Exploratory Data Analysis

Johannes Schauer
johannes.schauer@TUGraz.at

Institute of Statistics
Graz University of Technology
Steyrergasse 17/IV, 8010 Graz
www.statistics.tugraz.at

February 12, 2008

Introduction

Points of action:

- Presence of uncertainty
- Variation in data
- Very little / too much information
- . . .

Statistical methods $\Longrightarrow$ judgements and decisions

Populations and samples

We focus on a well-defined collection of objects, the **population**.

Examples:

- All cars produced on February 8, 2008 in factory A.
- All children in the European Union aged 8-12.

Usually - due to constraints on time, money, etc. - we only select a subset of the population, a **sample**.
For example we might select 10 cars out of the above car population to check the exhaust emissions.

Variables

We are interested in certain characteristics of the elements of the sample, e.g.:

- gender,
- age,
- weight or
- hair color.

A **variable** is any characteristic that may change from one object to another. There are **two types** of variables:

- categorical (male/female, blonde/brown/other) and
- numerical (age=9 years, weight=35.2 kg).

What can we do with a sample to begin with?

Exploratory Statistics:

- Graphical analysis
- Numerical summary measures

With the help of statistics software packages like `R` such calculations can easily be done.

Motivating example

Suppose we want to study the **age of our neighbors**. We get the following data:

26	34	35	13	4	20	74
50	14	48	14	53	9	39
36	40	41	56	16	41	17
46	43	18	35	38	35	45

Looking at the raw data it is difficult to see any distinctive features in it. Let us apply a simple graphical feature.

Stem-and-Leaf Display

```
> ages<-c(26,34,35,13,4,20,74,50,14,48,14,53,9,39,36,
+   40,41,56,16,41,17,46,43,18,35,38,35,45)
> stem(ages)

  The decimal point is 1 digit(s) to the right of the |

  0 | 49
  1 | 344678
  2 | 06
  3 | 4555689
  4 | 0113568
  5 | 036
  6 |
  7 | 4
```

Stem-and-Leaf Display - 2

- easy way to organize numerical data
- split observations in two parts
    1. stem (one or more leading digits)
    2. leaf (remaining digits)
- no loss of information
- very useful for samples with 20-200 observations

Stem-and-Leaf Display - 3

```
0 | 49
1 | 344678
2 | 06
3 | 4555689
4 | 0113568
5 | 036
6 |
7 | 4
```

What can we see now?

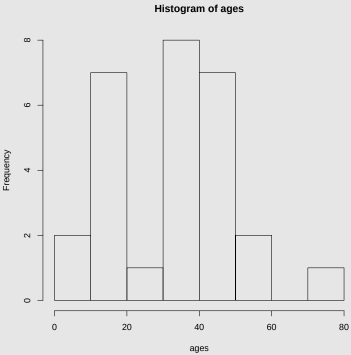
- Our youngest neighbor is 4, our oldest is 74.
- The groups of people in their 30s or 40s are the largest ones.
- Nobody is in his 60s.
- There are three neighbors aged 35.

Stem-and-Leaf Display - 4

Information in such a display:

- typical value
- extent of spread around the typical value
- presence of gaps
- symmetry
- number and location of peaks
- outliers

Histogram



- easy plot to see characteristics of the data
- more efficient than a stem-and-leaf display with larger data sets

Histogram - 2

Some numerical data is obtained by **counting** to get the value of a variable, e.g.

- the number of car accidents in one week or
- the number of customers in a shop on one day.

For other numerical data the **number of possible values** might be **finite**, for example University grades 1-5 (in Austria).

Variables as the ones above together with categorical variables are called **discrete**.

The set of possible values of discrete variables is either finite or can be listed as an infinite sequence $(1, 2, 3, \ldots)$.

Histogram - 3

Other numerical data is **obtained by measurements**, e.g.

- the weight of an individual,
- the computation time of a program or
- the difference between the actual fill quantity of a beer bottle and its target fill quantity.

These variables are called **continuous**.

The possible values of continuous variables consist of an entire interval on the real line.

Histogram - discrete variable

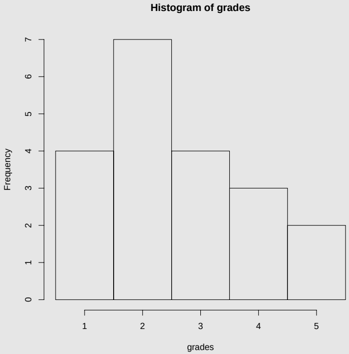
We look at a small sample consisting of the grades of 20 students in the statistics course.

4	2	2	1	3	2	5	4	2	1
3	1	4	5	2	1	2	3	3	2

To draw a histogram, we **count** how many times each grade has occurred and draw a **bar** with **height=(number of occurrences)** over that grade.

```
> grades<-c(4,2,2,1,3,2,5,4,2,1,
+   3,1,4,5,2,1,2,3,3,2)
> hist(grades,breaks=seq(0.5,5.5,1))
```

Histogram - discrete variable - 2



- The most common grade, the **mode**, is a 2.
- There are more bars on the right side of the mode than on the left side.
- Therefore there is no obvious symmetry around 2.

Histogram - discrete variable - 3

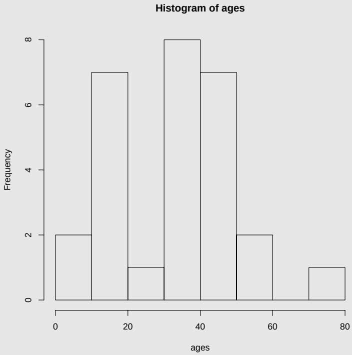
Take the ages sample from before. We **divide** the ages into 8 **groups** of equal length: 1-10, 11-20, . . ., 71-80. That is, we count how many observations lie in each of the groups.

1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
2	7	1	8	7	2	0	1

Now we can act as in the previous example.

```
> hist(ages)
```

Histogram - discrete variable - 4

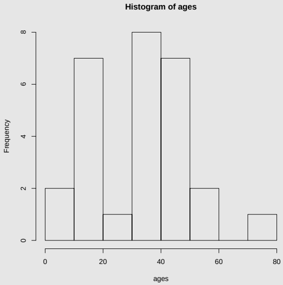


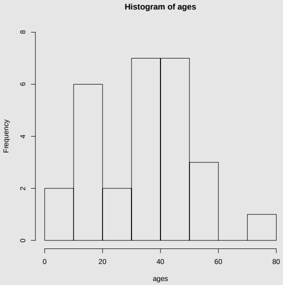
- similar to stem-and-leaf display
- common ages are 11-20 and 31-50
- two peaks
- gap in 61-70

Histogram - discrete variable - 5

- The **width** and the **positions** of the bars have to be chosen.
- They influence the shape of the histogram.

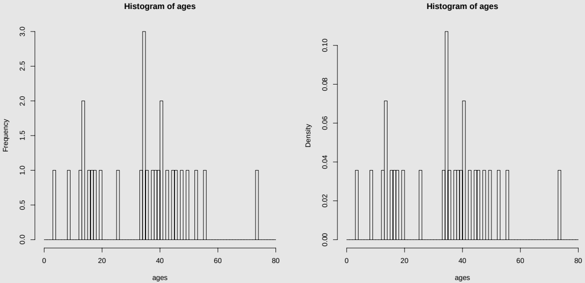
```
> hist(ages)
> hist(ages,right=FALSE)
```





Histogram - discrete variable - 6

If we put the **width** down to **one year**, there is a bar for each possible age.



Instead of the actual frequency, the height of the bars can also be the **relative frequency**:

$$\text{relative frequency} = \frac{\text{number of times the value occurs}}{\text{number of observations in the sample}}.$$

Histogram - discrete variable - 7

The **last histogram** with relative frequencies has the property, that the **total area** of all bars is **one**.

The **area of the bar** on top of 35 then **estimates** the **probability** that a neighbor is 35 years old.

In the discrete case we assume that there is a true probability of appearance for each age. They can also be plotted similar to a histogram.

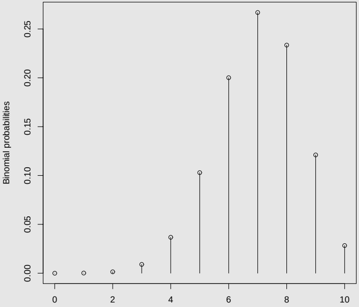
The **distribution** of the variable is **responsible for the shape** of that plot.

The **sum over all probabilities** is always **one**.

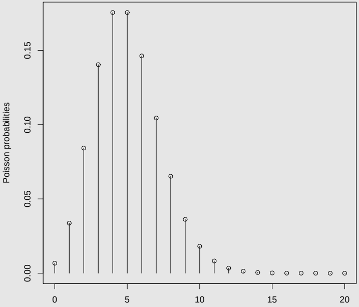
Discrete distributions

Examples of distributions with their probability mass functions:

Binomial distribution



Poisson distribution



There is **not always such a characteristic** in the distribution!

Discrete distributions - 2

The Binomial Distribution

Example: The number of students out of 10 that pass the statistics test (frequency). Usually 70 % pass the test.

- Items in **groups** of size $n = 10$.
- An item can be satisfactory (**success**) or unsatisfactory (**failure**).
- **Proportion** of all items that are **successes** is $\pi = 0.7$.
- **Proportion** of **failure** is $1 - \pi = 0.3$.

Therefore one observation can take values in $\{0, 1, 2, \ldots, n\}$, that is there is **a minimum and a maximum value**.

The average value, the **mean**, of such an observation is $n * \pi = 10 * 0.7 = 7$.
The **variance** is $n * \pi * (1 - \pi) = 10 * 0.7 * 0.3 = 2.1$.

Discrete distributions - 3

The Poisson Distribution

Examples:

1. The number of car accidents in Graz on one day.
2. The number of customers in a specific supermarket at 1 PM.

- There is **no explicit maximum value**.
- Usually used for **count data**.
- There is **only one parameter** λ that defines the shape of the distribution (mean=variance=λ).

Discrete distributions - 4

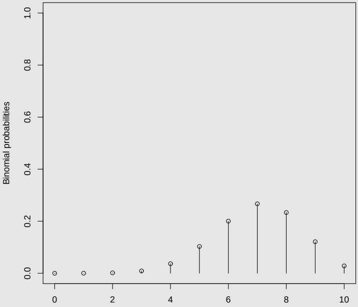
The probability mass function

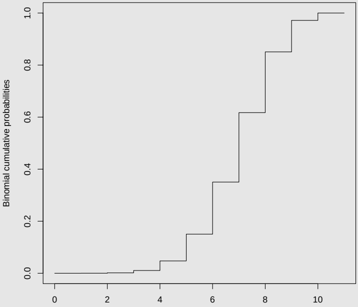
- $p_k = P(X = k)$ is the probability that the variable has value k.
- $p_k \geq 0$
- The sum over all p_k is one.

Discrete distributions - 5

If we **sum up** the **probabilities** to each value of a discrete variable and plot the resulting **step function**, we get the **distribution function** $F(x)$ of that random variable.

P(2 successes) $\Longrightarrow$ P(up to 2 successes)


Probability mass function

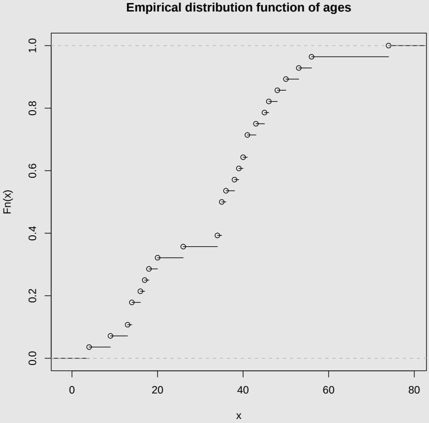

Distribution function $F(x)$

Discrete distributions - 6

If we already have a sample, we can calculate the **empirical distribution function**. This is a step function that makes a step of equal height $1/(\#$ of obs.$)$ at each observation in the sample.

Ages example:

```
> plot(ecdf(ages))
```



Some important measures of X

Mean: $E(X)$, average value, center of mass.

Variance: $\mathrm{Var}(X)$, measure for the variation in observed values around the mean.

Standard deviation: $\sqrt{\mathrm{Var}(X)}$, the square root of the variance.

Relation between the population and the sample:

$$
\begin{array}{ccc}
\textbf{Population} & \leftrightarrow & \text{Mathematical model assumptions} \\
\Downarrow & & \Uparrow \\
\textbf{Sample} & \leftrightarrow & \text{Check of adequacy of assumptions}
\end{array}
$$

Histogram - continuous variable

A **power company** needs information about the **power demand**. Therefore they study power consumption values of **90 gas-heated homes** in the US. The first 20 values in ascending order are the following.

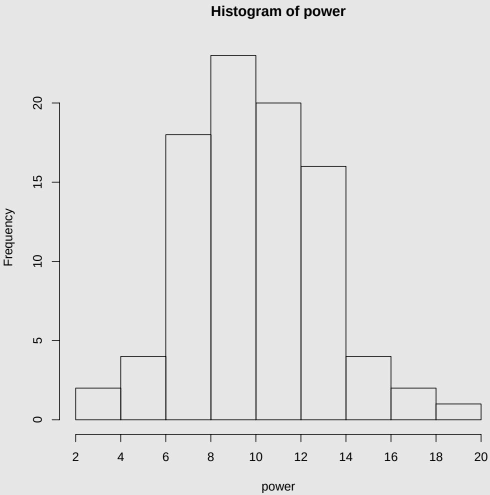
2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
...	...	...	...	...	...	...	...	...	...

Again we **divide** the values in **intervals of equal length**, this time e.g.

$$(2,4],(4,6],\ldots,(18,20],$$

and **count** the number of **observations** in each interval. Drawing bars as high as the number of occurrences gives us:

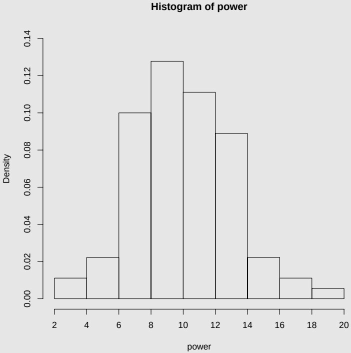
Histogram - continuous variable - 2



The histogram would look different for another choice of intervals!

Histogram - continuous variable - 3

If we **divide** the **height** of all **bars** by the **total area** of all bars, we obtain a histogram of relative frequencies:



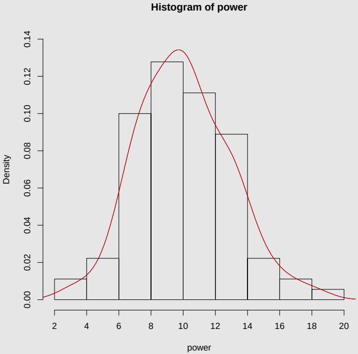
In this plot the **total area** of all bars is **one**. The bars can now be interpreted as **estimators for probabilities**:

The probability of a power consumption between 6 and 8 should be close to 10 %.

Histogram - continuous variable - 4

If we **smooth** such a **histogram** of relative frequencies we obtain a so called **density estimator** (red line):

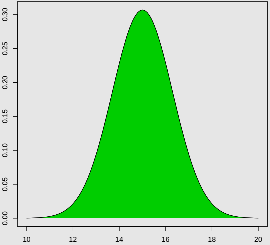
```
> lines(density(power))
```



For explaining the properties of a continuous variable a density estimator is **more appropriate** (because there are no jumps). This leads us to describing the distribution of continuous (random) variables.

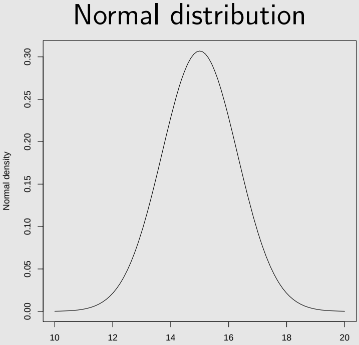
Continuous distributions

- In the **continuous** case variables **can take all values** in some **interval**, e.g. $[10, 20]$.
- The **probability** that the variable takes any **fixed value**, e.g. 15, is **0**.
- Therefore we need to find another concept of describing the probability mass: the **density function**.
- The **area between** the **density function** and the **x-axis** is **one**.

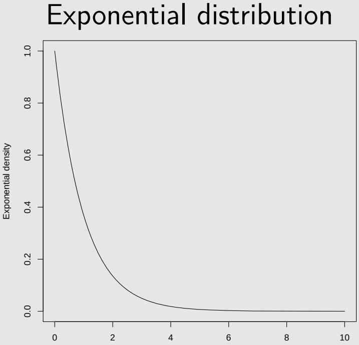


Continuous distributions - 2

Examples of distributions with their density functions:



Normal distribution



Exponential distribution

Continuous distributions - 3

The Normal Distribution

Examples:

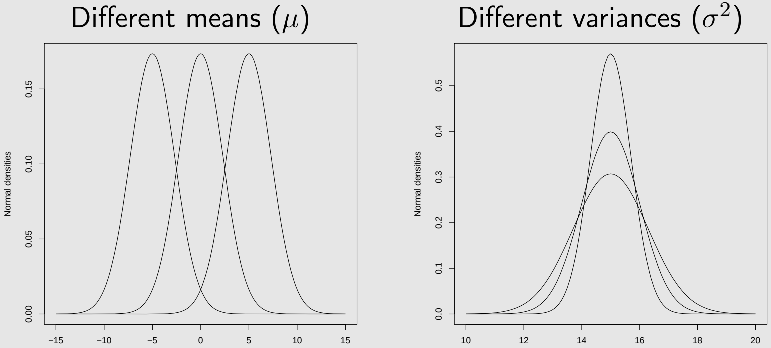
1. weights or heights of humans
2. measurement errors

Properties:

- most important distribution in statistics
- characterized by two parameters:
    1. the mean μ
    2. the variance σ^2
- sums or averages of variables are often approximately normally distributed

Continuous distributions - 4

Examples of normal density functions



Different means (μ)

Different variances (σ^2)

Continuous distributions - 5

The Exponential Distribution

Examples:

1. life time of a bulb
2. time between successive arrivals in a shop

Properties:

- can take all values larger than 0
- one parameter λ
- mean=λ
- variance=λ^2

Continuous distributions - 6

The density function

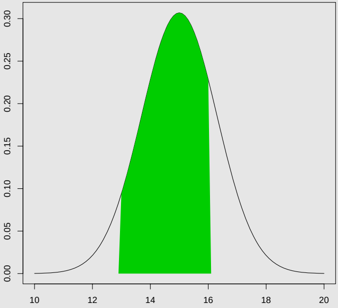
- usually called $f(x)$
- $f(x)$ is defined for all values x on the real line.
- $f(x) \geq 0$
- The area between $f(x)$ and the x-axis is one.

Continuous distributions - 7

To get the **probability** that a random variable X lies **between** a **and** b, we need to look at the **area** between the density function $f(x)$ and the x-axis from a to b.

For example the probability of a normal random variable X with $\mu = 15$ and $\sigma = 1.3$ to be between 13 and 16 is the green area:

$$P(13 \leq X \leq 16) = 0.72$$

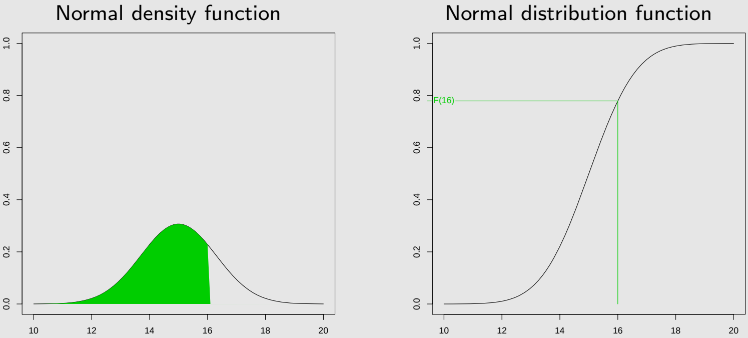


Continuous distributions - 8

The distribution function

If we study the probability that a random variable X is at most b, we are looking at its distribution function $F(b) = P(X \leq b)$.

For example we look at a normal random variable X as before ($\mu = 15$, $\sigma = 1.3$). Graphically looking at the green area is the same as $P(X \leq 16) = 0.78$.



Normal density function

Normal distribution function

Continuous distributions - 9

Properties of distribution functions:

- $F(x)$ is defined for all values x on the real line.
- $F(x)$ is a **nondecreasing** function.
- For x sufficiently **small** $F(x)$ is arbitrarily **close to 0**.
- For x sufficiently **large** $F(x)$ is arbitrarily **close to 1.**

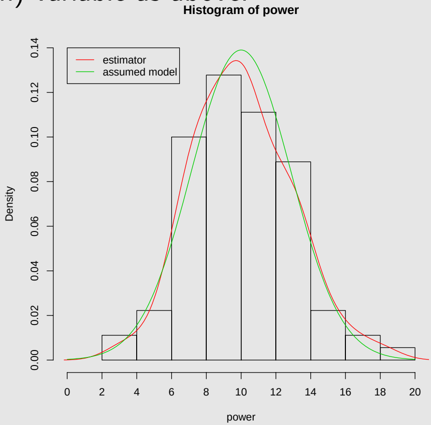
These properties apply for both continuous and discrete random variables' distribution functions.

The **empirical distribution function**, as it was introduced earlier, can also be calculated **in the same way** for a sample from a **continuous** population.

Histogram - application

Let's go back to the power demand example and **assume** that the **power demand** is **normally distributed** with $\mu = 10$ and $\sigma = 2.87$.

Compare the histogram of our sample with the density function of a normal (random) variable as above:



Histogram - shapes

Unimodal: One single peak.

Bimodal: Two different peaks.

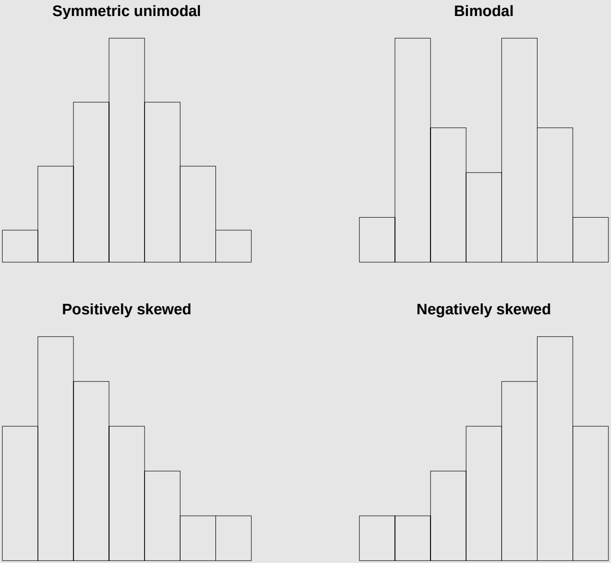
Symmetric: The left half is a mirror image of the right half.

Positively skewed: The right tail is stretched out further than the left one.

Negatively skewed: The left tail is stretched out further than the right one.

These shapes are also relevant for probability mass functions and density functions.

Histogram - shapes - 2



Symmetric unimodal

Bimodal

Positively skewed

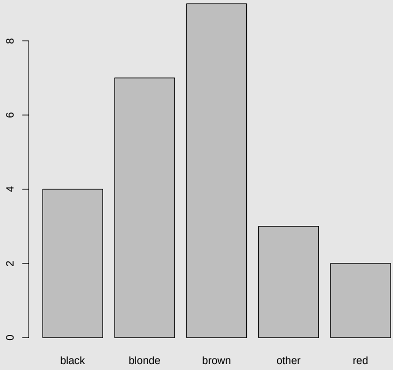
Negatively skewed

Barchart

- **Histogram** for **categorical data**.
- Some arbitrary order.

Example: Hair color of 25 children

```
> barplot(table(hair))
```

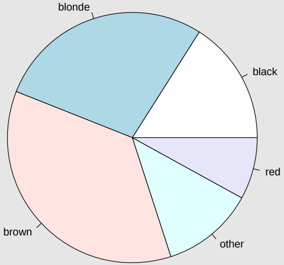


Piechart

- used for categorical data
- shows similar information as a barchart

Example: Hair color example from before

```
> pie(table(hair))
```



Numerical summary measures

Measures of the center

1. The **sample mean** $\bar{x}$.
2. The **sample median** $\tilde{x}$ (also called $q_{0.50}$).

We suppose we have a sample $x = (x_1, \dots, x_n)$, thus n is the **number of observations**.

The sample mean

The sample mean is defined by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

The sample mean **estimates** the **true mean** of the population. It is the **most common estimator**.

Problem: A **single outlier** (unusually large or small observation) **can** significantly **change** the sample **mean**!

The sample mean - Example

Remember the ages data set:

26	34	35	13	4	20	74
50	14	48	14	53	9	39
36	40	41	56	16	41	17
46	43	18	35	38	35	45

The sample mean is then

$$\bar{x} = \frac{26 + 34 + 35 + \ldots + 38 + 35 + 45}{28} = \frac{940}{28} = 33.57143.$$

```
> mean(ages)
[1] 33.57143
```

The sample median

- **Alternative measure** of the center.
- **Resists** the effects of **outliers**.
- **But:** the median **only estimates** the **true mean** for **symmetric distributions**!

The sample median **divides** the **observations** in **two equally big parts**:
Half of the observations are **smaller** than the sample median, **half** of them are **larger**.

The "**middle value**" of the ordered sample is the sample median:

1.2 3.4 4.2 5.1 5.9 6.9 8.3

The sample median - calculation

The following formula shows how to calculate the sample median $\tilde{x}$ from the ordered sample:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even.} \end{cases}$$

This mean, if n is **odd**, we simply take the **middle value**.

If n is **even**, there are **two middle values**, so we take the **average** of those two.

The sample median - examples

First look at the (already ordered) sample from before:

1.2 3.4 4.2 5.1 5.9 6.9 8.3

Here n is 7 (odd) and therefore $x_{(4)} = 5.1$ is the middle value and the sample median.

Suppose the sample consists of one more element, e.g. 8.7:

1.2 3.4 4.2 5.1 5.9 6.9 8.3 8.7

Then $n = 8$ (even) and the sample median is $\frac{1}{2}(x_{(4)} + x_{(5)}) = 5.5$.

The sample median - ages example

In the ages example we get the following sample median using R:

```
> median(ages)
[1] 35.5
```

Here we have $\tilde{x} = \frac{1}{2}(x_{(14)} + x_{(15)}) = \frac{1}{2}(35 + 36)$.

Median of a distribution

As we have just seen, the **sample median** has the property, that **50 %** of all observations are **smaller** or equal its value, while **50 %** are **larger**.

Looking at the mathematical model, the **median** $x_{0.50}$ of a continuous **random variable** X has the same property:

$$P(X \leq x_{0.50}) = P(X > x_{0.50}) = 0.50.$$

The probability of X being smaller than the median is 50 %.

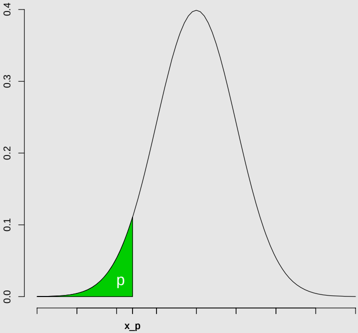
For discrete random variables the definition is similar.

Theoretical quantiles

As we just defined $x_{0.50}$, we can introduce x_p for any $0 < p < 1$ with the property

$$P(X \leq x_p) = p.$$

That is, the probability of X being smaller than x_p is $100 * p$ %.



Sample quantiles

To **estimate** the **theoretical quantiles** of a random variable from a sample $(x_1, \ldots, x_n)$ one usually uses for $0 < p < 1$

$$q_p = \begin{cases} \frac{1}{2}\big(x_{(np)} + x_{(np+1)}\big) & \text{if } np \text{ is an integer,} \\ x_{(\lfloor np \rfloor + 1)} & \text{otherwise.} \end{cases}$$

Note that the sample median $\tilde{x} = q_{0.50}$ is also included in this definition.

The 25 %, 50 % and 75 % quantiles are sometimes called the 1st, 2nd and 3rd **quartiles**.
The **interquartile range (iqr)** is the difference between the 3rd and the 1st quartile:

$$iqr = q_{0.75} - q_{0.25}.$$

Sample quantiles - 2

Example:

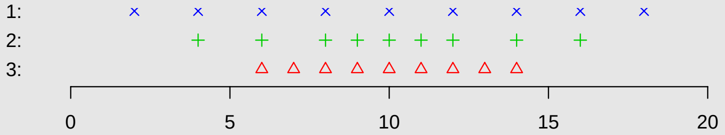
Let us calculate some quantiles of the **ages data set**.

```
> quantile(ages,c(0.05,0.25,0.75,0.95),type=2)
  5%  25%  75%  95%
 9.0 17.5 44.0 56.0
> median(ages)
[1] 35.5
> quantile(ages,0.50,type=2)
 50%
35.5
> IQR(ages)
[1] 25.75
```

Measures of variability

A **measure of the center** is only a **partial information** about a data set. Different samples may well have the same mean (or median), while they differ in other important ways.

The following plot shows three different samples with the **same sample mean** and **different variabilities** (sample 1 has the largest, sample 3 the smallest amount of variability):



The sample variance

The **variability** is responsible for the **spread** of observations **around the mean**. Therefore in studying the variability of a sample, the **deviations from the mean** $x_1 - \bar{x}, \ldots, x_n - \bar{x}$ are of great importance.

This leads us to the **sample variance**, the sum of the squared deviations:

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

The **sample standard deviation** is the (positive) square root of the variance:

$$s = \sqrt{s^2}.$$

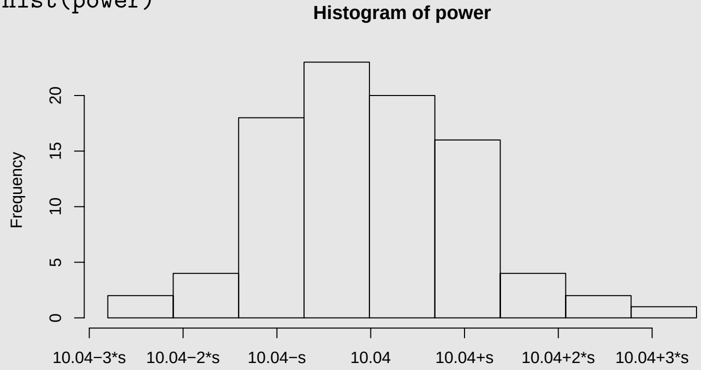
The sample variance - 2

The **scale for** s is the **same as the scale** for each x_i.

If **for example** the observations are **fill quantities** of beer barrels, then we might have $s = 2.0$ liters. Roughly speaking this means that a **typical deviation** is 2.0 liters.
If using another machine to fill the barrels, this could give $s = 3.0$, indicating a larger variability.

The sample variance - power demand example

```
> mean(power)
[1] 10.03844
> var(power) # the sample variance s^2
[1] 8.225368
> sd(power) # the sample standard deviation s
[1] 2.867990
> hist(power)
```



The sample skewness

The sample skewness is defined as

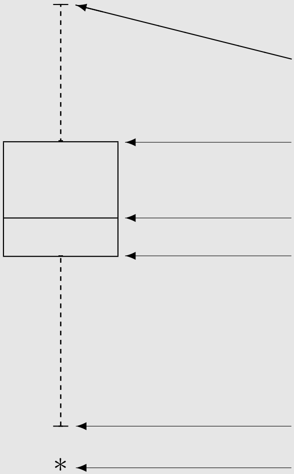
$$g_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}.$$

```
> skewness<-function(x){
   (sum((x-mean(x))^3)/(length(x)-1))/(sd(x)^3)}
> skewness(power)
[1] 0.2864764
```

	$g_1 \approx 0$	probably symmetric distribution
Interpretation:	$g_1 > 0$	long tail to the right
	$g_1 < 0$	long tail to the left

The boxplot

- visual summary of data
- based on different quantiles and extrema



$\min\left(x_{(n)}, q_{0.75} + 1.5\,iqr\right)$

$[iqr = q_{0.75} - q_{0.25}]$

$q_{0.75}$

$\tilde{x} = q_{0.50}$

$q_{0.25}$

$\max(x_{(1)}, q_{0.25} - 1.5\,iqr)$

Outlier $x_{(1)} < q_{0.25} - 1.5\,iqr$

The boxplot - 2

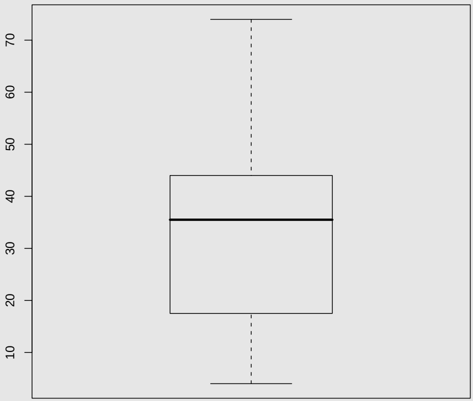
Advantages:

- **easy overview of characteristics** like the median and the range
- recognition of **skewness or symmetry**
- shows the **length of the tails** and **outliers**
- good tool to **compare** different **samples**
- **no need of** (subjective) choices of **parameters**

The boxplot - Example

This is the boxplot of the ages data set:

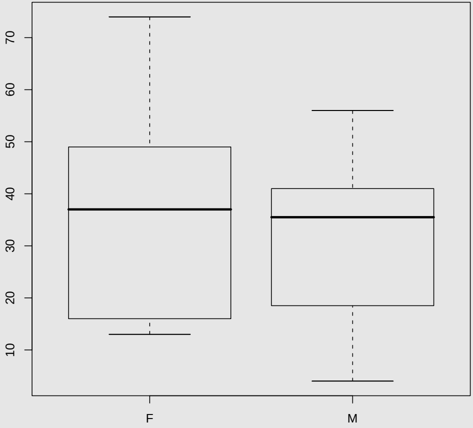
```
> boxplot(ages)
```



The boxplot - Example 2

Suppose we **also know** the **sex of our neighbors** (with **F=female** and **M=male**), then we might want to study and compare the ages of each group:

```
> boxplot(ages~sex)
```



The Q-Q plot

- Q-Q plot stands for **Quantile-Quantile** plot.
- It is used to **compare** a **sample** with a **theoretical population distribution**.
- x-axis: **theoretical quantiles**
- y-axis: **sample quantiles**

We use the $\frac{i-.5}{n}$th quantiles for $1 \leq i \leq n$, thus we compare

$$x_{(i)} \qquad \text{with} \qquad x_{\frac{i-.5}{n}}$$

of a possibly suitable distribution.

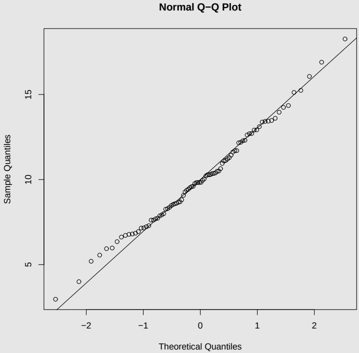
If the **chosen distribution** is **supported by the data**, the **points** in the plot should **nearly form a line**.

The Q-Q plot - example

Let us have a look at a Q-Q plot of the **power demand data**.

As we have already seen in its histogram, the data seems to be **close** to a **normal distribution**.

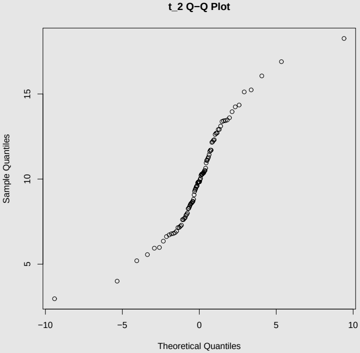
```
> qqnorm(power)
> qqline(power)
```



The points are **very close to a line**, this is an indicator of a **normal distribution** of the power demand population.

The Q-Q plot - example 2

Suppose we want to compare the same data with a different distribution, say the t-distribution with parameter 2 (longer tails).



The plot does not suggest an underlying t-distribution.

The scatter plot

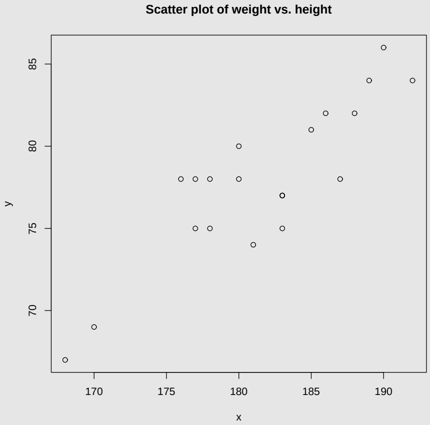
- used for **bivariate data**
- **data pairs** (x_i, y_i): (14.2,5.2), (27.5,6.3), ...
- each pair as a **dot** in a 2-dimensional coordinate system
- helps to find **relationships between variables** x and y

Example:
In a sample we have the height (x in cm) and weight (y in kg) of 20 male statistics students.

Obs.:	1	2	3	4	5	6	7	8	9	10
x:	178	177	190	186	183	168	188	180	177	192
y:	78	75	86	82	75	67	82	78	78	84
Obs.:	11	12	13	14	15	16	17	18	19	20
x:	185	181	183	176	189	170	183	187	178	180
y:	81	74	77	78	84	69	77	78	75	80

The scatter plot - 2

```
> plot(x,y,main="Scatter plot of weight vs. height")
```

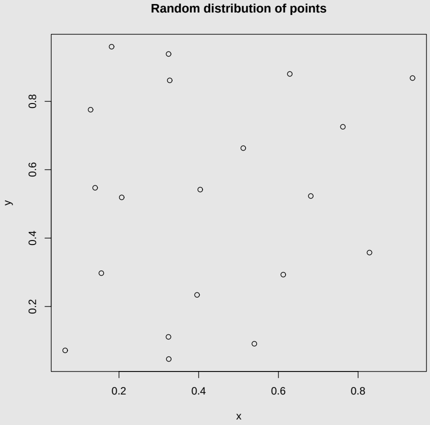


We see an indication of a linear relationship between the two variables:

The taller a student, the heavier he is.

The scatter plot - 3

A scatter plot may also show a seemingly **random distribution of points**:



Scatter plot matrix

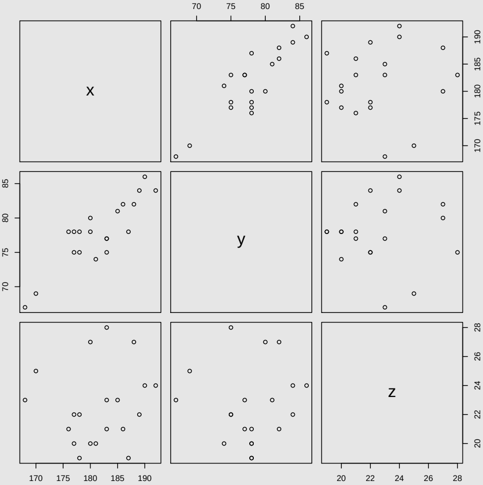
When looking at **more than two numerical variables**, we might like to study the **relationships between each pair** of variables. That is, we want a scatter plot of each variable pair. This is realized within a **scatter plot matrix**.

Example:
Suppose we also know the age of the 20 statistics students from before:

Obs.:	1	2	3	4	5	6	7	8	9	10
z:	19	22	24	21	28	23	27	20	20	24
Obs.:	11	12	13	14	15	16	17	18	19	20
z:	23	20	23	21	22	25	21	19	22	27

Scatter plot matrix - 2

```
> plot(data.frame(x,y,z))
```



Résumé of introduced numerical summary measures

- measures of the center
  1. sample mean $\bar{x}$
  2. sample median $\tilde{x}$
- measures of variability
  1. sample variance s^2
  2. sample standard deviation s
- sample skewness g_1
- sample quantiles q_p

Résumé of plots for a first graphical analysis

- Stem-and-Leaf Display
- Histogram
- Barchart
- Piechart
- Boxplot
- Q-Q Plot
- Scatter Plot (Matrix)

Literature

Devore, J., and Farnum, N. (2005):
Applied Statistics for Engineers and Scientists, 2nd ed., Thomson Learning.

Krämer, W. (2001):
Statistik verstehen, 7th ed., Piper.