Lesson 2:
Descriptive Statistics

Recap: Data Management

1. Import data: from excel table
2. Check, if read correctly
3. Transform variables
4. Select subsets

Important commands (recap)

COMMAND	EFFECT
head(data$variable)	prints first few rows of dataset, including variable names
ls(data)	creates a list of all variable names (alphabetized)
names(data)	prints all variables names (in order of appearance)
summary(data$variable)	summarizes descriptive statistics of variable
nrow(data)	number of rows
ncol(data)	number of columns
subset(data, data$variable == "condition")	creates a subset from dataset
quantile(data$numeric_variable)	(package stats) calculates quartiles of numeric variable

Descriptive Statistics

1. Categorical variables:

- Frequency tables: one / two variables
- Graphs: bar plots

2. Continuous variables:

- Common measures like mean, standard deviation, median, etc.
- Graphs: histograms

1. Categorical variables

Frequency table (one variable):

- Command `summary(data$cat_variable)` produces frequency table for categorical variables
- Alternative: `table(data$cat_variable)`

- Q: How would you produce a table with relative frequencies?
- `summary(data$cat_variable) / nrow(data)`
- Extra: round to two decimals with `round(object, 2)`

1. Categorical variables

Contingency table (two variables):

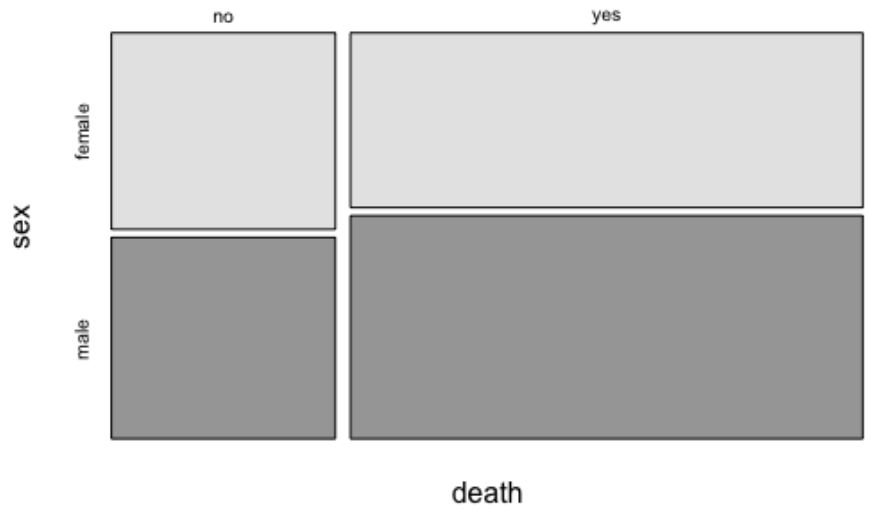
- Simply include both variables in table() command:

  `table(data$var_1, data$var_2)`

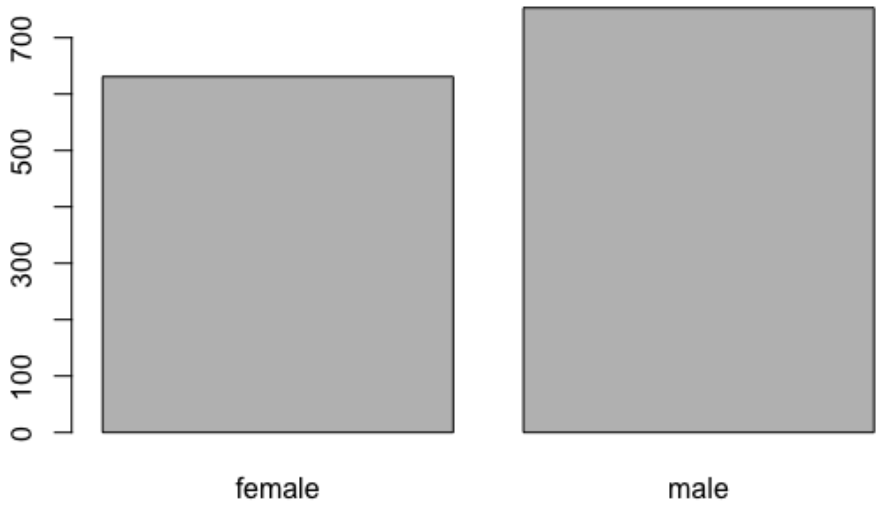
- More detailed alternative:

  `crosstab(data$var_1, data$var_2)` (package `descr`)

```
                data$death
data$sex          no    yes    Total
-----------------------------------
female           208    423      631
-----------------------------------
male             213    540      753
-----------------------------------
Total            421    963     1384
```



1. Categorical variables



Bar plot

```
barplot(data$variable)
```

```
barplot(data$variable/nrow(data),
     main = c("sex"),
     ylim = range(0,0.8))
```

- Q: what is the difference?

1. Categorical variables

- Q: what if you want three graphs in one plot?
- Use command `par(mfrow = c(1, 3))`
  - `mfrow`: vector with two variables `c(x, y)` where `x` defines number of rows and `y` number of columns in which graphs will be spread out
  - Be careful: using par() will spread all of the following graphs out in the same way. Use `c(1, 1)` to reset to only one graph per plot

2. Continuous variables

Description

- `summary(data$cont_variable)` gives overview, including:

```
   Min.  1st Qu.  Median   Mean   3rd Qu.  Max.
   4.00   6.00    8.00     15.55  8.50     100.00
```

- For more information: `describe()` from package `psych`

```
  vars  n  mean   sd     median trimmed  mad   min max range skew kurtosis   se
  1    11  15.55  28.07   8       7.44   2.97   4  100  96   2.45   4.46    8.46
```

2. Continuous variables

Description

Specific measures:

Median: `median()`

Mean: `mean()`

Maximum: `max()`

Variance: `var()`

Minimum: `min()`

Standard deviation: `sd()`

Range (Minimum and Maximum): `range()`

- in general: add `na.rm = TRUE` so NA's are excluded!

2. Continuous variables

Description for subgroups

- describe variable separately for a categorical variable
- use `describeBy(variable, group = cat_var)` from package `psych`

Example: `describeBy(colon$nodes, group = colon$sex)`

```
Descriptive statistics by group
group: 0
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 435 3.73 3.55      3    3.02 2.97   1  24    23  2.1     5.49 0.17
------------------------------------------------------------
group: 1
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 476 3.59 3.59      2     2.9 1.48   0  33    33 3.06       15 0.16
```

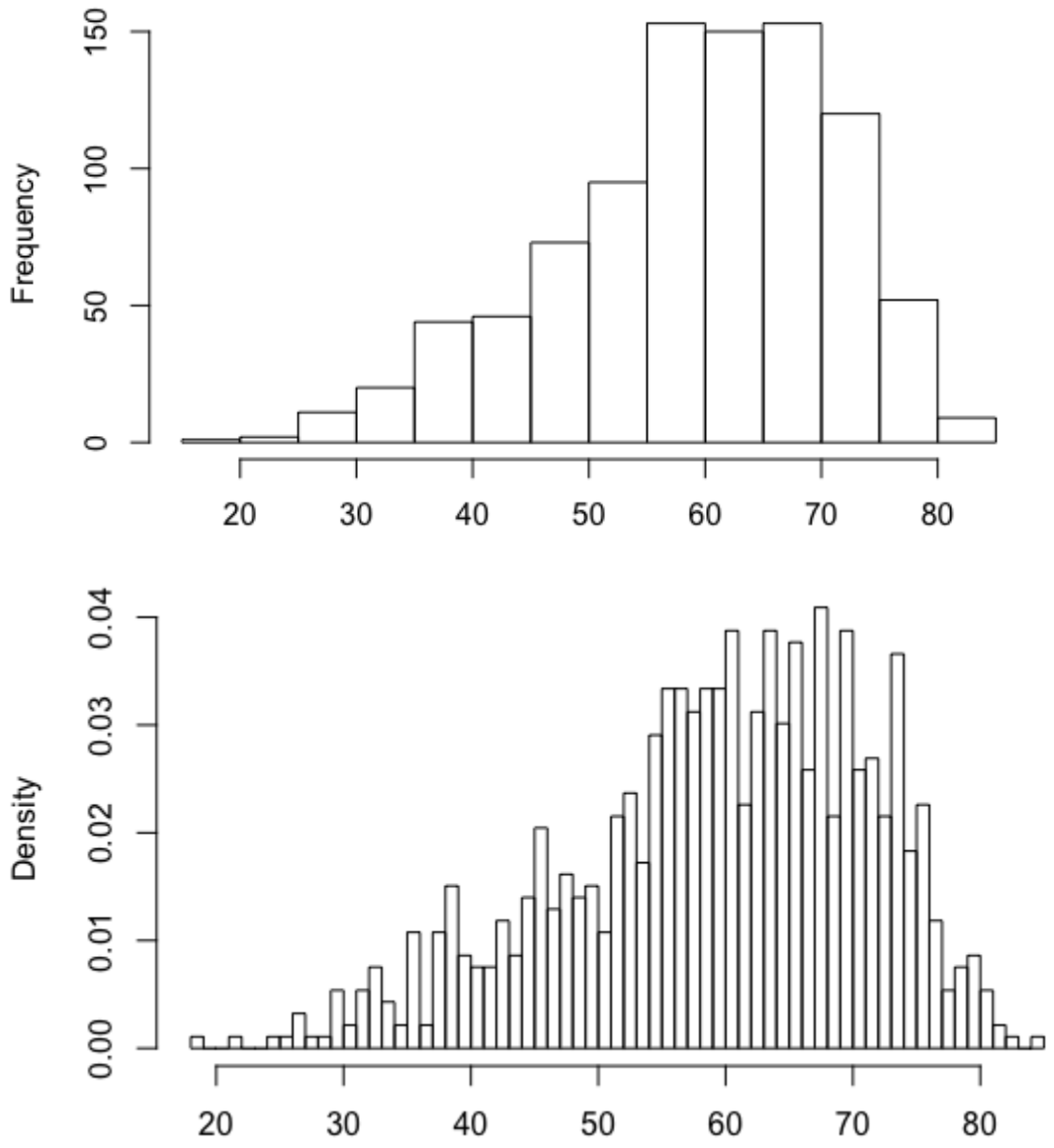
2. Continuous variables

Histograms

```
hist(data$time)
```

```
hist(data$time,
     breaks = 50,
     freq = FALSE)
```

- Q: what has changed?



2. Continuous variables

Histograms

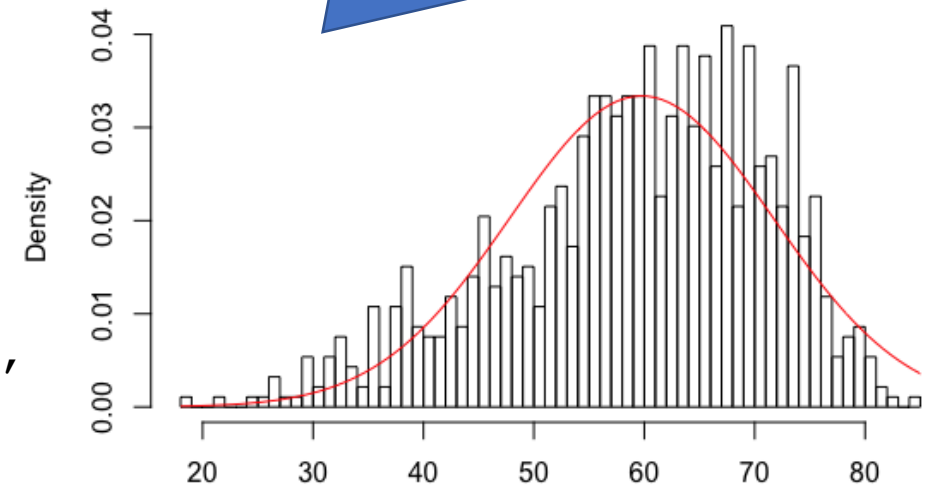
- is the variable normally distributed?

  → add normal curve with mean and sd of variable

```
hist(colon$age,
     breaks = 50,
     freq = FALSE)

curve(dnorm(x, mean = mean(colon$age, na.rm = TRUE),
               sd = sd(colon$age, na.rm = TRUE)),
      add = TRUE,
      col = "red")
```

only works if frequency is set to FALSE!



Preview: Graphs

What we already know:

- barplot() and hist()
- multiple graphs in one plot with par()
- how can graphs be specified?
- different kind of graphs?

Example: add main title to multiple graphs

```
par(mfrow = c(1, 2),
       oma = c(0, 0, 2, 0)
# add graphs
mtext("multiple graphs in one plot",
       cex = 1.5,
       outer = TRUE)
```

add 2 spaces at bottom

magnify font by 50%

Important commands

COMMAND	EFFECT
`round(object, x)`	round numbers in object to x decimal places
`par(mfrow = c(x, y), oma = (0,0,0,0))`	environment where graphs will be put in one plot of x rows, y columns
`barplot(var, main = c("title"), ylim = range())`	create barplot of variable, add main title and change range of y-axis
`table(var_1, (var_2))`	create frequency table for one or optionally two variables
`crosstab(var_1, var_2)`	package `descr`, more detailed cross table, including a graph
`mean(var, na.omit = TRUE), median(), sd(), quantile(), range()`	calculate mean/median/sd/quantile/range of variable. Add `na.rm = TRUE` to exclude missings from calculation
`describe(var)`	package `psych`, more extensive than summary() with skew, kurtosis etc.
`describeBy(var, group = cat_var)`	package `psych`, descriptive statistics for variable, separated for group variable
`hist(var, breaks = x, freq = TRUE)`	create histogram of variable, change count of bars, change `freq` to `FALSE` to get density function instead
`curve(dnorm(x, mean = , sd = , ), add = TRUE, col = "red")`	produces a curve that follows a normal distribution with mean and sd, added to last graphic, colour is read