# Chapter 1
# Introduction

What is statistics?

Statistics is the science of collecting, analyzing and interpreting data

The stages of a statistical study:

- Research question
- Experimental design
- Data collection
- Data analysis
- Interpretation of results
- Presentation of results and conclusions

Aim of this course
Give theoretical and practical insight in the last 3 stages.

Course overview

In each statistical study we need a statistical model.

Data analysis

- get an impression of data
- validate statistical model
- summarize data (descriptive statistics)
- analyze (e.g. estimate/test parameters in model)

Interpretation of results

- this is not always straightforward…

Presentation of results and conclusions

- translate back to the experimental context

Course overview

For the data analysis we discuss:

- Summarizing data (Chapter 2)
- Exploring distributions (Chapter 3)
- Bootstrap methods (Chapter 4)
- Robust estimators (Chapter 5)
  *Relatively insensitive to small deviations from the assumptions*
- Nonparametric tests (Chapter 6)
- Analysis categorical data (Chapter 7)
- Multiple linear regression (Chapter 8)

Interpretation and presentation of results and conclusions are practiced in the assignments.

# Chapter 2
# Summarizing data

## 2.1 Data

Data: quantified measurements of a study.

Data is typically stored in variables.

Variable: a property of an individual/object that can be measured

Variables can be

- measured on different scales,
- univariate, bivariate or multivariate,
- dependent or independent.

Measurement scales of variables

Qualitative variable

- nominal (e.g. gender)

*Location measures like median or mean and spread measures have no meaning.*

- ordinal (e.g. level of education)

*The categories can be ordered, without measurable distances.*
*The median and the mode can be useful, but the mean and spread measures have no meaning.*

*Nominal and ordinal variables are discrete by definition.*

Quantitative variable

discrete

- interval (e.g. date)
- ratio (e.g. counts)

continuous

- interval (e.g. temperature in Celsius)
- ratio (e.g. duration of this lecture, temperature in Kelvin)

*For quantitative variables the location measures (mean, mode and median) and the spread measures can all be used.*

*For interval scales only intervals are meaningful, ratios are not. Differences are meaningful.*

*For ratio scales both intervals and ratios are meaningful. There is a zero.*

Number of characteristics measured in one variable

- univariate (e.g. gender)
- bivariate (e.g. gender and level of education)
- multivariate (e.g. gender, level of education, shoe size, age, height)

Role of the variables

- dependent: variable of interest
- independent: variables containing background information

***Example 2.2*** *In a study about the dependence of political opinions on variables like age, sex, or religion, the political opinion is the dependent variable and answers to a question about political opinion are the values of the dependent variable. Age, sex, religion, and so on, are the independent variables.*

## 2.2 Summarizing data

A good summary shows at least

- location, scale
- range, extremes
- holes, modes
- symmetry

Additionally it may answer the following questions

- are data rounded?
- are data from a known distribution?
- do we need to divide the data into groups?
- is there influence of other variables, like time?
- what is the relation between variables?

### 2.2.1 Summarizing Univariate Data

Graphical summaries

- histogram

  *Too few or too many bin intervals gives a bad result. Sometimes it's better to choose a histogram with less bin intervals, although more data information gets lost. It gives a better impression of the global spread of the data.*

- stem-and-leaf-plot

*With stem-and-leaf plots little information about the data gets lost. Stem-and-leaf plots give an impression about the shape of the data distribution while retaining most of the numerical information.*

- empirical distribution function

*The empirical distribution function is defined as*

$$\hat{F}_n(x)=\frac{1}{n}\sum_{j=1}^{n} 1_{\{x_j\le x\}}$$

*If* $x<x_{(1)}$ *then* $\hat{F}_n(x)=0$ *, if* $x_{(1)}\le x<x_{(2)}$ *then* $\hat{F}_n(x)=\frac{1}{n}$ *, if* $x_{(2)}\le x<x_{(3)}$ *then* $\hat{F}_n(x)=\frac{2}{n}$ *, and so on.*

- boxplot

*combination of a graphical and a numerical summary (it also contains numerical info like quartiles, interquartile range, median)*

Numerical summaries

| | | |
|---|---|---|
| **sample size** | | $n$ |
| **location** | *mean* | $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ |
| | *median* | $\text{med}(x)=\begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ odd} \\ \frac{1}{2}(x_{(n/2)}+x_{(n/2+1)}), & \text{if } n \text{ even} \end{cases}$ |
| **scale** | *variance* | $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ |
| | *standard deviation* | $s=\sqrt{s^2}$ |
| | *coefficient of variation* | $cv = s/\bar{x}$ |
| | *range* | $(x_{(1)}, x_{(n)})$ |
| | *quartiles* | $\text{quart}(x)$, $3\text{quart}(x)$ |
| | *interquartile range* | $3\text{quart}(x) - \text{quart}(x)$ |
| **skewness** | *skewness* | $b_1 = \frac{\sqrt{n}\sum_{j=1}^{n}(x_j-\bar{x})^3}{\{\sum_{j=1}^{n}(x_j-\bar{x})^2\}^{3/2}}$ |
| **size of tails** | *curtosis* | $b_2 = \frac{n\sum_{j=1}^{n}(x_j-\bar{x})^4}{\{\sum_{j=1}^{n}(x_j-\bar{x})^2\}^{2}}$ |

**mode**: the location of the maximum of the probability density of the distribution.

The **skewness** and the **curtosis** give an idea of the asymmetry and the size of the tails, respectively, of the distribution.

Univariate Summaries Example

Example data incomes of 100 white families and 100 colored families in US.

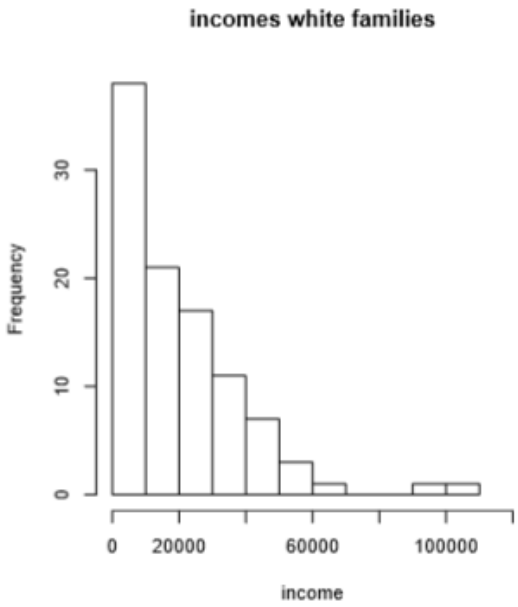
| | | | | | | |
|---|---|---|---|---|---|---|
| white | 148 | 202 | 540 | 541 | ··· | 102,909 |
| color | 129 | 237 | 288 | 294 | ··· | 49,185 |

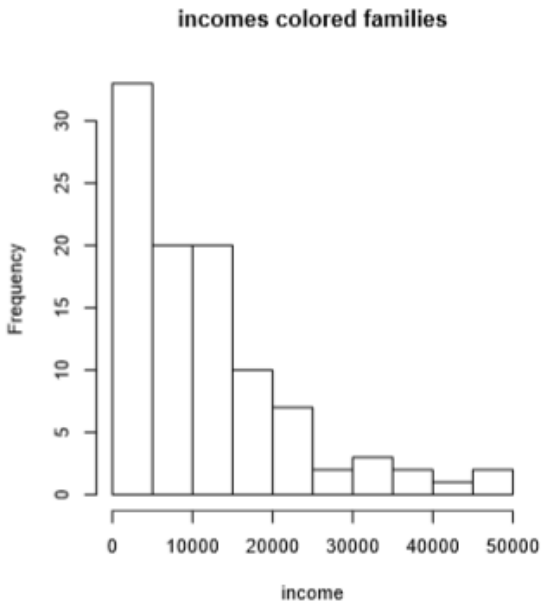
Two univariate data sets (not one bivariate data set).

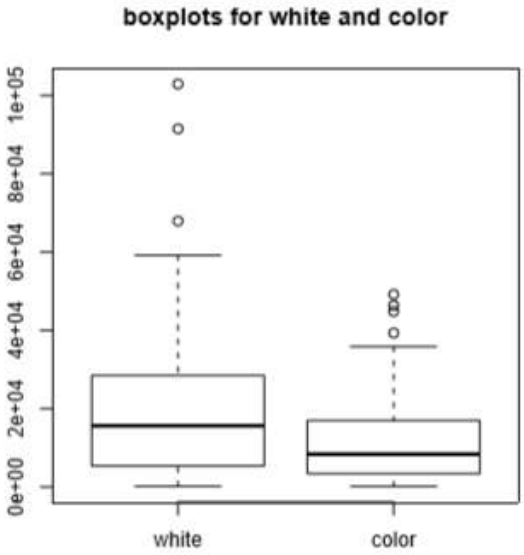
## Numerical summaries

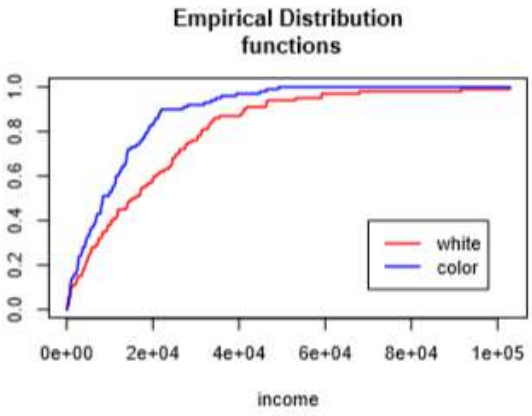
| white families | | colored families | |
|---|---|---|---|
| sample size | 100 | sample size | 100 |
| mean | 19,868 | mean | 11,670 |
| median | 15,614 | median | 8,372 |
| sd | 18,824 | sd | 10,811 |
| var | 354,344,281 | var | 116,884,650 |
| min | 148 | min | 129 |
| max | 102,909 | max | 49,180 |
| IQR | 22,814 | IQR | 13,421 |

## Graphical summaries



### 2.2.2 Summarizing Bivariate Data

Graphical summaries

- scatter plot

*scatter plots are used to plot the values of one quantitative variable against the corresponding values of the other quantitative variable. They can help us to detect relationships between the variables, like linear or quadratic relations, to find extreme values, or to determine clusters of observations.*

- time plot

*the time plot is a scatter plot of the data against time*

- contingency table

| blood group | stomach ulcer | stomach cancer | control | total |
|---|---|---|---|---|
| O | 983 | 383 | 2892 | 4258 |
| A | 679 | 416 | 2625 | 3720 |
| B | 134 | 84 | 570 | 788 |
| total | 1796 | 883 | 6087 | 8766 |

Table 2.3: $3 \times 3$ contingency table of blood group against disease of 8766 persons.

Whereas in a scatter plot the individual data values can still be recognized, in a contingency table this information may get lost (when the categories consist of more than one value). The advantage of contingency tables is that they can be used not only to summarize data that are measured on a quantitative scale, but also to summarize data that are measured on a nominal or ordinal scale.

Numerical summaries

| | |
|---|---|
| *mean* | $(\bar{x}, \bar{y})$ |
| *covariance* | $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ |
| *correlation coefficient* | $r_{xy} = \frac{s_{xy}}{s_x s_y}$ |
| *covariance matrix* | $\Sigma = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$ |
| *Spearman's rank correlation coefficient* | $r_s = \frac{\sum_{i=1}^{n} (r_i - \frac{1}{2}(n+1))(t_i - \frac{1}{2}(n+1))}{\sqrt{\sum_{i=1}^{n} (r_i - \frac{1}{2}(n+1))^2 \sum_{i=1}^{n} (t_i - \frac{1}{2}(n+1))^2}}$ |
| *Kendall's rank correlation coefficient* | $\tau = \frac{\sum\sum_{i \neq j} \operatorname{sgn}(r_i - r_j)\operatorname{sgn}(t_i - t_j)}{n(n-1)} = \frac{4N_\tau}{n(n-1)} - 1$ |

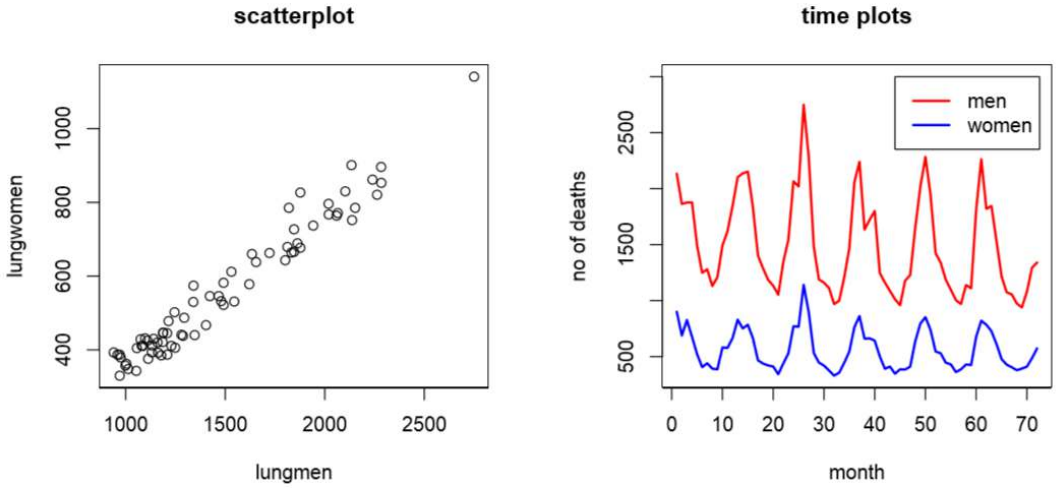
Table: Numerical summaries of bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$ . $(r_1, \ldots, r_n)$ and $(t_1, \ldots, t_n)$ are the rank vectors of $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ , in the ordered samples $(x_{(1)}, \ldots, x_{(n)})$ and $(y_{(1)}, \ldots, y_{(n)})$ , respectively. The quantity $N_\tau$ is the number of pairs $(i, j)$ with $i<j$ for which either $x_i<x_j$ and $y_i<y_j$ , or $x_i>x_j$ and $y_i>y_j$ ("concordant"). Let $z>0$ . The *sign function* $sgn(z)=1, sgn(-z)=-1, sgn(0)=0$ .

The **sample correlation coefficient** $r_{xy}$ is a measure of the strength of the linear relationship between $x$ and $y$ . It can take values from -1 to 1. A $r_{xy}$ -value close

to−1 means that there is a strong negative linear relation between x and y. Equality to −1 or 1 means that the relationship is exactly linear.

Bivariate Summaries Example

Example data number of deaths due to lung diseases in the UK between 1974 and 1979, registered monthly




### 2.2.3 Summarizing Multivariate Data

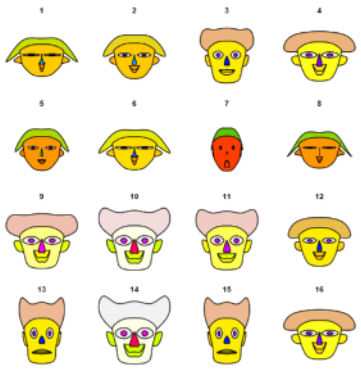
Graphical summaries

- scatter plot
- contingency tables

For a multivariate data set it is often useful to make scatter plots or contingency tables for all (relevant) **pairs** of variables. With 7 variables this yields already 21 graphs to study. Moreover, it can be misleading to project higher-dimensional data into two dimensions.

Multivariate Summaries Examples

Example: Chernoff faces display multivariate data in the shape of a human face.



Chernoff faces handle each variable differently: the individual parts, such as eyes, ears, mouth and nose represent values of the variables by their shape, size, placement and orientation.
Idea humans easily recognize faces and notice small changes without difficulty.

# Chapter 3
# Exploring distributions

In order to explore distributions we will discuss

- Quantile function
- Location-scale family
- QQ-plots and symplots

Distribution functions

Let us focus on a certain (real-valued) random variable $X$, e.g. body weight of a random individual in this lecture room.

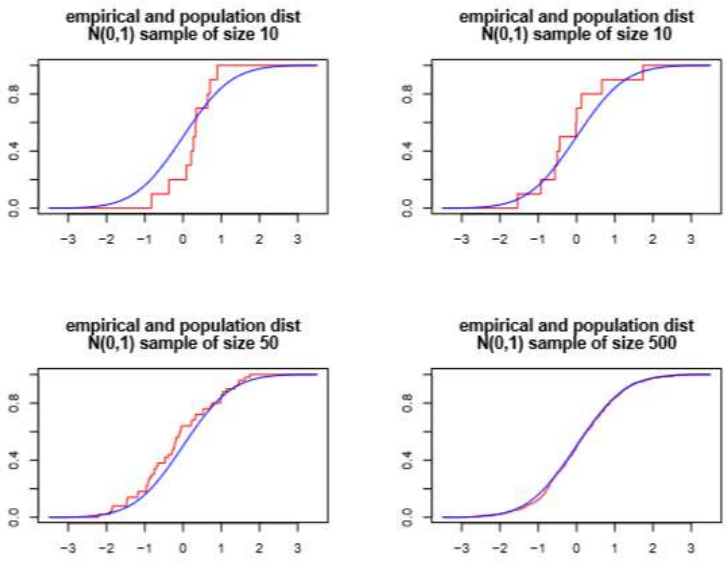
The population distribution function $F$ is the underlying distribution of the variable in the population. That is, $F(x)$ is the probability that $X$ isn't greater than $x \in R$ in this population.

The empirical distribution function $\hat{F}_n$ is the distribution of the variable in the sample $x_1, \ldots, x_n$

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{x_j \le x\}}.$$

When generating a sample from a given distribution, the empirical distribution $\hat{F}_n$ will vary around the population distribution $F$. The larger the sample, the smaller this variation.

Empirical and population distribution function of a sample with underlying distribution $N(0,1)$




Goal

The empirical distribution helps to determine the underlying (population) distribution. This underlying distribution is usually part of a (parametric) statistical model. Hence, the empirical distribution helps to check or set up a statistical model.

Goal: find underlying distribution

Type of questions that we deal with in this chapter:

One sample data

- Do data originate from a specific distribution? (QQ-plot, goodness-of-fit tests)
- Is the underlying distribution symmetric? (symplot)

Two sample data

- Do both data sets originate from same distribution? (QQ-plot)

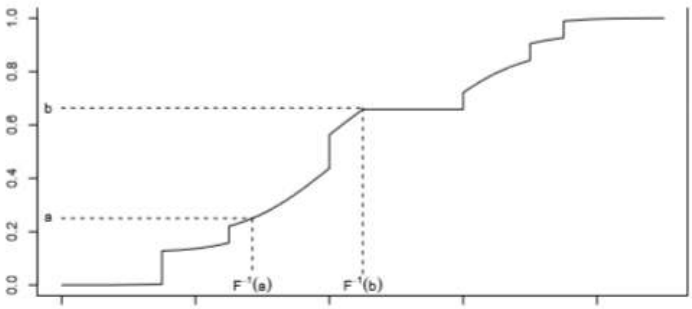
## 3.1 The quantile function and location-scale families

Quantile functions

If for a given $\alpha \in (0,1)$ there exists exactly one $x_a \in R$ such that $F(x_a) = \alpha$, then $x_a$ is called the $\alpha$-quantile of F. The $\alpha$-quantile is denoted by $F^{-1}(\alpha)$. As this notation suggests, the quantile function is the function $\alpha \mapsto F^{-1}(\alpha)$, the inverse of F, if this inverse is well-defined. This is the case when F is a strictly increasing function.

Apart from strictly increasing pieces, a cumulative distribution function can have jumps as well as constant pieces. Therefore, for fixed $\alpha$ the equation $F(x) = \alpha$ has exactly one, none or infinitely many solutions. In order to be able to define the $\alpha$-quantile in the latter two cases, the quantile function of F is in general defined by

Definition $$F^{-1}(\alpha) = inf\{x : F(x) \geq \alpha\}, \alpha \in (0,1).$$




*R: qnorm, qexp, qpois etc.*

Location-scale family

If a random variable $X$ has distribution $F$, then the distribution of $Y = a + bX$ is $F_{a,b}$, $a \in R, b > 0$, given by

$$F_{a,b}(x)=F\left(\frac{x-a}{b}\right)$$

The collection of distributions $\{F_{a,b}: a\in R, b>0\}$ is called the location-scale family corresponding to $F$.

If $EX=0$ and $varX=1$ then $EY=a$ and $varY=b^2$.

Quantiles of F and Fa,b

Claim There is a linear relation between $F^{-1}(\alpha)$ and $F_{a,b}^{-1}(\alpha)$:

$$F_{a,b}^{-1}(\alpha)=a+b\,F^{-1}(\alpha)$$

Proof (for invertible $F$)

$$\alpha=F_{a,b}\left(F_{a,b}^{-1}(\alpha)\right)=F\left(\frac{F_{a,b}^{-1}(\alpha)-a}{b}\right)$$

$$F^{-1}(\alpha)=F^{-1}\left(F\left(\frac{F_{a,b}^{-1}(\alpha)-a}{b}\right)\right)=\frac{F_{a,b}^{-1}(\alpha)-a}{b}$$

$$F_{a,b}^{-1}(\alpha)=a+b\,F^{-1}(\alpha)$$

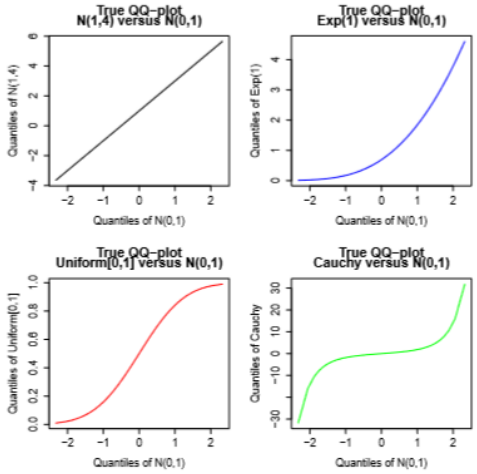
Because $F_{a,b}^{-1}(\alpha)=a+b\,F^{-1}(\alpha)$, the points $\{(F^{-1}(\alpha),F_{a,b}^{-1}(\alpha)):\alpha\in(0,1)\}$ are on the straight line $y=a+bx$.

$N(0,1)$ and $N(1,4)$ are in the same location-scale family.

$N(0,1)$ and $\exp(1)$ are not in the same location-scale family.

$N(0,1)$ and $Uniform(0,1)$ are not in the same location-scale family. Normal distributions have heavier tails.

$N(0,1)$ and $Cauchy(1)$ are not in the same location-scale family. Cauchy distributions have heavier tails.




### 3.2 QQ-plots

For independent random variables $X_1,\ldots,X_n$ with (continuous) distribution $F$, we have $E\,F(X_{(i)})=\frac{i}{n+1}$. So $X_{(i)}\approx F^{-1}(\frac{i}{n+1})$.

If $Y_1,\ldots,Y_n$ with $Y_i=a+bX_i$ have distribution $F_{a,b}$, we have $EF_{a,b}(Y_{(i)})=\frac{i}{n+1}$.

So $Y_{(i)}=F_{a,b}^{-1}\left(\frac{i}{n+1}\right)=a+bF^{-1}\left(\frac{i}{n+1}\right)$. Hence, the points

$$\left\{\left(F^{-1}\left(\frac{i}{n+1}\right),y_{(i)}\right):i=1,\ldots,n\right\}$$

are approximately on the straight line $y=a+bx$.

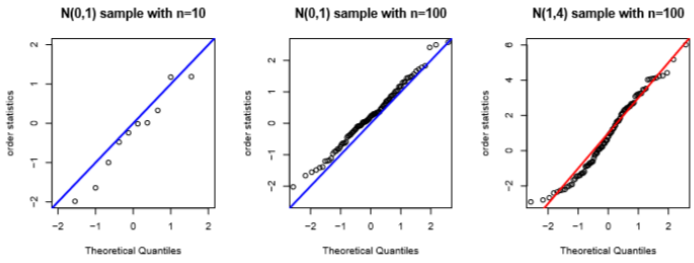
In practice $F$ is unknown. A QQ-plot is a plot of these points for some chosen distribution $F$.

*R: qqnorm, qqexp, qqunif, etc.*

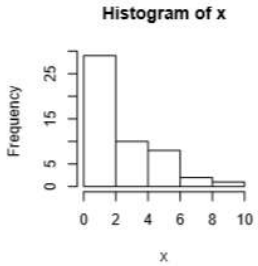
A QQ-plot yields a method to judge whether the data come from a certain distribution by only looking at the plot. When the plot yields approximately the line $y=x$, this is an indication that the data come from the distribution $F$. Deviations from the line $y=x$ indicate differences between the true distribution of the data and $F$. The kind of deviation from $y=x$ suggests the kind of difference between the true distribution and $F$. The simplest case of such a deviation is when the QQ-plot is a straight line but not the line $y=x$, as in Figure 3.3. This is an indication that the data do not originate from $F$, but come from another member of the location-scale family of $F$. Interpreting a bent curve is more complicated. Such QQ-plots mainly yield information about the relative thickness of the tails of the distribution of the data with respect to those of $F$.

Example QQ-plot

Example QQ-plot of $X_1,\ldots,X_n$ $N(\mu,\sigma^2)$ against $N(0,1)$ for varying $n,\mu$ and $\sigma^2$.

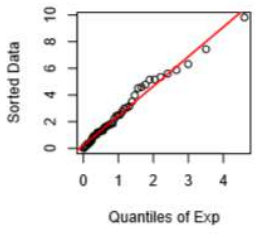



$y=x$ $y=x$ $y=1+2x$




Using QQ-plots

- plot a histogram
- plot different QQ-plots and choose the most linear one



- determine location (a) and scale (b)
  - by visually fitting a straight line $y=a+bx$
  - by fitting sample mean and variance to theoretical values

Example

- $y=0+2x$
- $\acute{X}=1.98 \quad , \quad var(X)=4.2$

$\exp(1/2)$ is a suitable distribution for this sample.

### 3.3 Symplots

The symmetry plot is used to investigate symmetry or skewness of a distribution.

A random variable $X$ is symmetrically distributed around $\theta$ if $X-\theta$ and $\theta-X$ follow the same distribution.

To judge whether or not a sample originates from a symmetric distribution, a **histogram** or a **stem-and-leaf plot** can be used. Naturally, the **skewness parameter** gives information about symmetry too, although in spite of its name one should not overestimate its usefulness. Also a large **difference between mean and median** indicates a skewed distribution.

Skewness can also be assessed with the quantile function.

If $F$ is symmetric around $\theta$ , we have

$$F^{-1}(1-\alpha)-\theta=\theta-F^{-1}(\alpha), \alpha \in (0,1).$$

Hence, the points $\{(\theta-F^{-1}(\alpha), F^{-1}(1-\alpha)-\theta): \alpha \in (0,1)\}$ lie on the straight line $y=x$ .

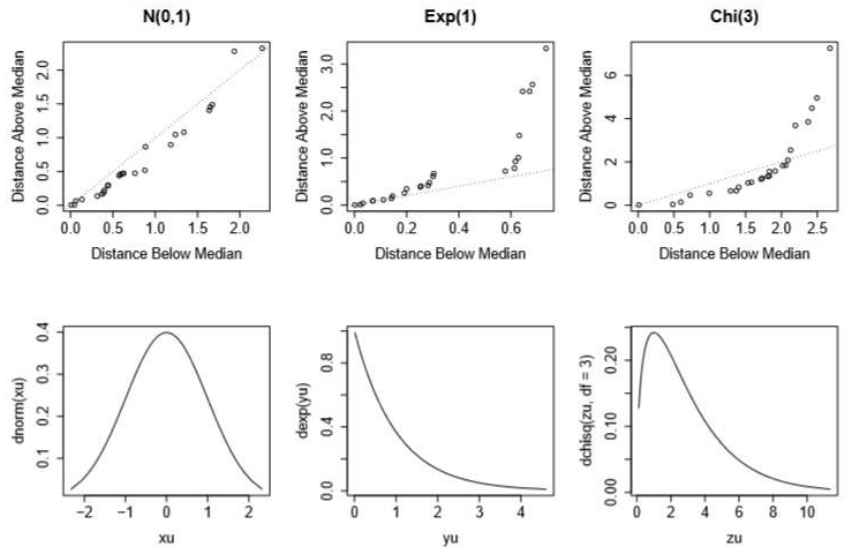
For data from a symmetric distribution we expect that the

$$\left\{\left(med(x)-x_{\lfloor i \rfloor}, x_{\lfloor n-i+1 \rfloor}-med(x)\right): i=1,\ldots,\lfloor \frac{n}{2} \rfloor\right\}.$$

also lie on the straight line $y=x$ . A plot of these points is called a symmetry plot or, briefly, a symplot.

*R: symplot*

Example symplot

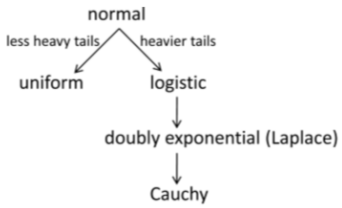



As in the case of QQ-plots, sample size matters!

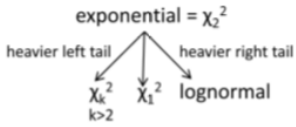
First is symmetrical, second and third clearly have skewed distributions since they don't follow the line $y=x$ .

Systematic search for underlying distribution

- Investigate symmetry with symmetry plot (and histogram)
- Try several QQ-plots
  - If symmetric:




  - If not symmetric:




- If not satisfactory, try transformations

### 3.4 Two-sample QQ-plots

Two sample QQ-plots (also called an empirical QQ-plot) are used to investigate whether two samples $x_1,\ldots,x_m$ and $y_1,\ldots,y_n$ originate from the same distribution, or from distributions in the same location-scale family.

If $m=n$ , then this empirical QQ-plot is a plot of the points $\{(x_{(i)},y_{(i)}):i=1,2,..,n\}$.

If $m<n$ , it is a plot of the points $\{(x_{(i)},y_{(i)}^{¿}):i=1,2,..,m\}$ , where

$$\hat{y}_{(i)} = \frac{1}{2}\left( y_{\left(\left\lfloor i\frac{n+1}{m+1}\right\rfloor\right)} + y_{\left(\left\lfloor i\frac{n+1}{m+1}+\frac{m}{m+1}\right\rfloor\right)} \right).$$

The idea is just to match $x_{(i)}$ with $y_{(j)}$ for which $\frac{i}{m+1} \approx \frac{j}{n+1}$ .

The reasoning is similar to one-sample QQ-plots: roughly a straight line indicates that it is plausible to assume that the two samples are from distributions in the same location-scale family.

*R: qqplot*

### 3.5 Goodness of fit tests

Idea: Assume we consider a sample $x_1, \ldots, x_n$ from an **unknown** distribution $F$ .

Another (more formal) way to check whether this data comes from a 'known' distribution is a goodness of fit test. With these tests, the null hypothesis that the data comes from a certain distribution $F$ , or from a member of a certain family of distributions, can be tested against the alternative hypothesis that this is not the case:

$H_0 : F \in F_0$

$H_1 : F \notin F_0$

where either $F_0$ = { $F_0$ } (simple $H_0$ ) or $F_0$ is a small collection of distributions (composite $H_0$ ), e.g. a location-scale family.

We look for an omnibus test that has reasonable power in most of the alternatives. When such a test does not reject the null hypothesis, this is considered as an indication that the null hypothesis may be correct.

In a lot of situations we use a goodness-of-fit test to show that $H_0$ is plausible, i.e. we're happy if we don't reject, it "confirms" our statistical model. (Warning: this is actually never true with real data!)

Different tests

The tests we consider:

- Shapiro-Wilk test for $H_0 : F \in \{ N(\mu, \sigma^2); \mu \in R, \sigma^2 > 0 \}$
- Kolmogorov-Smirnov test for simple $H_0$ and adjusted for composite $H_0$
- Chi-square test for simple $H_0$

These tests use different test statistics, with different distributions under $H_0$ .

When you perform a test, state clearly the null hypothesis $H_0$, the alternative hypothesis $H_1$, the chosen significance level α, the test statistic, its distribution under $H_0$, the p-value or critical region and the conclusion.

### 3.5.1 Shapiro-Wilk Test

The Shapiro-Wilk test for $H_0: F \in \{N(\mu, \sigma^2); \mu \in R, \sigma^2 > 0\}$
The Shapiro-Wilk test is meant for testing the null hypothesis that the observations are independent and originate from a normal distribution. This is a composite $H_0$.

The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \acute{X})^2} \in ¿$$

with $a_1, \ldots, a_n$ constants based on the covariance of the order statistics.

Possible values for test statistic $W$ are $W \in ¿$. $H_0$ is rejected for "small" values of $W$.

The denominator of the test statistic $W$ is equal to: $\sum_{i=1}^{n} (X_i - \acute{X})^2 = (n-1) S_n$, where $S_n$ is the sample variance.

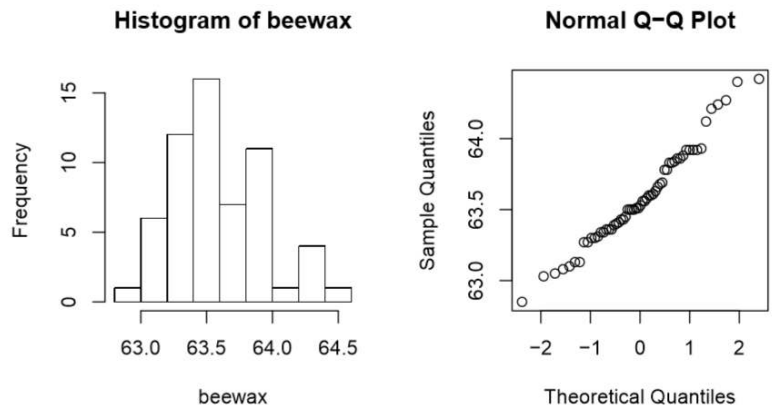
Distribution of $W$ under $H_0$ is known from tables (or R).

*"Suppose that the distribution of $W$ is not available and we have to use a bootstrap test. Describe the steps that are made in a bootstrap test for the $H_0$ using $W$ as test statistic."*
Simulate the distribution of $W$ under $H_0$. Test statistic $W$ is non-parametric under $H_0$, in other words, W has the same distribution for any underlying distribution.
Procedure: For $i = 1, \ldots, B$ we do the following:

1. Generate a sample $X_1^{¿}, .., X_n^{¿}$ $N(0,1)$
2. Compute $W_i^{¿} = W(X_1^{¿}, .., X_n^{¿})$
3. Compute the bootstrap p-value
$$\frac{¿\, W_i^{¿} : W_i^{¿} < W(X_1, \ldots, X_n)}{B}$$

Example Shapiro-Wilk test

Example Beewax data consisting of melting points (in Celsius) of 59 samples of beewax.




Is normality an adequate assumption?

The Shapiro-Wilk test applied with significance level α=0.05 on beewax data

```
> shapiro.test(beewax)

        Shapiro-Wilk normality test

data: beewax
W = 0.9748, p-value = 0.2579
```

Since the p-value is bigger than the significance level (0.2579 > 0.05) we do not reject our null hypothesis.

*R: shapiro.test*

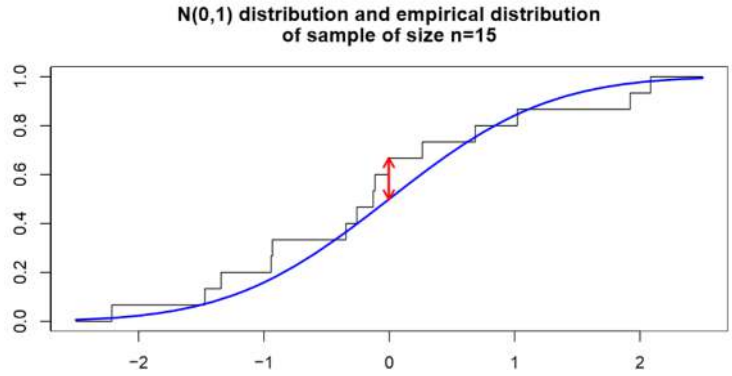
### 3.5.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test for

$H_0: F = F_0$

$H_1: F \neq F_0$

This is a simple $H_0$ . The test statistic is based on the maximum vertical distance between $\hat{F}_n$ and $F_0$ :




The test statistic is :

$D_n = \sup_{-\infty < x < \infty} \left| \hat{F}_n(x) - F_0(x) \right|.$

$H_0$ is rejected for large values of $D_n$ .

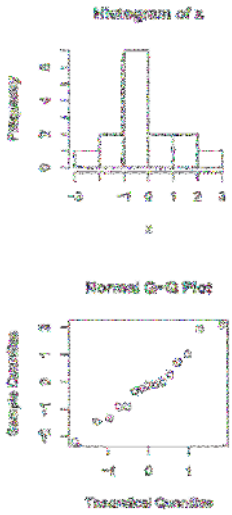
The distribution of $D_n$ under $H_0$ depends on $n$, though is independent of $F_0$ for $F_0$ a continuous distribution,

$$D_n = \max_{1 \le i \le n} \max\left\{ \left| \frac{i}{n} - F_0\left(X_{(i)}\right) \right|, \left| \frac{i-1}{n} - F_0\left(X_{(i)}\right) \right| \right\}.$$

Therefore, this test is nonparametric, or distribution free over the class of continuous functions. The Kolmogorov-Smirnov test can only be used to test a simple hypothesis about a continuous distribution.

*R: ks.test*

Example Kolmogorov-Smirnov test




Test $H_0: X_1, \ldots, X_n \sim N(0,1)$

> ks.test(x,pnorm,**0,1**)

One-sample Kolmogorov-Smirnov test

data: x
D = 0.1681, p-value = 0.73
alternative hypothesis: two-sided

$H_0$ is not rejected, since the p-value is large.

How not to use the Kolmogorov-Smirnov test

In order to test the composite $H_0$ of normality (i.e. the complete location-scale family), the KS-test cannot be used.

The next application of the KS-test IS REALLY WRONG!!

> ks.test(x,pnorm,**mean(x),sd(x)**)

One-sample Kolmogorov-Smirnov test

data: x
D = 0.1287, p-value = 0.9378
alternative hypothesis: two-sided

An adjusted (bootstrap) version of the KS-test for testing normality will be discussed.

### 3.5.3 Chi-square test

The chi-square goodness-of-fit test for

$H_0: F = F_0$
$H_1: F \neq F_0$

This is a simple $H_0$ .

For a chi-square test, the real line is divided into adjacent intervals $I_1, \ldots, I_k$ .
The test statistic is based on the difference between observed and expected number of observations in intervals $I_1, \ldots, I_k$ .


The test statistic for a sample size $k$ is

$$X^2 = \sum_{i=1}^{k} \frac{[N_i - n p_i]^2}{n p_i}$$

where $N_i$ is observed number of observations in $I_i$ and $p_i = F_0(I_i)$ , so that $n p_i$ is the expected number of observations in $I_i$ .

$H_0$ is rejected for large values of $X^2$ .

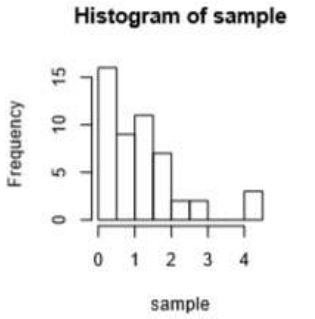
The distribution of $X^2$ under $H_0$ is asymptotically $\chi^2_{k-1}$ . This approximation is reliable when $n p_i \geq 5$ for all $i$ (Rule of thumb). The chi-square test is distribution free, because the distribution of $X^2$ does not depend on $F_0$ .

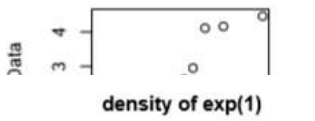
*R: chisquare (on Canvas)*

Example chi-square test (1)

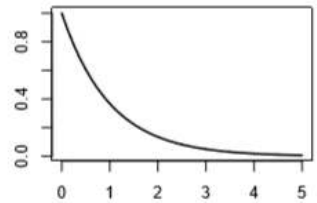
```
> range(sample)
[1] 0.02910324  4.46345348
> length(sample)
[1] 50
> chisquare(sample, pexp, 10, 0, 5)
$chisquare
[1] 26.30088
$pr
[1] 0.001823704
$N
(0,0.5] (0.5,1] (1,1.5] (1.5,2] (2,2.5]
16        9        11       7        2
(2.5,3] (3,3.5] (3.5,4] (4,4.5] (4.5,5]
   2        0        0        3        0
```

$np
[1] 19 11 7 4 2 1 0 0 0 0



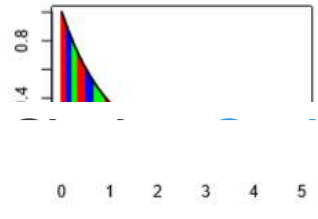

density of exp(1)




Example chi-square test (2)

```
> b
[1] 0.0 0.1 0.2 0.4 0.5 0.7 0.9 1.2 1.6 2.3 Inf
> chisquare(sample, pexp, 10, 0, 5,b)
```

probability mass
devided in 10 parts



```
$chisquare
[1] 13.6
$pr
[1] 0.1372824
$N
      (0,0.105] (0.105,0.223] (0.223,0.357]
              2             5             6
(0.357,0.511] (0.511,0.693] (0.693,0.916]
              3             1             5
   (0.916,1.2]     (1.2,1.61]     (1.61,2.3]
              6            11             6
      (2.3,Inf]
              5
$np
[1] 5 5 5 5 5 5 5 5 5 5
```

# Chapter 4
# The bootstrap

The bootstrap is a technique which can be used for:

- investigating the variance of an estimator
- computing confidence intervals
- determining critical values of test statistics.

## 4.1 Bootstrap estimators for a distribution

Why bootstrap?

Suppose that a set of random variables $X_1, \ldots, X_n$ is available and that one is interested in a function of these random variables, the random variable $T_n = T_n(X_1, \ldots, X_n)$, and in particular in its distribution. This random variable $T_n$ is, for example, an estimator or a test statistic, but it may also depend on an unknown parameter. When $T_n$ is an estimator, then from its distribution a measure for its accuracy, like its variance, can be derived. In the case that $T_n$ is a test statistic, the critical values of the test follow from the distribution of $T_n$ under the null hypothesis. In general, the distribution of $T_n$ is unknown, so that it is an important problem to estimate it from the data

Summarized:

$X_1, \ldots, X_n$ from an unknown distribution $P$ .

The location of $P$ can be described by the population mean $\mu_P$ .

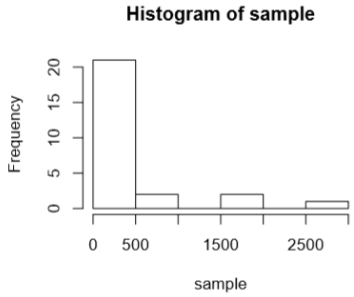
$T_n = \acute{X}$ is an unbiased estimator of $\mu_P$ .
$T_n$ is a random variable (stochastic)
$T_n$ has a distribution $Q$ that depends on $P$ , so we call it $Q_P$ .

What is this distribution (e.g. the variance) of $T_n$ ?

We can use bootstrap techniques to estimate $Q_P$ .

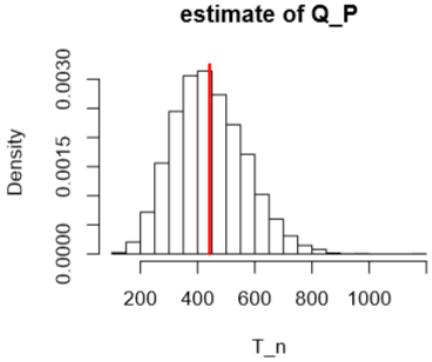
Bootstrap example

Example $X_1, \ldots, X_n$ are data from cloud seeding.




> mean(sample)
[1] 441.9846
Estimate of $\mu_P$ is $\acute{X} = 442$ .

How accurate is this number? Does it mean $\mu_P \in [420, 460]$ or $\mu_P \in [300, 580]$?




We can use bootstrap techniques to estimate $Q_P$ .

Bootstrap scheme

Given data $X_1, \ldots, X_n$ $P$
Bootstrap scheme in 3 steps:

1. Samples
2. Estimators
3. Estimate distribution $Q_P$

Use data $X_1, \ldots, X_n$ to estimate $P$ by $\widetilde{P}$ *(empirical or parametric)*.

Bootstrap sampling scheme

The bootstrap sampling scheme in 3 steps

1. Generate $B$ times a sample $X^*_1, \ldots, X^*_n$ from $\widetilde{P}$
2. Compute for the $i$ -th sample the value $T^*_i = T_n(X^*_1, \ldots, X^*_n)$ , for $i = 1, \ldots, B$
3. Use the empirical distribution $\widetilde{Q}_{\widetilde{P}}$ of the sample $T^*_1, \ldots, T^*_B$ as estimate for $Q_{\widetilde{P}}$ (which is an approximation of $Q_P$ ).

In the last step you can use e.g. the sample variance of $T^*_1, \ldots, T^*_B$ as estimate of the variance of $T$ .

Bootstrap types

Estimate $P$ by $\widetilde{P}$

We have two possibilities:

- $\widetilde{P} = \hat{P}_n$ (empirical distribution)

Empirical bootstrap

*The empirical distribution $\hat{P}_n$ is a simple estimator for P: new samples are created by simply resampling from the original sample. This is the best estimator when nothing is known about the unknown distribution.*

- $\widetilde{P} = P_{\hat{\theta}}$ (estimated parametric distribution)

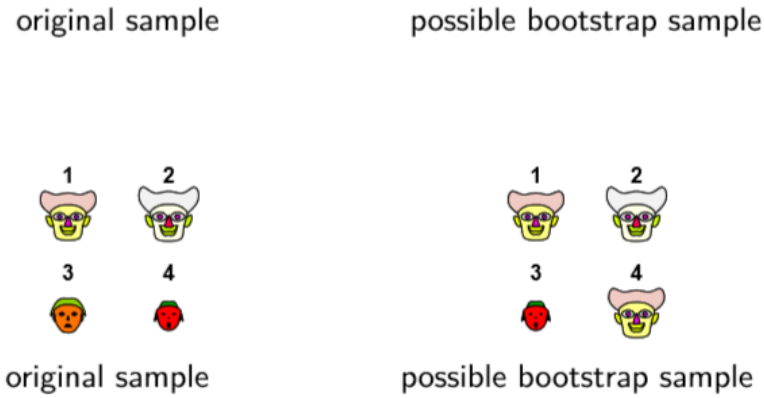
Parametric bootstrap

*This estimator is appropriate in situations where the unknown distribution $P$ is known to belong to a parametric family like the normal or exponential family, but its parameter value $\theta$ is unknown. In this case it is natural to first find an estimator $\hat{\theta}_n$ of $\theta$ , and then estimate $P$ by the distribution in the family which has $\hat{\theta}_n$ as its parameter value. This parametric estimator of $P$ will be denoted by $P_{\hat{\theta}_n}$ . The estimator $Q_{P_{\hat{\theta}_n}}$ for the distribution of $T_n$ is a parametric bootstrap estimator.*

Remarks on bootstrap notation

- $P$ depends on $n$ , so we should write $\widetilde{P}_n$ and $Q_{\widetilde{P}_n}$
- $\hat{\theta}$ depends on $n$ , so we should write $\hat{\theta}_n$ , $P_{\hat{\theta}_n}$ and $Q_{P_{\hat{\theta}_n}}$
- $T^*_i$ depends on $n$ , so we should write $T^*_{n,1}, \ldots, T^*_{n,B}$

Example empirical bootstrap

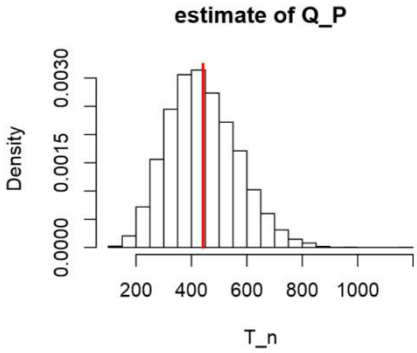



Sample with replacement.

Example empirical bootstrap

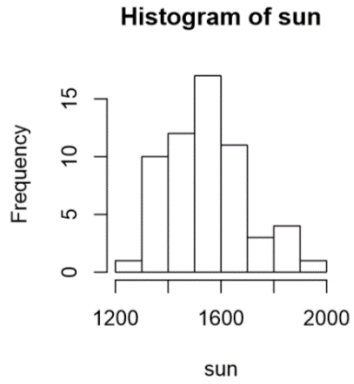
Example The empirical bootstrap scheme (i.e. using $\widetilde{P}=\widehat{P}_n$ ) in the case of the clouds data:

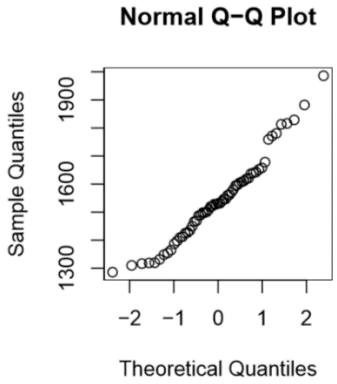
```
> B=1000
> Tstar=numeric(B)
> for(i in 1:B){
+ xstar=sample(clouds[,1], replace=TRUE)
+ Tstar[i]=mean(xstar)
+ }
> hist(Tstar)
> sd(Tstar)
[1] 125.5883
```




Example parametric bootstrap

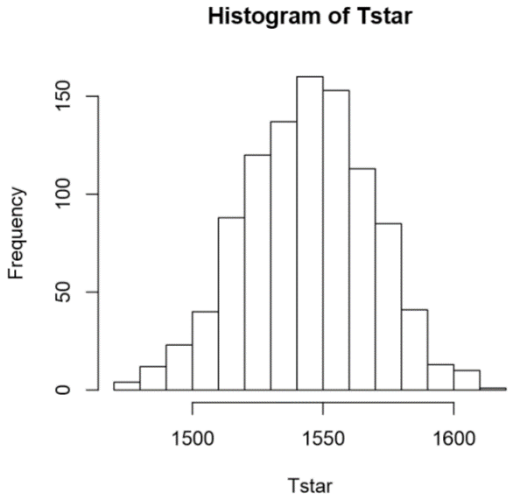
Example The yearly number of sun hours in De Bilt in 1920-1978. We want to estimate the standard deviation of the sample median.




We assume the numbers are normally distributed.

```
> median(sun)
[1] 1531
> mean(sun)
[1] 1543.797
> sd(sun)
[1] 153.9447
> var(sun)
[1] 23698.97
> length(sun)
[1] 59
> B=1000
> Tstar=numeric(B)
> for(i in 1:B){
+     xstar=rnorm(59,1543.797,153.9447)
+     Tstar[i]=median(xstar)}
> hist(Tstar)
> sd(Tstar)
[1] 24.36120
```




<u>Two types of bootstrap errors</u>

We make estimation errors in the two steps:

1. Estimate $P$ by $\widetilde{P}$ (and hence, $Q_P$ by $Q_{\widetilde{P}}$ )
2. Estimate $Q_{\widetilde{P}}$ by the empirical distribution of a sample $T^{\acute{}}_1,\ldots,T^{\acute{}}_B$ from $Q_{\widetilde{P}}$ .

The first error is unavoidable. Fitting a wrong parametric distribution $P_\theta$ will make this error big. It is usually safer to use the empirical estimate for $P$ .

The second error depends on $B$ . The larger $B$ , the smaller this error. Usually $B=1000$ is sufficient.

<u>Example sun hours (1)</u>

Example Assume we want to estimate the *var* of the sample mean of the sun hours $var\left(\acute{X}_n\right)$ .

### Option 1

**Assuming** $X_1,\ldots,X_{59}\sim N(\mu,\sigma^2)$ we've estimated $P$ by $P_{\hat{\theta}}=N(\hat{\mu},\hat{\sigma}^2)=N(1544,23699)$ ¿). We know from theory that the distribution of $\acute{X}_n$ is $Q_P=N(\mu,\sigma^2/n)$ so we don't need to take bootstrap samples $X^{¿}$ 's and compute $T^{¿}$ 's here!! ☺

After having done the first step, $\widetilde{P}=N(\hat{\mu},\hat{\sigma}^2)$ we can immediately estimate $var(\acute{X}_.)$ . by $\hat{\sigma}^2/n=23699/59=402$ .

### Option 2

**Assuming** $X_1,\ldots,X_{59}\sim N(\mu,\sigma^2)$ we've estimated $P$ by $P_{\hat{\theta}}=N(\hat{\mu},\hat{\sigma}^2)=N(1544,23699)$ ¿). We can still compute the bootstrap samples. Using $B=1000$ this yields $399$ .

This option is really worse than option 1 because we estimate the variance of $Q_{P_{\hat{\theta}}}$ using a sample, while we know the parameters of this normal distribution.

### Option 3

We don't assume normality and use the empirical bootstrap scheme

*> var(bootstrap(sun,mean,1000))*
*[1] 380.055*
*> var(bootstrap(sun,mean,1000))*
*[1] 425.2614*
*> var(bootstrap(sun,mean,10000))*
*[1] 402.4628*
*> var(bootstrap(sun,mean,10000))*
*[1] 390.7697*

Part of the variation is due to the first bootstrap error.

Summarizing the 3 options:

1. Parametric "bootstrap" with normal theory (without $T^{¿}$ 's)
2. Parametric bootstrap with bootstrap sampling (with $T^{¿}$ 's)
3. Empirical bootstrap sampling

The first option is only possible in special cases. For example, in the example of the *sd* of the sample median, sampling of $T^{¿}$ 's was really necessary in a parametric bootstrap set up, since the distribution of median( $X$ ) is not easily derived from the distribution of the $X_i$ 's.

Is option 1 or 3 the best? In this case there isn't much difference. Normality is a little doubtful (SW-test: p=0.08), so the empirical bootstrap estimator would slightly be preferred.

## 4.2 Bootstrap confidence intervals

Set-up: estimate an unknown parameter $\theta$ by the estimator $T$ (with unknown distribution $Q_P$ ). The accuracy of T can be expressed in terms of

- $var(T)$ or $sd(T)$ (we've seen bootstrap estimators for this)
- confidence interval for $\theta$, based on $T$.

A confidence interval is a correct manner to present the accuracy of the estimator T. The interval is based on the distribution $Q_P$ of $T$. Since we don't know $Q_P$, we can use the bootstrap estimate (the empirical distribution of a sample from $Q_{\tilde{P}}$) as an approximation.

The confidence interval, before bootstrapping

$T$ estimates $\theta$, so we hope that the distribution of $T-\theta$ is concentrated around 0. Denote the distribution of $T-\theta$ by $G$.

By the definition of quantiles:

$$P\left(G^{-1}(\alpha)\le T-\theta\le G^{-1}(1-\alpha)\right)$$
$$¿\, P\left(T-\theta\le G^{-1}(1-\alpha)\right)-P\left(T-\theta\le G^{-1}(\alpha)\right)$$
$$\ge 1-\alpha-\alpha=1-2\alpha$$

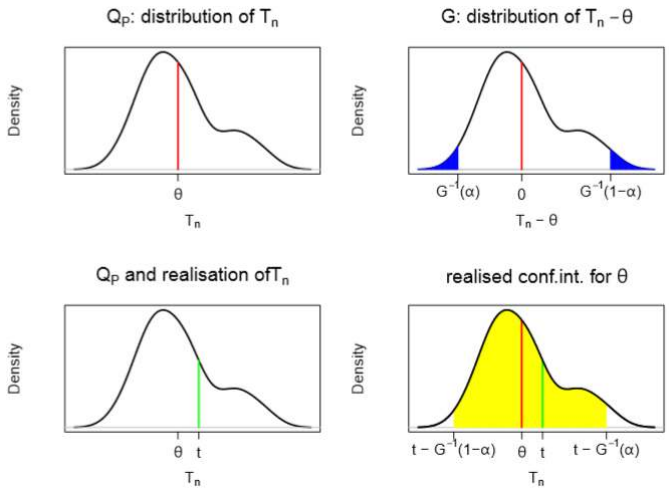
This can be written as:

$$P\left(T-G^{-1}(1-\alpha)\le\theta\le T-G^{-1}(\alpha)\right)\ge 1-2\alpha.$$

Hence, a $(1-2\alpha)$ confidence interval for $\theta$ is:

$$\left[T-G^{-1}(1-\alpha),T-G^{-1}(\alpha)\right].$$

In pictures:




The bootstrap confidence interval

-

In confidence interval $\left(1-\alpha\right),T-G^{-1}¿$ $T-G^{-1}¿$ $¿$ unknown are:

- $G$, i.e. the distribution of $T-\theta$,
- $Q_P$, i.e. the distribution of $T$,
- $\theta$, the parameter of interest.

Hence, estimate

the distribution $G$ of $Z=T-\theta$
by the empirical distribution of $Z_i^{\iota}=T_i^{\iota}-T$ , $i=1,\ldots,B$,

with $T_1^{\iota},\ldots,T_n^{\iota}$ a bootstrap sample (empirical or parametric).

$G^{-1}(\alpha)$ is estimated by $Z^{\iota}_{([\alpha B])}$ .
$G^{-1}(1-\alpha)$ is estimated by $Z^{\iota}_{([(1-\alpha)B])}$ ..

Hence, the unknown interval $[T-G^{-1}(1-\alpha),T-G^{-1}(\alpha)]$ is estimated by the computable interval

$$[T-Z^{\iota}_{([(1-\alpha)B])},T-Z^{\iota}_{([\alpha B])}]$$

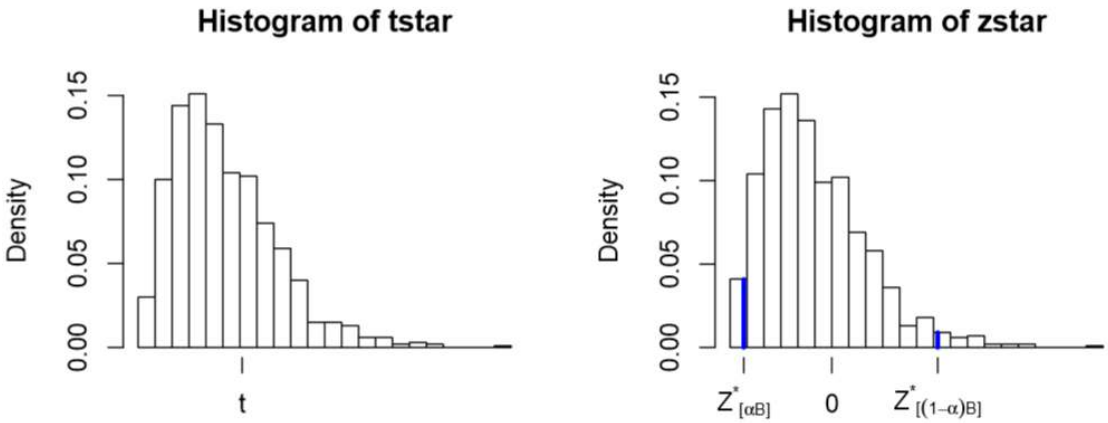
which is equal to

$$[2T-T^{\iota}_{([(1-\alpha)B])},2T-T^{\iota}_{([\alpha B])}]$$

Because $Z_i^{\iota}=T_i^{\iota}-T$ .

*R: quantile*

In pictures:




```
> zstar=tstar-tn
> c(tn-quantile(zstar,0.975),tn-quantile(zstar,0.025))
      97.5%       2.5%
-0.1129308 11.3179222
> 2*tn-quantile(tstar,c(0.975,0.025))
      97.5%       2.5%
-0.1129308 11.3179222
```

Reliability of a confidence interval

Problem Several possibilities to construct confidence intervals (e.g. empirical or parametric bootstrap). Which one works best? That really depends on the situation (statistic $T$ , distribution $P$ of $X_i$ , ...).

Approach Simulate actual coverage probability of nominal $(1-\alpha)$ confidence intervals! Repeat the following procedure many (e.g. $K=1000$ ) times:

1. generate a random sample $x_1,\dots,x_n$ according to $P_\theta$ ,
2. derive statistic $T_n$ , generate bootstrap samples $T^{*}_{1.},\dots,T^{*}_{B}$ ,
3. construct confidence interval as explained before,
4. is $\theta$ part of the interval? If yes, output: 1, else: 0.

Number of 1's divided by $K\approx$ coverage probability.

### 4.3 Bootstrap tests

Situation Consider a sample $X_1,\dots,X_n$ from an unknown distribution $P$ . Suppose we want to test the goodness-of-fit hypothesis

$$H_0: P\in P_0 \qquad \text{versus} \qquad H_1: P\notin P_0$$

for some collection of distributions $P_0$ . Use a test statistic $T$ .

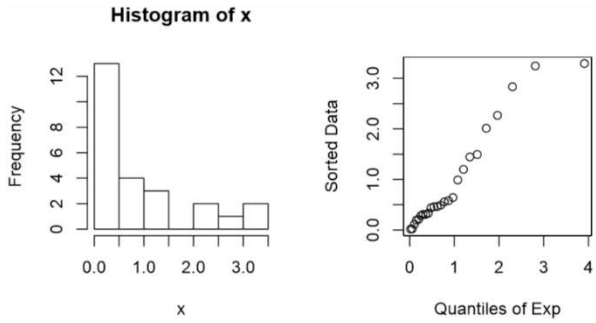
Problem: we don't know the distribution $Q_P$ of $T$ for $P\in P_0$ .

We can use a bootstrap test.

Idea Estimate the unknown distribution of $T$ (that is $Q_P$ ) by the empirical distribution of a sample $T^{*}_{1.},\dots,T^{*}_{B}$ .

Example(1)

Consider the data (histogram and QQ-plot against $\exp(1)$ ).




and the hypotheses

$H_0: X_1,\dots,X_n \sim \exp(\lambda)$ for some λ > 0

$H_1: X_1,\dots,X_n$ are not *exp* distributed

which we want to test using test statistic

$$T=\frac{median(X)}{mean(X)}$$

We simulate $T$ under $H_0$ , because we don't know the distribution of $T$ under $H_0$ .

The distribution $Q_P$ of $T$ under $H_0$ does not depend on $\lambda$ . Thus, $Q_P$ is the same regardless which exponential distribution the $X_i$ come from. Therefore, $T$ is called nonparametric.

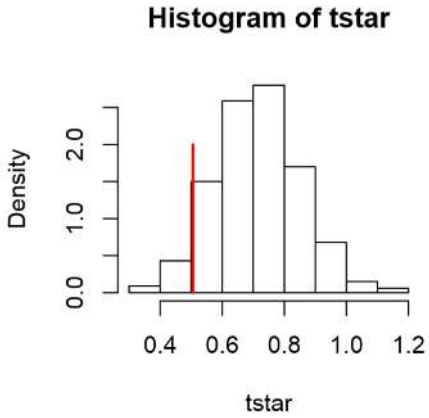
We simulate the distribution $Q_P$ of $T$ using the (parametric) bootstrap scheme:

- Generate $B$ times a sample $X^{*}=(X^{*}_1,\dots,X^{*}_n)$ from $\exp(1)$

- Compute for each sample the value $T^{i} = \frac{median(X^{i})}{mean(X^{i})}$

Remark The label bootstrap is actually inappropriate for **this** test, because we do not use the data in the generation of the $T^{i}$ 's (cf. empirical and parametric bootstrap estimation).

```
> for(i in 1:B) {
+ xstar=rexp(n)
+ tstar[i]=median(xstar)/mean(xstar) }
> median(x)/mean(x)
[1] 0.5058572
> p=2*min(sum(tstar<=0.5058572)/B,sum(tstar>=0.5058572)/B)
> p
[1] 0.112
```




$H_0$ is not rejected, since the (two-sided) p-value is 0.112.

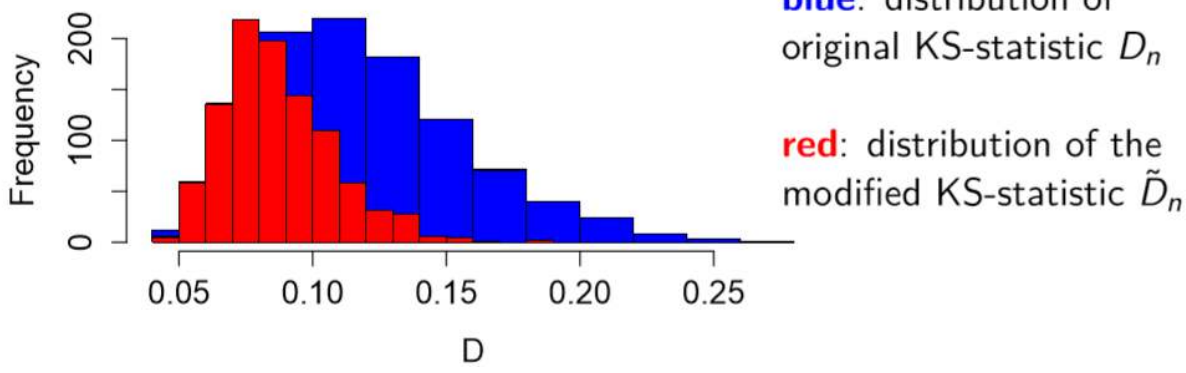
Example(2)

Remember how not to use the Kolmogorov Smirnov test for the composite null hypothesis

$$H_0: X_1, \ldots, X_n \quad N(\mu, \sigma^2) \quad \text{for some} \quad \mu \quad \text{and} \quad \sigma^2$$

```
> ks.test(x,pnorm,mean(x),sd(x))
```

This R-command tests the simple $H_0: X_1, \ldots, X_n \quad N(\acute{X}, S_X^2)$ .




The adjusted KS-test statistic $\widetilde{D}_n$ is informative, and therefore a sensible test statistic. However, the reported p-value of

```
> ks.test(x,pnorm,mean(x),sd(x))
```

is wrong, since it is based on the blue distribution of $D_n$ , whereas it should be based on the red distribution of $\widetilde{D}_n$ . Using bootstrap testing, one can simulate the red distribution of $\widetilde{D}_n$ .

Bootstrap: Warnings

Warning The bootstrap does not always “work” (well)!

- Typical risk: if you use a parametric bootstrap but sample contains outliers and parameter estimator is sensitive ⇒ possibly bad bootstrap estimates (see exercise 4.1).
- Extreme order statistics: the empirical bootstrap usually fails to approximate the distribution of $X_{(1)}=min\left(X_1,\ldots,X_n\right)$ or $X_{(n)}=max\left(X_1,\ldots,X_n\right)$ .
- If the distribution underlying your sample has heavy tails (e.g. non-existent first moment), the empirical bootstrap fails. (It always produces no tails, i.e. light tails.) (See Example 4.6 in syllabus with the Cauchy distribution).

# Chapter 5
# Robust estimators

Idea Most methods from statistics are based on assumptions on the probability distribution from which the observations originate. Such assumptions are almost never completely correct. It is therefore important to investigate how sensitive a method is for deviations from the assumptions. Statistical methods which are relatively insensitive to small deviations from the model assumptions are called **robust methods**.

Robust methods are insensitive to small deviations from the assumptions, e.g.

- large deviations from the assumptions in some observations
- small deviations from the assumptions in all observations
- small deviations from the assumed (in)dependence structure

An outlier is a data point that deviates from the other observations, e.g. extremely large or extremely small values. In some cases outliers are 'wrong' data, from which mistakes are easily made. A number of 'incorrect' observations up to 10% is not abnormal.

Sometimes it is possible to identify, or even correct, mistakes by careful screening of the data. In general, it is recommended to perform a statistical data analysis both with and without suspect data points and compare the results. Obviously, it is wrong to delete observations from a data set just for subjective reasons. A good robust procedure decreases the influence of incorrect observations in an objective manner.

## 5.1 Robust estimators for location

### 5.1.1 Trimmed means

So far we have seen the mean and the median as location estimators.

The mean is very sensitive to outliers, i.e. is not robust.
The median is very insensitive to outliers, i.e. is very robust.

Both are examples of $\alpha$ -trimmed means:

$$T_{n,\alpha}=\frac{1}{n-2\lfloor \alpha n \rfloor}\sum_{j=\lfloor n\alpha \rfloor+1}^{n-\lfloor \alpha n \rfloor} X_{(j)}, 0\leq \alpha < \frac{1}{2}$$

In words: the $\lfloor \alpha n \rfloor$ largest and smallest observations are deleted and the average is taken over the remaining values. It is clear that outliers have a smaller influence on $T_{n,\alpha}$ then on $\acute{X}$ , because they are 'trimmed away' when $T_{n,\alpha}$ is used (if $0\leq \alpha < \frac{1}{2}$ ).

$\acute{X}$ is the 0-trimmed mean, whereas the other extreme, the (almost) $\frac{1}{2}$ -trimmed mean is close to the median. The median is very robust against the occurrence of outliers. Even when almost all observations are extreme, the median is a useful location estimator.

The concept of location

What do these location estimators estimate?

For independent random variables $X_1,\ldots,X_n$ from a distribution $F$ with density $f$, the usual mean $\acute{X}_n$ estimates the population mean

$$E\,X=\int x\,dF(x)$$

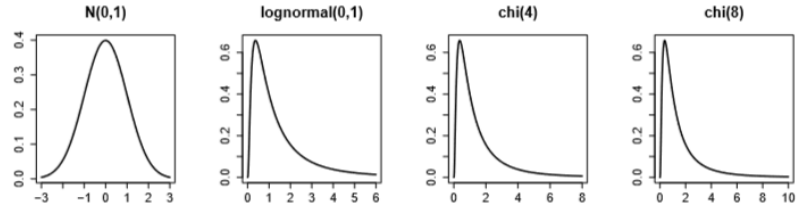
and the trimmed mean $T_{n,\alpha}$ estimates the trimmed population mean

$$T_\alpha(F)=\frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x\,dF(x).$$

where $dF(x)$ denotes integration with respect to the (continuous/discrete) distribution measure $F$.

Example: Some trimmed means of distributions

| $\alpha$ | $N(0,1)$ | lognorm(0,1) | $\chi^2_4$ | $\chi^2_8$ |
|---|---|---|---|---|
| 0 | 0 | 1.65 | 4 | 8 |
| 0.1 | 0 | 1.24 | 3.64 | 7.63 |
| 0.2 | 0 | 1.11 | 3.50 | 7.49 |
| 0.3 | 0 | 1.04 | 3.41 | 7.40 |
| 0.4 | 0 | 1.01 | 3.37 | 7.36 |
| 0.5 | 0 | 1 | 3.36 | 7.34 |




Example: location estimators

Location estimators for Newcomb's data: original set, and altered set (add an extra outlier).

```
> mean(newcomb)
 [1] 26.21212
 > mean(newcomb,trim=0.1)
[1] 27.42593
> mean(newcomb,trim=0.2)
[1] 27.35
> mean(newcomb,trim=0.3)
[1] 27.25
> mean(newcomb,trim=0.4)
[1] 27.28571
```

```
> nnew=c(newcomb,-60)
> mean(nnew)
[1] 24.92537
> mean(nnew,trim=0.1)
[1] 27.29091
> mean(nnew,trim=0.2)
[1] 27.2439
> mean(nnew,trim=0.3)
[1] 27.14815
> mean(nnew,trim=0.4)
[1] 27.2
```

> median(newcomb)
[1] 27

> median(nnew)
[1] 27

The mean is very sensitive to outliers, trimmed means much less. This influence of outliers can be quantified in the influence function.

Asymptotic influence function

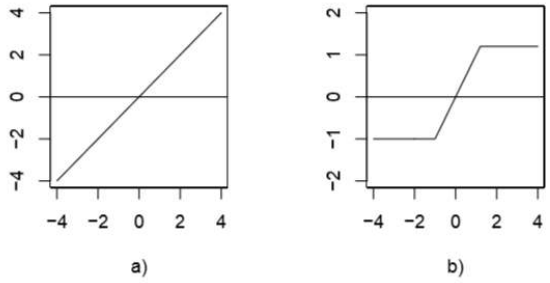
The sensitivity of the mean for outliers can be quantified by an influence function (IF). Suppose that we have a random sample of $n$ observations $x_1,\ldots,x_n$ and that we next obtain one additional observation of size $y$. When we calculate the mean for both cases, the difference of these means, 'the influence of an additional observation in y', is:

$$\sum_{i=1}^{n} \frac{x_i+y}{n+1} - \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$IF(y) = \frac{1}{n+1}$$

Multiply by $n+1$ and take $n \to \infty$ to obtain $IF(y) = y - EX$.
The function $y \mapsto IF(y) = y - EX$ is the (asymptotic) influence function of the mean.

The influence function for the mean (left) and the α-trimmed mean (right).




For the mean the IF is not bounded, whereas for the median the IF is bounded.

Boundedness of the influence function is an important prerequisite for an estimator to be robust. Estimators with an unbounded IF, like the mean $\acute{X}$, are not robust. Estimators with a bounded IF are called B-robust; the median is B-robust. (Note, however, that the influence of an extreme observation is not zero!) **All $\alpha$ -trimmed means with $\alpha > 0$ are B-robust!**

The gross error sensitivity is equal to $\sup_y |IF(y)|$. The smaller the gross error sensitivity, the more B-robust the estimator.

### 5.1.2. $M$ -estimators

There are many other robust measures for location. An important class of such measures is formed by the so-called $M$ **-estimators**. An $M$ -estimator can be regarded as a

generalization of a maximum likelihood estimator. For a given function $\rho$, the $M$-estimator $M_n$ can be defined as the value of $M_n$ that maximizes the expression

$$\prod_{i=1}^{n} \rho(X_i - M_n)$$

Most of the time an $M$-estimator $M_n$ is not computed as the solution of a maximization problem, but as the solution of an equation of the form

$$\psi(¿ X_i - M_n) = 0$$

$$\sum_{i=1}^{n} ¿$$

for some function $\psi$.

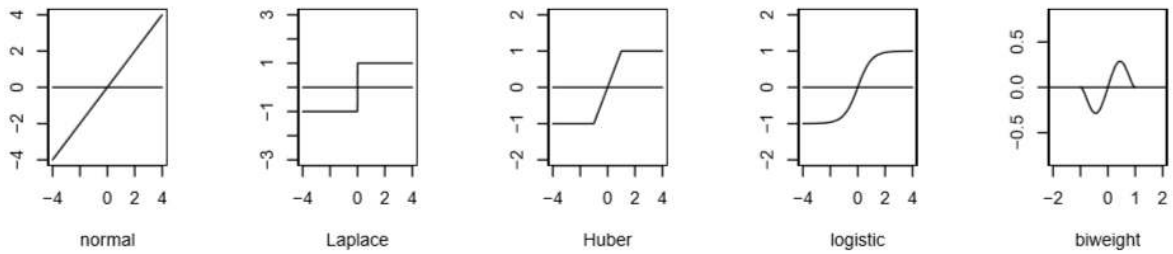
When $\psi(x) = x$, then $M_n = \acute{X}_n$, ML estimator for normal distribution.
When $\psi(x) = sign(x)$, then $M_n = median(X)$, ML estimator for Laplace distribution.

The influence function of an $M$-estimator is equal to:

$$\int \psi'(x - ¿ T_\psi(F))\, dF(x)$$

$$IF(y, F) = \frac{\psi(y - T_\psi(F))}{¿}$$

and has the same shape as the function $\psi$. Hence, to obtain a B-robust $M$-estimator it is sufficient to choose a bounded function $\psi$. In other words, bounded $\psi$-functions yield B-robust location estimators.




IFs for some $\psi$-functions (that yield ML estimator for denoted distribution).