# Beschreibende Statistik

André Meichtry[1]

2019

[1] Statistical Consulting

# Statistics

- Descriptive statistics
  - describe data
- Inferential statistics
  - what is the best guess of the truth, given some data? $\rightarrow$ point estimation
  - what is a range of plausible truths, given the data? $\rightarrow$ estimation, confidence intervals
  - is a specific truth plausible? $\rightarrow$ hypothesis testing

# Terminology

- statistical unit: one member of a set of entities being studied
- variable: quantified aspect of the statistical unit
- population: arbitrary defined set of units (with in- and exclusion criteria)
- sample: subset of the population, in the ideal case, randomly chosen from the population

# Read in data `read.table()`

Example data *stroke.csv*

- ```r
  rm(list=ls())
  mydata  <-  read.csv("https://raw.githubusercontent.com/mcdr65/StatsRsource/master/Data/stroke.csv")
  ```

  ```r
  head(mydata)
  ```

  ```
  ##     SEX AGE DGN COMA DIAB MINF HAN
  ## 1   man  76 INF   no   no  yes  no
  ## 2   man  58 INF   no   no   no  no
  ## 3   man  74 INF   no   no  yes yes
  ## 4 women  77 ICH   no  yes   no yes
  ## 5 women  76 INF   no  yes   no yes
  ## 6   man  48 ICH  yes   no   no yes
  ```

- for `Excel`-data: Look at helpfiles `?read.table` and `?read.csv` for further information
- see preparation exercise!!
  ```

# Look at data

Let us look at a simple dataset (`chickwts`) available in the R environment:

```
head(chickwts) ##only the first observations
```

```
##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

# help()

Very important: help files for R-functions.
For example, the help for the function `median`

```
help(median)
## or ?median
```

# Look at object, `str()`

look at the structure of R-objects

```
str(chickwts)
```

```
## 'data.frame': 71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
```

# Measurement scales

the scale of measurement depends on the preserved property during the mapping of the empirical world into the numerical world:

| scale | preserves | example | operations |
|---|---|---|---|
| nominal | categories | gender | $=, \neq$ |
| ordinal | order | independence score | $\leq, \geq$ |
| interval | equidistance | celcius | $+, -$ |
| ratio | ratios | kelvin | $\cdot, /$ |

Table: measurement scales

# R, creation of new variables

```
##weight in kilograms
chickwts$weightkg <- chickwts$weight/1000
str(chickwts)

## 'data.frame': 71 obs. of  3 variables:
##  $ weight  : num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed    : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ weightkg: num  0.179 0.16 0.136 0.227 0.217 0.168 0.108 0.124 0.143 0.14 ...
```

# Look at first observations, `head()`

```
head(chickwts)
```

```
##   weight      feed weightkg
## 1    179 horsebean    0.179
## 2    160 horsebean    0.160
## 3    136 horsebean    0.136
## 4    227 horsebean    0.227
## 5    217 horsebean    0.217
## 6    168 horsebean    0.168
```

# Sample

- $n$ observations of a random variable $X$

$$X_1 = x_1, \quad X_2 = x_2, \quad \ldots, \quad X_n = x_n.$$

- $n$: sample size
- $X$: random quantity which is observed $n$-times
- $X_1, X_2, \ldots, X_n$ constitutes a sample
- $X$ is random, because each unit is viewed as randomly chosen unit from the population
- How many observations has `chickwts`, how many variables has `chickwts`

# Frequencies

- absolute frequency $f$ [$f$: "‘frequency"’]: The number of observations of a specific value on the variable $X$
- relative frequency $f_{rel}$: The proportion of observations that have a specific value on the variable $X$
- absolute cumulative frequency: $F$: The number of observations that are smaller or equal ($\leq$) than a specific value on the variable $X$
- relative cumulative frequency $F_{rel}$: The proportion of observations that are smaller or equal ($\leq$) than a specific value on the variable $X$

# Frequency distribution, histogram, `hist()`

```
hist(chickwts$weightkg,xlab="Gewicht(kg)",freq=FALSE)
```

# Empirical cumulative distribution, `ecdf()`

```
cumfreq<-ecdf(chickwts$weightkg)
plot(cumfreq,xlab="weight",main="cumul.freq.")
```

# Did we understand `ecdf()`?

```
cumfreq(max(chickwts$weightkg))
```

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

```
cumfreq(median(chickwts$weightkg))
```

```
## [1] 0.507
```

# summary()

```
summary(chickwts)
```

```
##      weight            feed       weightkg
##  Min.   :108   casein   :12   Min.   :0.108
##  1st Qu.:204   horsebean:10   1st Qu.:0.204
##  Median :258   linseed  :12   Median :0.258
##  Mean   :261   meatmeal :11   Mean   :0.261
##  3rd Qu.:324   soybean  :14   3rd Qu.:0.324
##  Max.   :423   sunflower:12   Max.   :0.423
```

# boxplot()

```
boxplot(weight~feed,data=chickwts,
        ylab="weight",xlab="feed",las=2)
```

# summary measures of data subsets, aggregate()

```
aggregate(chickwts$weightkg,by=list(chickwts$feed),FUN="summary")
```

```
##     Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
## 1    casein  0.216     0.277    0.342  0.324     0.371  0.404
## 2 horsebean  0.108     0.137    0.151  0.160     0.176  0.227
## 3   linseed  0.141     0.178    0.221  0.219     0.258  0.309
## 4  meatmeal  0.153     0.249    0.263  0.277     0.320  0.380
## 5   soybean  0.158     0.207    0.248  0.246     0.270  0.329
## 6 sunflower  0.226     0.313    0.328  0.329     0.340  0.423
```

# aggregate()

```
aggregate(chickwts$weight,by=list(chickwts$feed),FUN="quantile")
```

```
##     Group.1 x.0% x.25% x.50% x.75% x.100%
## 1    casein  216   277   342   371    404
## 2 horsebean  108   137   152   176    227
## 3   linseed  141   178   221   258    309
## 4  meatmeal  153   250   263   320    380
## 5   soybean  158   207   248   270    329
## 6 sunflower  226   313   328   340    423
```

# Summary measures of data subsets, `split()`

```
subgr <- split(chickwts[,c(1,3)],f=chickwts$feed)
lapply(subgr,summary)
```

```
## $casein
##      weight       weightkg
##  Min.   :216   Min.   :0.216
##  1st Qu.:277   1st Qu.:0.277
##  Median :342   Median :0.342
##  Mean   :324   Mean   :0.324
##  3rd Qu.:371   3rd Qu.:0.371
##  Max.   :404   Max.   :0.404
##
## $horsebean
##      weight       weightkg
##  Min.   :108   Min.   :0.108
##  1st Qu.:137   1st Qu.:0.137
##  Median :152   Median :0.151
##  Mean   :160   Mean   :0.160
##  3rd Qu.:176   3rd Qu.:0.176
##  Max.   :227   Max.   :0.227
##
## $linseed
##      weight       weightkg
##  Min.   :141   Min.   :0.141
##  1st Qu.:178   1st Qu.:0.178
##  Median :221   Median :0.221
##  Mean   :219   Mean   :0.219
##  3rd Qu.:258   3rd Qu.:0.258
##  Max.   :309   Max.   :0.309
##
## $meatmeal
##      weight       weightkg
##  Min.   :153   Min.   :0.153
##  1st Qu.:250   1st Qu.:0.250
##  Median :263   Median :0.263
##  Mean   :277   Mean   :0.277
##  3rd Qu.:320   3rd Qu.:0.320
##  Max.   :380   Max.   :0.380
##
## $soybean
```

# Kreuztabellen, table()

```r
head(mydata)
```

```
##     SEX AGE DGN COMA DIAB MINF HAN
## 1   man  76 INF   no   no  yes  no
## 2   man  58 INF   no   no   no  no
## 3   man  74 INF   no   no  yes yes
## 4 women  77 ICH   no  yes   no yes
## 5 women  76 INF   no  yes   no yes
## 6   man  48 ICH  yes   no   no yes
```

```r
table(mydata[,c(1,5)])
```

```
##        DIAB
## SEX      no yes
##   man   291  28
##   women 441  69
```

# Boxplot

- median: 0.5-quantile
- the box goes from the 0.25-quantile to the 0.75-quantile. This distance is the interquartile range (IQR)
- the lines go to the value of the data point which lies still below of 0.75-quantile $+$ 1.5 IQR resp. above 0.25-quantiles - 1.5 IQR

# Quantile: quantile()

```r
quantile(chickwts$weightkg)
```

```
##    0%   25%   50%   75%  100%
## 0.108 0.205 0.258 0.324 0.423
```

```r
quantile(chickwts$weightkg,prob=c(.33,.66))
```

```
##   33%   66%
## 0.226 0.310
```

# summary()

```
summary(chickwts)
```

```
##      weight             feed       weightkg
##  Min.   :108   casein   :12   Min.   :0.108
##  1st Qu.:204   horsebean:10   1st Qu.:0.204
##  Median :258   linseed  :12   Median :0.258
##  Mean   :261   meatmeal :11   Mean   :0.261
##  3rd Qu.:324   soybean  :14   3rd Qu.:0.324
##  Max.   :423   sunflower:12   Max.   :0.423
```

```
table(chickwts$feed)
```

```
##
##    casein horsebean   linseed  meatmeal   soybean sunflower
##        12        10        12        11        14        12
```

# Measures of central tendency/dispersion

| measure / scale | nominal | ordinal | metric |
|---|---|---|---|
| central tendency | mode | mode, median | mode, median, mean |
| dispersion | | range, IQR | $s$,$s^2$ |

# `mean`, `mean()`

- empirical mean oder arithmetic mean:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- $\bar{x}$ is the most frequent measure of central tendency
- however, $\bar{x}$ is not a robust measure of central tendency
- when we do not have metric data, the arithmetic mean gives no sense

# median, `median()`

- median: makes sense if variable $X$ is at least ordinal scaled
- to calculate the median, we order the values of $X$
- the median of an ordered sample $(x_{(1)}, ..., x_{(n)})$ of $n$ observations is:

$$\text{median} = \begin{cases} 0.5(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ even} \\ x_{((n+1)/2)} & \text{if } n \text{ uneven.} \end{cases}$$

- this corresponds to the percentile $P50$ or to the quantile $Q_{.5}$.
- the median is more robust with respect to extreme values than the mean $\bar{x}$

# mode

- the mode is the value which has the maximal absolute frequency

$$\text{mode} = \arg\max_x f(x)$$

there is no `mode`-function in R.

```
x <- table(chickwts$feed)
names(x)[which.max(x)]

## [1] "soybean"
```

# Measures of dispersion, `range()`, `IQR()`

- the range is the distance between the maximal and the minimal value of a sample: max($X$)-min($X$)
- the *p*-quantile ($0 \leq p \leq 1$) or $Q_p$ for a random variable $X$ is that value of $X$ which corresponds to $p \cdot 100\%$ of cumulative frequency. For example, the median is the $Q_{0.5}$
- the interquartil range (IQR) is the distance between $Q_{.25}$ and $Q_{.75}$. (see boxplot)

# Empirical variance, `var()`

- the sample variance or empirical variance is the mean squared deviation from the mean of the sample
- the empirical sample variance $s^2$ is therefore given by

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$

# Sample standard deviation, `sd()`

- if our variable $X$ was body weight, measured in kg, then the sample variance has as a unit $\text{kg}^2$
- to return to the original scale, we take the square root of the variance
- this quantity is known as the sample standard deviation or as the empirical standard deviation $s$:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

# Measures of location and dispersion

```r
colMeans(chickwts[,c("weight","weightkg")])
```

```
##   weight weightkg
##  261.310    0.261
```

```r
colMeans(chickwts[,c(1,3)])
```

```
##   weight weightkg
##  261.310    0.261
```

```r
sd(chickwts$weight)
```

```
## [1] 78.1
```

```r
var(chickwts$weight)
```

```
## [1] 6096
```

```r
range(chickwts$weight)
```

```
## [1] 108 423
```

```r
IQR(chickwts$weight)
```

```
## [1] 119
```

# Variable $A$ with $n = 200$ observations

- `rnorm(n,mu,sigma)` creates $n$ observations from a normal distribution with true mean $\mu$ and true standard deviation $\sigma$
- ```
  A <- rnorm(200,100,20)
  ```
- TASK: repeat all we have done so far in univariate statistics in R with the variable $A$
- Why the empirical mean (and the empirical SD) is not exactly equal to $\mu = 100$ ($\sigma = 20$), the known truth from simulation
- Why do we all have slightly different results?
  ```

# Standardisation

- to be able to compare deviations from the mean, we can normalise these deviations with the standard deviation of the sample
- this is the z-transformation

$$z_i = \frac{x_i - \bar{x}}{s} \qquad i = 1, ..., n.$$

- the new variable $Z$ has mean 0 and standard deviation 1
- the unit of $Z$ is the standard deviation

# Standardisation

- for questions like: what is extreme or normal?
- if a variable $X$ is approximately normal distributed, then $Z$ is standard normal distributed with mean 0 and standard deviation 1
- we can reduce calculations on $X$ on calculations on $Z$
- $Z$ is nothing else than normalized version von $X$

# scale()

```
X <- chickwts$weight
Z <- (X-mean(X))/sd(X)
mean(Z)
```

```
## [1] -2.71e-16
```

```
sd(Z)
```

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

```
Z2 <- scale(X) ## direct version
```

# Quantiles

- important quantiles of the standard normal distribution

| $p$ | 50% | 75% | 90% | 95% | 97.5% | 99% |
|---|---|---|---|---|---|---|
| $z_p$ | 0(median) | 0.67 | 1.28 | 1.64 | 1.96 | 2.33 |

Table: quantiles of the $z$-distribution

- lets look at some important quantiles of $A$:

```
quantile(A,c(.5,.9,.95,.975,.99));

##   50%   90%   95% 97.5%   99%
##  98.2 122.9 130.4 141.7 144.8
```

- and the normalized versions

```
z.p <- quantile(scale(A),c(.5,.9,.95,.975,.99))
z.p

##     50%     90%     95%   97.5%     99%
## -0.0255  1.1733  1.5395  2.0850  2.2393
```

# Quantiles

- the quantiles of $x_p$ can be obtainend by the the inverse transformation of $z_p$:

$$x_p = \bar{x} + s \cdot z_p$$

- ```
  x.p <- mean(A)+sd(A)*z.p
  x.p
  
  ##   50%   90%   95% 97.5%   99%
  ##  98.2 122.9 130.4 141.7 144.8
  ```

- the quantiles of the theoretical standard normal distribution (the table above) can be obtained with `qnorm()`

```
p <-c(.5,.75,.90,.95,.975,.99)
qnorm(p)

## [1] 0.000 0.674 1.282 1.645 1.960 2.326
```
```

# Intervals of the normal distribution, `pnorm()`

- 68% of all observations lie in the interval $\bar{x} \pm s$ (or standardised: $0 \pm 1$)
- 95% of all observations lie in the interval $\bar{x} \pm 1.96s$ (or standardised: $0 \pm 1.96$)
- 99% of all observations lie in the interval $\bar{x} \pm 2.6s$ (or standardised: $0 \pm 2.6$)

```
pnorm(1)-pnorm(-1)

## [1] 0.683

pnorm(1.96)-pnorm(-1.96)

## [1] 0.95

pnorm(2.58)-pnorm(-2.58)

## [1] 0.99
```

# Bivariate data

- joint distribution of two variables $X$ und $Y$
- Quantify the correlation between two variables $X$ und $Y$

# Bivariate data, example

- $X$: alcohol concentration with observations $x_1, x_2, \dots, x_n$
- $Y$: reaction time with observations $y_1, y_2, \dots, y_n$
- sample size: $n = 9$ data pairs

```
A <- c(0.00, 0.20, 0.50, 0.70, 1.00, 1.40, 1.80, 2.25, 2.50);
R <- c(554, 581, 589, 628, 623, 687, 692, 734, 812);
alc <- data.frame(Alc=A,Rct=R)
summary(alc)

##       Alc             Rct
##  Min.   :0.00   Min.   :554
##  1st Qu.:0.50   1st Qu.:589
##  Median :1.00   Median :628
##  Mean   :1.15   Mean   :656
##  3rd Qu.:1.80   3rd Qu.:692
##  Max.   :2.50   Max.   :812
```

- univariate description:
  - $\bar{x} = 1.15$ per mill, $s_x = 0.894$ per mill
  - $\bar{y} = 655.56$ ms, $s_y = 82.97$ ms

# Covariance, `cov()`

- do the two variables present covariance?
- the covariance is a generalisation of the variance and can be quantified as:
$$\widehat{\text{Cov}}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$
- variance is a special case of covariance for the univariate case. (control: set for every $y$ a $x$ in the above formula)
- in our example the covariance is $\widehat{\text{Cov}}(X, Y) = 72.15$

# Correlation, `cor()`

- the size of the covariance is difficult to interpret
- we standardise the covariance by the sample standard deviations $s_x$ und $s_y$
- this leads to the pearson correlation coefficient:

$$r_{X,Y} = \frac{\widehat{\text{Cov}}(X, Y)}{s_X \cdot s_Y}$$

- $r$ is a number between -1 and $+1$ and quantifies the size of the correlation
- the sign gives the direction of the correlation
- in our example, we calculate a large correlation, as expected:

$$r_{X,Y} = \frac{72.15}{0.894 \cdot 82.97} = 0.973$$

# plot()

```
plot(A,R,main="Reaction time versus Alcohol",
     xlab="Alc. concentration",ylab="Reaction time")
```

# built-in data *women*

```
women
```

```
##    height weight
## 1      58    115
## 2      59    117
## 3      60    120
## 4      61    123
## 5      62    126
## 6      63    129
## 7      64    132
## 8      65    135
## 9      66    139
## 10     67    142
## 11     68    146
## 12     69    150
## 13     70    154
## 14     71    159
## 15     72    164
```

# women$height and women$weight

```
plot(women$height,women$weight,xlab="height",
     ylab="weight",main="weight versus height")
```

# Ausblick Schätzen und Testen: `cor.test()`

```
cor(women$height,women$weight,method="pearson")
```

```
## [1] 0.995
```

```
cor.test(women$height,women$weight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  women$height and women$weight
## t = 38, df = 13, p-value = 1e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.986 0.999
## sample estimates:
##   cor
## 0.995
```

# More than two variables, `pairs()`

```r
pairs(swiss)
```

# Other correlation techniques

| Y / X | intervall | ordinal | dichotom |
|---|---|---|---|
| intervall | pearson | spearman | point-biserial |
| ordinal | spearman | spearman | biserial |
| dichotom | point-biserial | biserial | phi-coefficient |

Table: correlation techniques

```r
citation()
```

To cite R in publications use:
R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

# Bibliographie