# Descriptive statistics and graphics

August 22, 2017

Before going into the actual statistical modelling and analysis of a data set, it is often useful to make some simple characterizations of the data in terms of summary statistics and graphics.

```
> x <- rnorm(50)
 x
 [1] -0.309270366 -1.193744540 -1.371289095  0.099782810  0.201376186 -2.326314652 -0.60577991
 [10] -0.068659988 -0.263489063  1.607827954 -0.884346937 -1.450494465 -0.161458140  1.43151878
 [19] -1.184462248  0.311736055  0.043895519 -0.873953248 -0.081652654 -2.173983874 -1.57404385
 [28] -0.551099497  0.003053438  0.035236063  0.277447881  0.191386276 -0.337250847  0.37413719
 [37] -0.767269692 -0.317717418  1.979043918  0.917119797 -1.508397175 -0.301541621 -0.23035740
 [46]  0.822813163 -0.635045606 -1.172923973 -1.415615259  2.141197337



> mean(x)
[1] 0.03301363
> sd(x)
[1] 1.069454
> var(x)
[1] 1.143731
> median(x)
[1] -0.08682795
> quantile(x)
0% 25% 50% 75% 100%
-2.60741896 -0.54495849 -0.08682795 0.70018536 2.98872414
> pvec <- seq(0,1,0.1)
> pvec
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> quantile(x,pvec)
0% 10% 20% 30% 40%
-2.60741896 -1.07746896 -0.70409272 -0.46507213 -0.29976610
50% 60% 70% 80% 90%
-0.08682795 0.19436950 0.49060129 0.90165137 1.31873981
100%
2.98872414
```

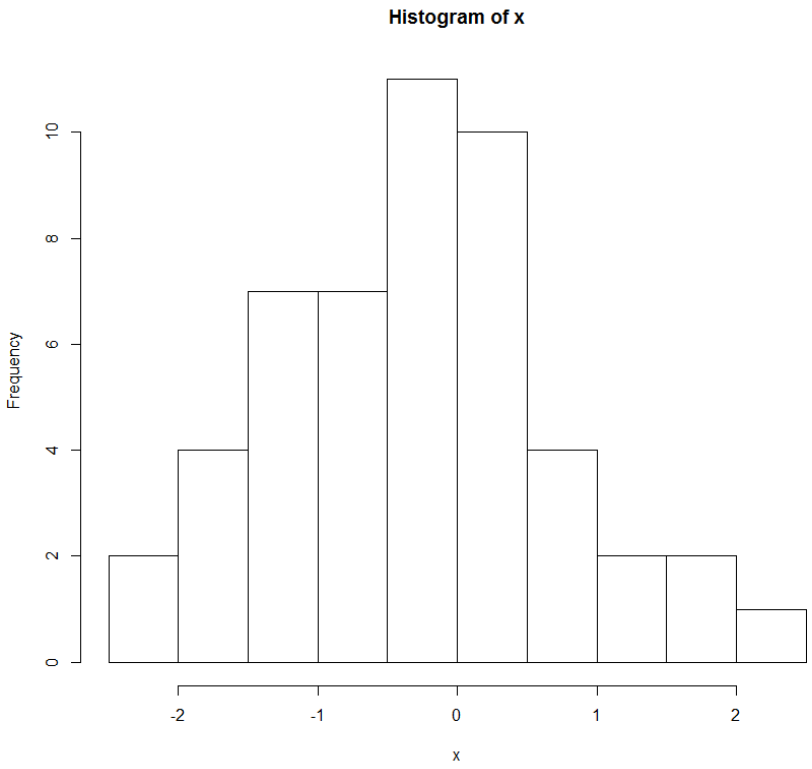


Figure 1: Histogram of 50 normal observations

## 0.1 Missing value

If there are missing values in data, things become a bit more complicated. For illustration, we use the following example.
Example: The data set juul contains variables from an investigation performed by Anders Juul (Rigshospitalet, Department for Growth and Reproduction) concerning serum IGF-I (insulin-like growth factor) in a group of healthy humans, primarily schoolchildren. The data set is contained in the **ISwR package** and contains a number of variables, of which we only use igf1 (serum IGF-I) for now, but later in the chapter we also use tanner (Tanner stage of puberty, a classification into five groups based on appearance
of primary and secondary sexual characteristics), sex, and menarche (indicating whether or not a girl has had her first period). Attempting to calculate the mean of igf1 reveals a problem.

```
> attach(juul)
> mean(igf1)
[1] NA
```

**R will not skip missing values unless explicitly requested to do so.**
The mean of a vector with an unknown value is unknown. However, you can give the na.rm argument (not available, remove) to request that missing values be removed:

```
> mean(igf1,na.rm=T)
[1] 340.168
```

There is one slightly annoying exception: The length function will not understand na.rm, so we cannot use it to count the number of nonmissing measurements of igf1. However, you can use

```
> sum(!is.na(igf1))
[1] 1018
```

A nice summary display of a numeric variable is obtained from the summary function:

```
> summary(igf1)
Min. 1st Qu. Median Mean 3rd Qu. Max. NAs
25.0 202.2 313.5 340.2 462.8 915.0 321.0
```

The 1st Qu. and 3rd Qu. refer to the empirical quartiles (0.25 and 0.75 quantiles). In fact, it is possible to summarize an entire data frame with

```
> summary(juul)
age menarche sex
Min. : 0.170 Min. : 1.000 Min. :1.000
1st Qu.: 9.053 1st Qu.: 1.000 1st Qu.:1.000
Median :12.560 Median : 1.000 Median :2.000
Mean :15.095 Mean : 1.476 Mean :1.534
3rd Qu.:16.855 3rd Qu.: 2.000 3rd Qu.:2.000
Max. :83.000 Max. : 2.000 Max. :2.000
NAs : 5.000 NAs :635.000 NAs :5.000
igf1 tanner testvol
Min. : 25.0 Min. : 1.000 Min. : 1.000
1st Qu.:202.2 1st Qu.: 1.000 1st Qu.: 1.000
```

```
70 4. Descriptive statistics and graphics
Median :313.5 Median : 2.000 Median : 3.000
Mean :340.2 Mean : 2.640 Mean : 7.896
3rd Qu.:462.8 3rd Qu.: 5.000 3rd Qu.: 15.000
Max. :915.0 Max. : 5.000 Max. : 30.000
NAs :321.0 NAs :240.000 NAs :859.000
```

The data set has menarche, sex, and tanner coded as numeric variables even though they are clearly categorical. This can be mended as follows:

```
> detach(juul)
> juul$sex <- factor(juul$sex,labels=c("M","F"))
> juul$menarche <- factor(juul$menarche,labels=c("No","Yes"))
> juul$tanner <- factor(juul$tanner,
+ labels=c("I","II","III","IV","V"))
> attach(juul)
> summary(juul)
age menarche sex igf1
Min. : 0.170 No :369 M :621 Min. : 25.0
1st Qu.: 9.053 Yes :335 F :713 1st Qu.:202.2
Median :12.560 NAs:635 NAs: 5 Median :313.5
Mean :15.095 Mean :340.2
3rd Qu.:16.855 3rd Qu.:462.8
Max. :83.000 Max. :915.0
NAs : 5.000 NAs :321.0
tanner testvol
I :515 Min. : 1.000
II :103 1st Qu.: 1.000
III : 72 Median : 3.000
IV : 81 Mean : 7.896
V :328 3rd Qu.: 15.000
NAs:240 Max. : 30.000
NAs :859.000
```

Notice how the display changes for the factor variables. Note also that juul was detached and reattached after the modification. This is because modifying a data frame does not affect any attached version. It was not strictly necessary to do it here because summary works directly on the data frame whether attached or not. In the above, the variables sex, menarche, and tanner were converted to factors with suitable level names (in the raw data these are represented using numeric codes). The converted variables were put back into the data frame juul, replacing the original sex, tanner, and menarche variables.We might also have used the transform function (or within):

```
> juul <- transform(juul,
+ sex=factor(sex,labels=c("M","F")),
+ menarche=factor(menarche,labels=c("No","Yes")),
+ tanner=factor(tanner,labels=c("I","II","III","IV","V")))
```

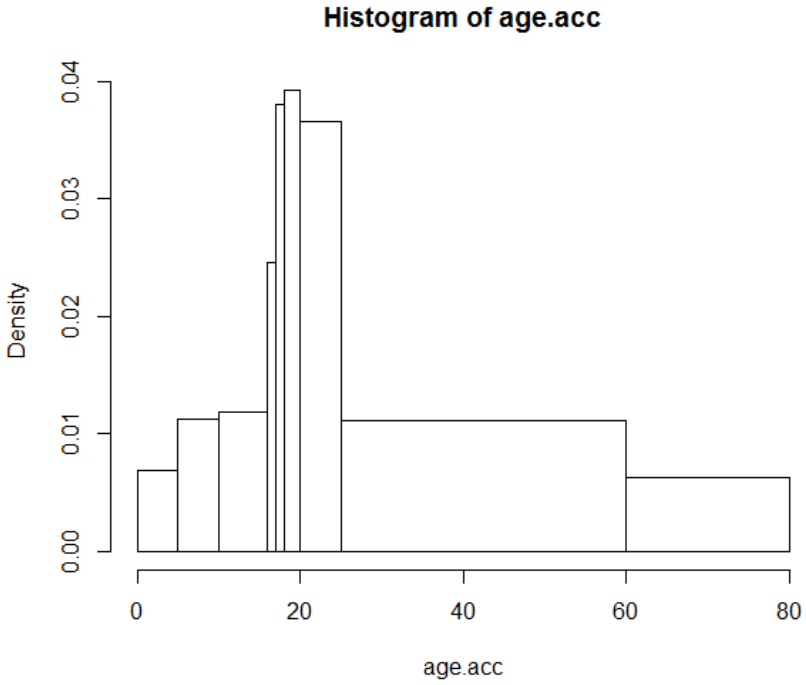


Figure 2: Histogram with unequal divisions

## 0.2 Graphical display of distributions

### 0.2.1 Histograms

You can get a reasonable impression of the shape of a distribution by drawing a histogram.

```
> hist(x)
```

Altman (1991, pp. 2526) contains an example of accident rates by age group. These are given as a count in age groups04,59,1015,16,17,1819,2024,2559,and6079years of age. The data can be entered as follows:

```
mid.age <- c(2.5,7.5,13,16.5,17.5,19,22.5,44.5,70.5)
 acc.count <- c(28,46,58,20,31,64,149,316,103)
 age.acc <- rep(mid.age,acc.count)
 brk <- c(0,5,10,16,17,18,20,25,60,80)
 hist(age.acc,breaks=brk)
```

### 0.2.2 Empirical cumulative distribution

```
n <- length(x)
 plot(sort(x),(1:n)/n,type="s",ylim=c(0,1))
  qqnorm(x)
  par(mfrow=c(1,2))
```

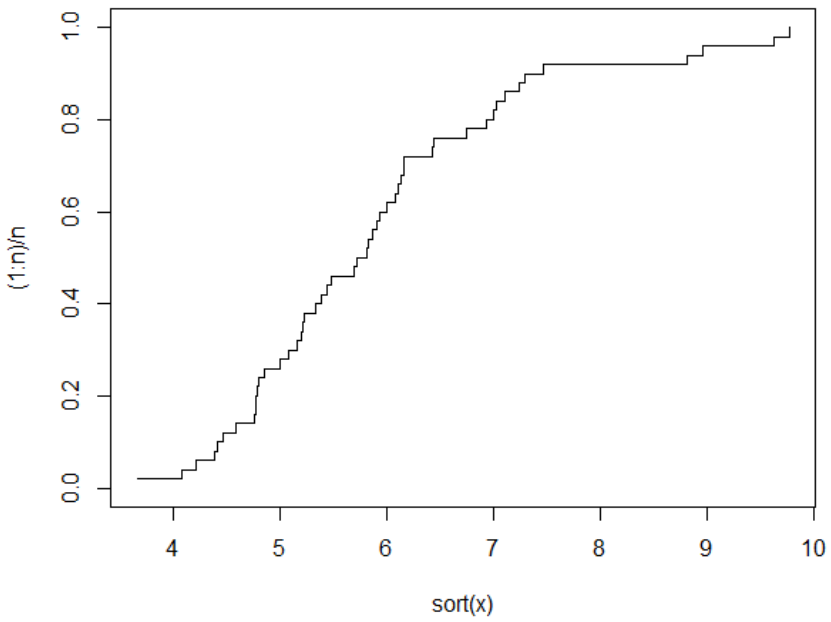


Figure 3: Empirical cumulative distribution function.

```
boxplot(IgM)
boxplot(log(IgM))
par(mfrow=c(1,1))
```

Here is how a boxplot is drawn in R. The box in the middle indicates hinges (nearly quartiles; see the help page for boxplot.stats) and median. The lines (whiskers) show the largest or smallest observation that falls with in a distance o f1.5 times the box size from the nearesthinge. If any observations fall farther away, the additional points are considered extreme values and are shown separately.

```
tapply(igf1, tanner, mean) I II III IV V NA NA NA NA NA
```

We need to get tapply to pass na.rm=T as a parameter to mean to make it exclude the missing values. This is achieved simply by passing it as an additional argument to tapply.

```
tapply(igf1, tanner, mean, na.rm=T)
 I            II       III      IV       V
 207.4727 352.6714 483.2222 513.0172 465.3344
```

Thenaturalwayofstoringgroupeddatainadataframeistohavethedata themselves in one vector and parallel to that have a factor telling which data are from which group. Consider, for instance, the following data set on energy expenditure for lean and obese women.

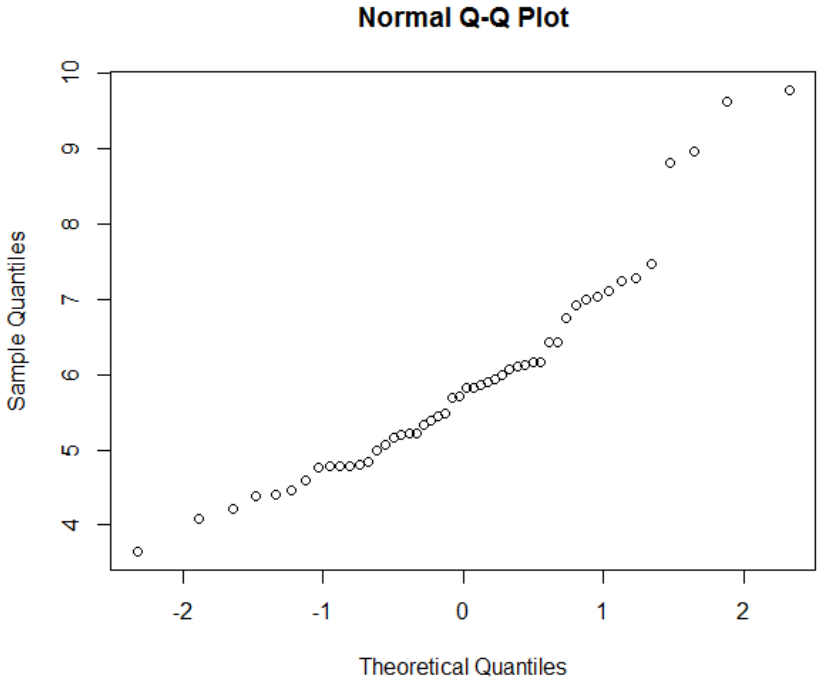


Figure 4: QQ plot using qqnorm(x).

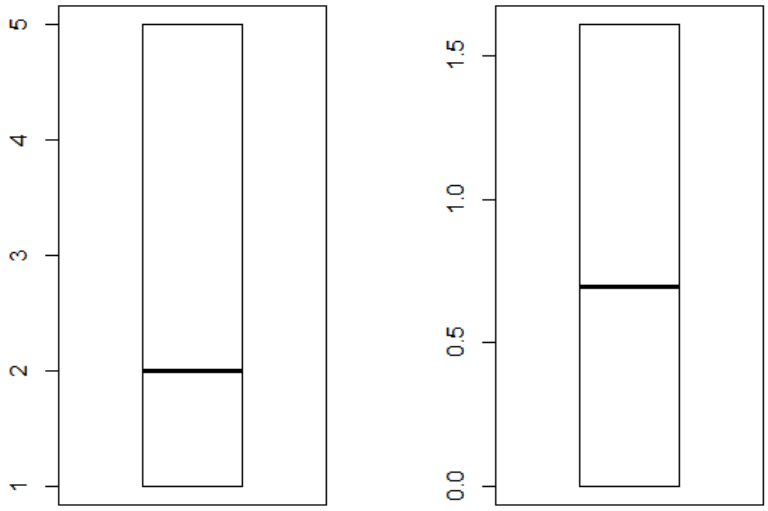


Figure 5: Boxplots for tunner in juul data set

```
mid.age <- c(2.5,7.5,13,16.5,17.5,19,22.5,44.5,70.5)
acc.count <- c(28,46,58,20,31,64,149,316,103)
age.acc <- rep(mid.age,acc.count)
brk <- c(0,5,10,16,17,18,20,25,60,80)
hist(age.acc,breaks=brk)
n <- length(x)
plot(sort(x),(1:n)/n,type="s",ylim=c(0,1))
plot(sort(x),(1:n)/n,ylim=c(0,1))
qqnorm(x)
data("juul")
par(mfrow=c(1,2))
boxplot( juul$tanner )
boxplot(log( juul$tanner ))
par(mfrow=c(1,2))
```