##   primary                      71         22
##   secondary                    39         16
##   tertiary                     28          6
##   vocational                   14          8
```

In the case of qualitative variables, the `summary()` function has a similar effect to the `table()` function. The only difference is that `table()` ignores `NA` data, whereas `summary()` provides their count.

```
summary(socData$education)
```

```
##    primary  secondary   tertiary vocational
##         93         55         34         22
```

The `summary()` function can also take an argument of `data.frame` type. In this case, summaries are given for each column of the data frame.

```
summary(socData[,1:4])
```

```
##       age             education              civil_status     sex
##  Min.   :22.00   primary   :93   in a relationship: 84   female: 55
##  1st Qu.:30.00   secondary :55   single           :120   male  :149
##  Median :45.00   tertiary  :34
##  Mean   :43.16   vocational:22
##  3rd Qu.:53.00
##  Max.   :75.00
```