Statistics & Econometrics

for CS|DS@UCU

University of Augsburg
Chair of Statistics and Data Science
Prof. Dr. Yarema Okhrin

Introduction

Statistics deals with the analysis of processes that are driven by random factors. For this purpose we collect real data on the process. There are numerous methods and tools developed to help us to collect, describe, analyze, and draw conclusions from data (observations).

Wikipedia: Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data.

Examples: number of clicks on a ad-banner, number of orders of a particular product, price of a particular financial assets, number and size of insurance claims, creditability of a particular company, customer churn, etc

Econometrics deals with the modelling of causal dependence between one or several dependent variable and set a set of explanatory variables. Thus it aims to “explain” the relationship. Special tools for forecasting, modelling specific types of data and specific functional relationships.

Examples: impact of expenditures on ad campaigns, training for employees, quality assurance, research, etc. on the sales/profit

Time series analysis deals with modelling and forecasting of time ordered data.

Examples: modelling the dynamics of sales, asset prices, website traffic

Bulding blocks of Statistics

Descriptive Statistics

- Presentation of data using tables and graphs
- Characterizing the data using a few but powerful measures

Probability Theory

- The concept of probability, conditional probability
- random variables, distribution and density function, characterization of RV's

Inferential Statistics

- Inference about the population on the basis of a sample
- Testing statistical hypothesis, building confidence intervals, measuring reliability of tests

Additional advanced components

- Theory of point estimation
- Nonparametric statistics
- Large sample theory
- Bayesian statistics

Chapter 1

Descriptive Statistics

Descriptive Statistics

Basic statistical concepts

- real world problem $\rightsquigarrow$ statistical analysis
- The complete set of the objects that are subject of the analysis is called population and is usually denote by Ω. We denote the elements of Ω by ω.
- Note: we are not interested in the population itself, but more in the properties of the population measured by one or several quantities of interest X (characteristics/attributes).

  $X\colon \Omega \to S$, where S is the space of possible values of X.

  $x = X(\omega)$ is called a realization or an observation.

Universität Augsburg

Example: public appeal of a new movie

$\Omega =$ the set of all audience members,

$X =$ (assessment of the movie, age, gender, occupation)

	assessment	age	gender	occupation
1	good	23	m	student
2	very good	14	m	pupil
3	good	19	f	shop assistant
4	satisfactory	35	m	worker
5	adequate	29	f	school teacher

Data sampling

- complete survey: we collect and analyze all elements of Ω (for example, population census).

  Disadvantage: too expensive, too costly, not always feasible in practice (for example, life expectancy of bulbs)

- partial survey: we collect only a small part of the elements of the population.

- The set of the considered elements is called sample.

Classification of variables I

- nominal scale:
  Let x and y denote two realizations of an attribute. If the attribute is nominal, then we can only conclude that either
  $$x = y \text{ (equality) or } x \neq y \text{ (inequality)}$$
  Example: marital status, gender, occupation

- ordinal scale: the realizations can be naturally ordered, i.e. statements with „smaller/less " and „larger/more " have clear interpretation. This implies that for all realizations x and y
  $$x = y \quad \text{or} \quad x > y \quad \text{or} \quad x < y.$$
  Examples: grades, rankings

Classification of variables II

- interval scale: if the differences between two realizations of an ordinally scaled attribute has natural meaning.

  Example: temperature values in Celsius, year of birth

- ratio scale: additionally to definition of the interval scale we require that there is a meaningful non-arbitrary zero in the set of realizations.

  Examples: income, price, turnover, age

- absolute scale: in addition to the interval scale we have a natural, scale-independent unit.

  Examples: quantity, number of students enrolled at a university

Classification of variables III

- An attribute is called qualitative, if it has a finite set of possible realizations and is at most ordinally scaled. The realizations reflect the difference/strength, but not the magnitude (e.g. gender, colour).
- If, however, the realizations reflect both the difference and the magnitude, then we speak about quantitative attributes (for example, age, income, price).
- We observe an increasing informational content by moving from nominal to interval scale, but the observations may suffer from assessment errors.

Classification of variables IV

A variable/attribute is discrete, if the set of possible realizations is a countable set. The attribute/variable is continuous, if it is has uncountably many possible realizations.

Examples: height, speed, time, grade, quality

Note:

- Despite of the fact that many variables are continuous by nature, it is **not** possible to measure them with an arbitrary precision.
- Often a discrete attribute has very many realizations (for example, prices, income). In this case it is reasonable to treat them as continuous attributes.

Long Example : largest companies (2000)

```
## ## install.packages('HSAUR')
data("Forbes2000", package = "HSAUR")
## ??Forbes2000
head(Forbes2000)
##   rank                name        country             category  sales profits
## 1    1           Citigroup  United States              Banking  94.71   17.85
## 2    2    General Electric  United States        Conglomerates 134.19   15.59
## 3    3 American Intl Group  United States            Insurance  76.66    6.46
## 4    4          ExxonMobil  United States Oil & gas operations 222.88   20.96
## 5    5                  BP United Kingdom Oil & gas operations 232.57   10.27
## 6    6     Bank of America  United States              Banking  49.01   10.81
##    assets marketvalue
## 1 1264.03      255.30
## 2  626.93      328.54
## 3  647.66      194.87
## 4  166.99      277.02
## 5  177.57      173.54
## 6  736.45      117.55
## View(Forbes2000)
```

```
G7 <- c("Germany", "France", "Italy", "Japan", "Canada", "United Kingdom", "United States")
ForbesG7 <- Forbes2000[Forbes2000$country %in% G7, ]
ForbesG7 <- ForbesG7[1:500, ]
ForbesG7 <- droplevels(ForbesG7)
str(ForbesG7)
## 'data.frame': 500 obs. of  8 variables:
##  $ rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ name       : chr  "Citigroup" "General Electric" "American Intl Group" "ExxonMobil" ...
##  $ country    : Factor w/ 7 levels "Canada","France",..: 7 7 7 7 6 7 6 5 7 7 ...
##  $ category   : Factor w/ 27 levels "Aerospace & defense",..: 2 6 16 19 19 2 2 8 9 20 ...
##  $ sales      : num  94.7 134.2 76.7 222.9 232.6 ...
##  $ profits    : num  17.85 15.59 6.46 20.96 10.27 ...
##  $ assets     : num  1264 627 648 167 178 ...
##  $ marketvalue: num  255 329 195 277 174 ...
```

```
summary(ForbesG7)
##       rank           name                        country
##  Min.   :  1.0   Length:500         Canada        : 23
##  1st Qu.:162.8   Class :character   France        : 33
##  Median :315.5   Mode  :character   Germany       : 31
##  Mean   :325.1                      Italy         : 14
##  3rd Qu.:493.2                      Japan         : 83
##  Max.   :664.0                      United Kingdom: 51
##                                     United States :265
##                     category        sales             profits
##  Banking                : 66   Min.   :  1.470   Min.   :-25.830
##  Utilities              : 42   1st Qu.:  8.375   1st Qu.:  0.360
##  Insurance              : 37   Median : 14.190   Median :  0.650
##  Consumer durables      : 32   Mean   : 23.605   Mean   :  1.086
##  Diversified financials: 28   3rd Qu.: 27.540   3rd Qu.:  1.383
##  Food drink & tobacco   : 28   Max.   :256.330   Max.   : 20.960
##  (Other)                :267
##      assets          marketvalue
##  Min.   :   3.36   Min.   :  0.940
##  1st Qu.:  13.91   1st Qu.:  8.828
##  Median :  26.02   Median : 14.560
##  Mean   :  85.85   Mean   : 28.805
##  3rd Qu.:  64.99   3rd Qu.: 29.858
##  Max.   :1264.03   Max.   :328.540
##
```

Characteristics of univariate data sets

Starting point: the quantity of interest X

- the sample $x_1, .., x_n$ with $x_i \in \mathbb{R}$ (univariate);
- let $a_1, ..., a_k$ denote all possible but different realizations

absolute frequency of a_i:
$n(a_i) =$ frequency of the occurrence of the realization a_i in the sample

relative frequency of a_i: $h(a_i) = n(a_i)/n$

Example A firm observed the following delivery times (in days) for the last 50 orders.

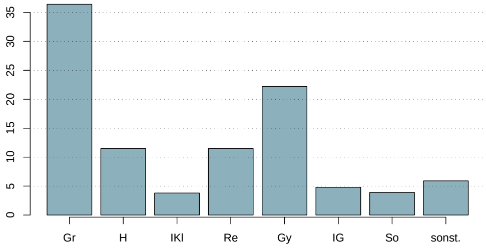
7 8 7 3 8 7 5 7 8 9 9 8 8 7 10 7 9 8 9 7 8 7 10 8 8
9 10 7 10 9 9 10 7 8 7 10 10 8 8 8 8 9 9 7 8 5 8 7 10 8

Realisations (ordered)	a_j	3	5	7	8	9	10	$\sum$
abs. frequency	$n(a_j) = n_j$	1	2	13	17	9	8	50
rel. frequency	$h(a_j) = n(a_j)/n$	$\frac{1}{50}$	$\frac{2}{50}$	$\frac{13}{50}$	$\frac{17}{50}$	$\frac{9}{50}$	$\frac{8}{50}$	1

Graphical presentation of the frequencies I

bar plot: for each realization we draw bars/sticks. The height of the bars equals the absolute OR relative frequency.

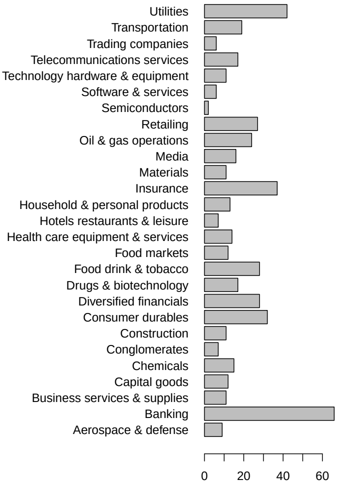
Example:
Out of 9 558 455 pupils in Germany (in 1993) 36.4% went to elementary school, 11.5% to secondary modern, 3.8% to integrated secondary and junior high school, 11.5% to junior high school, 22.2% to "Gymnasium", 4.8% to integrated school, 3.9% to special schools and 5.9% to other types of schools.

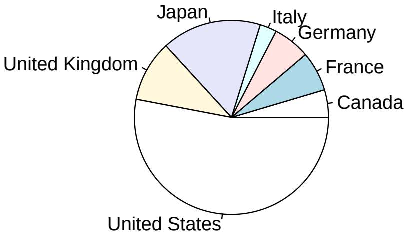


Pie chart

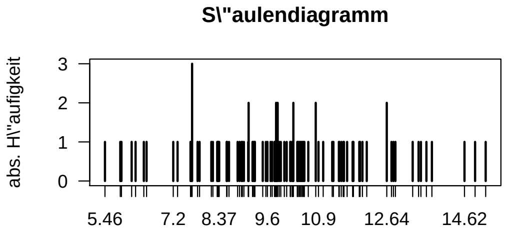
Angle: the square is proportional to the frequency:

$$w_j = 360^\circ h(a_j)$$





Problem: if we have a continuous variable or a discrete one with many outcomes, then the bar plot is not informative.



Solution: histogram

Histogram

(a) Let $K_j : [x_0 + (j-1)h, x_0 + jh), \quad j \in \mathbb{Z}$ be the classes of possible values with starting point y_0 and bandwidth h;

(b) count the observations in each K_j (class frequency $n(Kj)$);

(c) calculate the relative class frequency $h(K_j) = n(K_j)/n$, where n is the sample size;

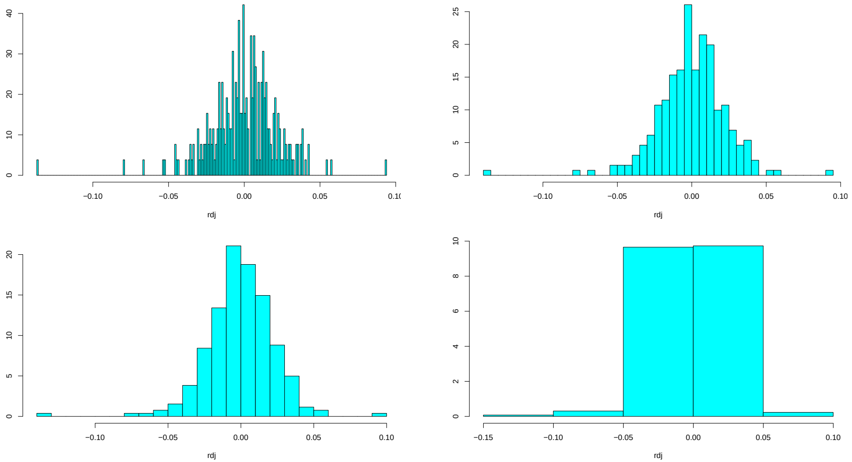
(d) normalise to 1: $f_j = \frac{n(K_j)}{nh}$ (relative class frequency divided by h);

(e) plot rectangles of height f_j for each class K_j.

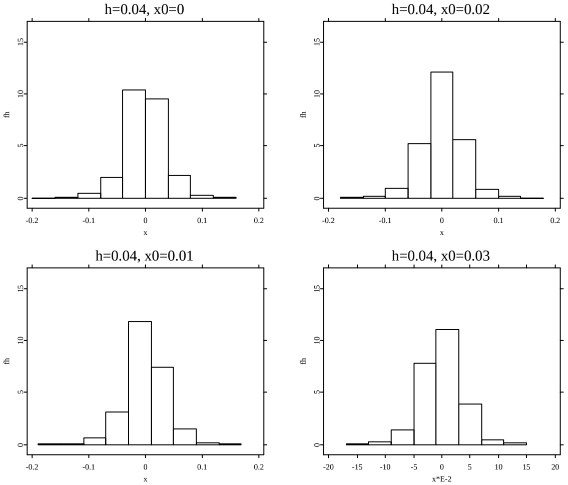
Histogram

$$\hat{f}_h(x) = h(K_j)/h \qquad \text{for} \qquad x \in K_j$$

Here: Dow Jones index returns with the bandwidth $h = 0.001, 0.005, 0.01, 0.05$



Four histograms for the same data with different starting points: $x_0 = 0$, $x_0 = 0.01$, $x_0 = 0.02$, $x_0 = 0.03$; bandwidth $h = 0.04$



conditions on the classes:

- disjunct classes
- each realization falls in one of the classes
- desirable: all classes have equal width
- the square above the class K_i: $h(K_i)/|K_i| \cdot |K_i| = h(K_i)$, i. e. the key information about the histogram is revealed by the squares of the rectangles!
- $$\int_{-\infty}^{\infty} \hat{f}(x)\, dx = \sum_{i=1}^{k} h(K_i) = 1$$
- special method are required to determine the “best” bandwidth

Empirical cumulative distribution function

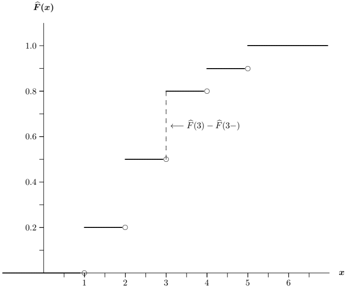
Requirement: at least the ordinal scale

empirical cumulative distribution function (ECDF):

$\hat{F}(x) =$ relative number of observations equal to or less than x

$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$

Example: public appeal of a movie (grades: $1, 1, 2, 2, 2, 3, 3, 3, 4, 5$)



Universität Augsburg

Properties of the ECDF:

a) $\hat{F}(x) = 0$ for $x < x_{(1)}$, $\hat{F}(x) = 1$ for $x \geq x_{(n)}$

b) $\hat{F}(x)$ is increasing

c) $\hat{F}(x)$ is continuous from the right

d) $\hat{F}(x_j) - \hat{F}(x_j-) =$ relative frequency of x_j

Note: The ECDF contains all the information about the sample in an aggregated form.

Characteristics/Parameters

Parameters are measures, that quantify important characteristics of the empirical distribution function.

Important parameters are e.g.:

- Location parameter: Gives insights into the central tendency of the the data.
- Dispersion measure: Contains information about the variability of the data.

Overview

parameters of the variables

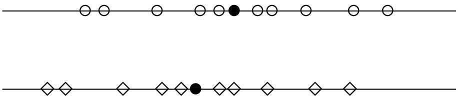
- location
  - arithmetic mean
  - mode
  - median
  - quantile
- dispersion measure
  - range
  - interquartile range
  - mean absolute deviation
  - variance and standard deviation

Location measure

Mean characterizes the central location of the data.

Example: Monthly personal income of elves and orcs in €

Elves: 1000, 1200, 1750, 2200, 2400, 2800, 2950, 3300, 3800, 4150 ($\circ$)
$\bar{x}_{elf} = 2555$ € ($\bullet$)
Orcs: 600, 800, 1350, 1800, 2000, 2400, 2550, 2900, 3400, 3750 ($\diamond$)
$\bar{x}_{orc} = 2155$ € ($\bullet$)



i) Mean (arithmetic mean, average)

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} n(a_i)\, a_i \;=\; \sum_{i=1}^{k} h(a_i)\, a_i$$

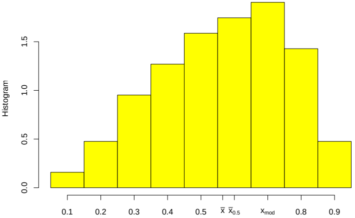
Properties:

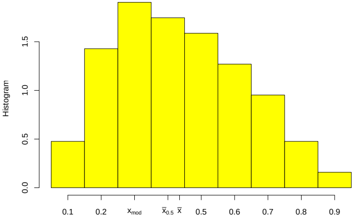
- The mean is the value with the smallest possible mean-squared deviation, i. e. it holds for all $a \in \mathbb{R}$

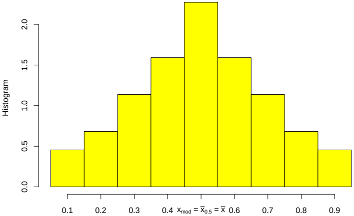
$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \leq \quad \sum_{i=1}^{n}(x_i - a)^2 .$$

- The mean is very sensitive to outliers (for example, monthly income of 1000.0, 1000.0, 1000.0, 10000.0 returns $\bar{x} = 3250$).
- Note: the mean is meaningful **only** for symmetric data. Otherwise it is difficult to draw conclusions.

Symmetric and nonsymmetric distributions







ii) α– trimmed mean $\bar{x}_\alpha$

$x_{(i)}$ is the i–th order statistics, if $x_{(i)}$ is on the i-th position in the ordered sample.

α–trimmed mean

$$\bar{x}_\alpha = \frac{1}{n - 2\,[n\,\alpha]} \sum_{i=[n\,\alpha]+1}^{n-[n\,\alpha]} x_{(i)}$$

with $\alpha \in [0, 0.5)$, $[z]$ denotes the largest natural number that is smaller than z

Example: grades 2.7, 3.0, 3.0, 3.0, 3.3, 3.3, 3.3, 3.7, 4.0, 6.0
It holds that $\bar{x} = 3.53$, but $\bar{x}_{0.1} = 26.6/8 = 3.325$.
Note: it is much more robust to outliers compared to the simple mean

iii) p–quantile $\tilde{x}_p$

p–quantile

$$\tilde{x}_p = \begin{cases} x_{([n\,p]+1)} & \text{for } n\,p \notin \mathbb{Z} \\ \left(x_{(n\,p)} + x_{(n\,p+1)}\right)/2 & \text{for } n\,p \in \mathbb{Z} \end{cases}, \quad p \in (0,1]$$

$\tilde{x}_{0.25}$ is called the lower quartile , $\tilde{x}_{0.5}$ is the median and $\tilde{x}_{0.75}$ is the the upper quartile

- The arithmetic mean is not robust to outliers.
- The median is, however, **robust**, as it is determined by the ranks of the observations and not by the exact values.

Sample quantiles correspond to $\hat{F}^{-1}(p)$ (in some sense)

Example: Demand for a particular commodity, $n = 10$

sample x_i									
10	23	20	33	50	20	20	13	50	33

ordered sample $x_{(i)}$									
10	13	20	20	20	23	33	33	50	50

Thus:

$$\begin{aligned}
\tilde{x}_{0.25} &\underset{10\cdot 0.25=2.5\notin\mathbb{Z}}{=} x_{(\lfloor 2.5\rfloor+1)} = x_{(3)} = 20,\\
\tilde{x}_{0.5} &\underset{10\cdot 0.5=5\in\mathbb{Z}}{=} \frac{1}{2}\left(x_{(5)} + x_{(5+1)}\right) = \frac{1}{2}(20+23) = 21.5,\\
\tilde{x}_{0.75} &\underset{10\cdot 0.75=7.5\notin\mathbb{Z}}{=} x_{(\lfloor 7.5\rfloor+1)} = x_{(8)} = 33.
\end{aligned}$$

Properties:

- The number of observations, which are smaller than $\tilde{x}_p$ or equal to $\tilde{x}_p$, is larger or equal to $[n\,p]$.
- It holds that $x_{([n\,p])} \leq \tilde{x}_p \leq x_{([n\,p]+1)}$.
- It holds for $a \in \mathbb{R}$ that

$$\sum_{i=1}^{n} |x_i - med| \quad \leq \quad \sum_{i=1}^{n} |x_i - a|$$

  i.e. the median minimizes the mean absolute deviation to all data points.
- The median can also be used to characterize asymmetric data.

Linear transformation of location measures

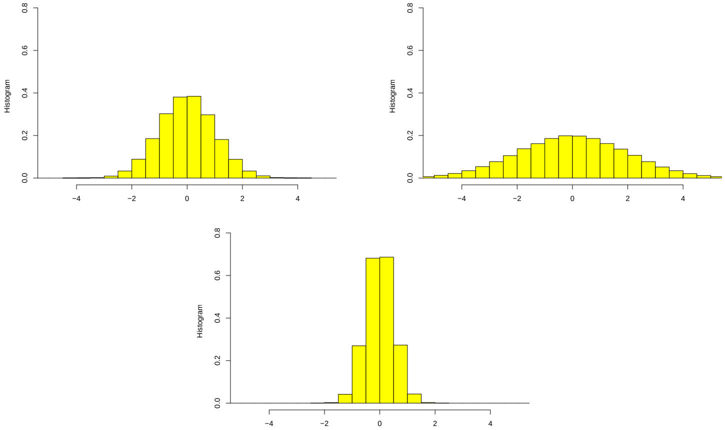
If we transform the data linearly

$$y_i = a + b \cdot x_i$$

then the same holds for the location measures too:

$$\begin{aligned}
\bar{y} &= a + b \cdot \bar{x} \\
\bar{y}_\alpha &= a + b \cdot \bar{x}_\alpha \\
y_{\text{Med}} &= a + b \cdot x_{\text{Med}} \\
\tilde{y}_p &= a + b \cdot \tilde{x}_p
\end{aligned}$$

Dispersion/Volatility/Variability



Volatility measures

Problem: the location measures do not characterize the data sufficiently
Aim: statements about the variation of the data around the center (a location measure)

i) range

$$\tilde{R} = x_{(n)} - x_{(1)}$$

Note: the range is *extremely* sensitive to the data/outliers.
ii) interquartile range

$$QA = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

Properties:
a) the interquartile range is robust to outliers.
b) There are at least $[n/2]$ of all observations in the interval $[\tilde{x}_{0.25}, \tilde{x}_{0.75}]$

iii) empirical variance

$$\tilde{s}^2 \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n(a_i)\,(a_i - \bar{x})^2 \;=\; \sum_{i=1}^{k} h(a_i)\,(a_i - \bar{x})^2$$

$\tilde{s^2}$ is the average squared deviation of the observations from the mean.

iv) sample variance

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

- s is the sample standard deviation. $\tilde{s} = \sqrt{\tilde{s^2}}$ is the empirical standard deviation.
- The empirical/sample variance/standard deviation is very sensitive to outliers .
- The empirical/sample variance/standard deviation is only reasonable for symmetric data.

Example: price of pizza $x_1 = (6,\ 8,\ 5,\ 5,\ 6)$ mit $\bar{x}_1 = 6$

$$\tilde{s}^2 = \frac{2 \cdot 5^2 + 2 \cdot 6^2 + 8^2}{5} - 6^2 = \frac{186}{5} - 36 = 37.2 - 36 = 1.2\,,$$

$$\tilde{s} \approx 1.095\,.$$

Example: price of pizza with an outlier $x_2 = (6,\ \mathbf{18},\ 5,\ 5,\ 6)$ mit $\bar{x}_2 = 8$

$$\tilde{s}^2 = \frac{2 \cdot 5^2 + 2 \cdot 6^2 + 18^2}{5} - 8^2 = \frac{446}{5} - 64 = 89.2 - 64 = 25.2\,,$$

$$\tilde{s} \approx 5.02\,.$$

iv) MAD - median of the absolute deviation from the median

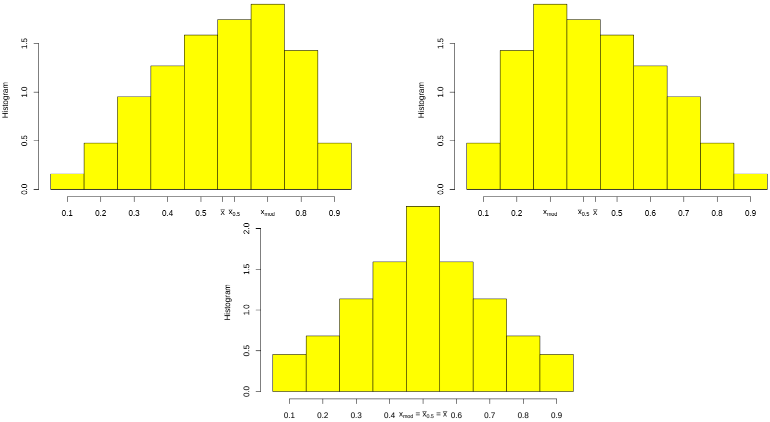
$$\text{mad} = \text{Median of } |x_i - \tilde{x}_{0.5}|\,,\ i = 1, \ldots, n$$

Linear transformations of volatility measures: $y_i = a + b \cdot x_i$

$$\tilde{R}_y = |b| \cdot \tilde{R}_x$$
$$\tilde{s}_y^2 = b^2 \cdot \tilde{s}_x^2$$
$$\tilde{s}_y = |b| \cdot \tilde{s}_x$$
$$MAD_y = |b| \cdot MAD_x$$

Measures of skewness

Symmetric and nonsymmetric distributions



Aim: statements about the asymmetry of a sample

Note: it is reasonable only for unimodal distributions.

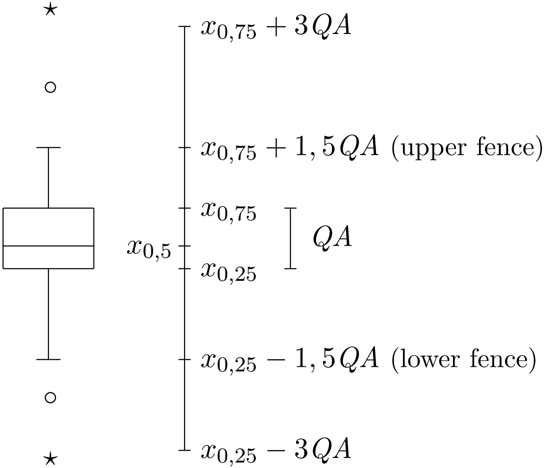
A distribution is right-skewed, if the peak is located at the left part of the distribution. Otherwise the distribution is left-skewed.

Sample skewness (empirical skewness)

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\tilde{s}}\right)^3$$

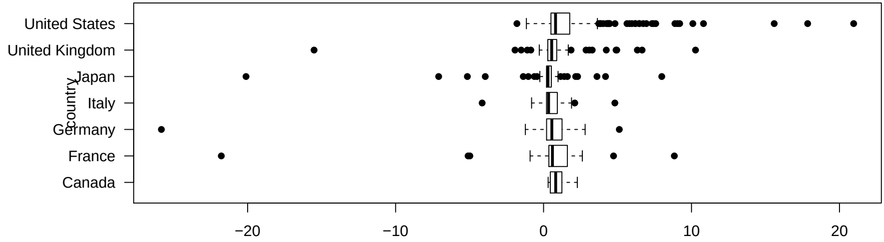
If it is larger (smaller) than zero, then we conclude that the distribution is right-skewed (left-skewed).

Boxplot - Graphical representation of some measures of location and variation



Example: Forbes

```
apply(ForbesG7[, stetigeVar], 2, sd)
##       sales     profits      assets marketvalue
##   29.463847    3.043527  169.299051   41.041540
apply(ForbesG7[, stetigeVar], 2, function(x) max(x) - min(x))
##       sales     profits      assets marketvalue
##      254.86       46.79     1260.67      327.60
apply(ForbesG7[, stetigeVar], 2, IQR)
##       sales     profits      assets marketvalue
##     19.1650      1.0225     51.0750     21.0300
boxplot(profits ~ country, data = ForbesG7, horizontal = TRUE, las = 1, pc
```



Measures of concentration/inequality

Example: 5 companies and 25M customers. If every company has 5M customers, then no concentration. If one has 20M, then strong concentration.

Idea: how much does a single observation contribute to the total?

Aim: Which fraction of the total sum make the $u\%$ of the smallest observations?

Note: ordered data $x_i \mapsto x_{(i)}$!

Lorenz curve:

Streckenzug: $(0,0),\ (u_1, v_1),\ \ldots,\ (u_n, v_n) = (1,1)$ mit

$$u_i = \text{fraction of the } i \text{ smallest observ.} \;= \frac{i}{n}$$

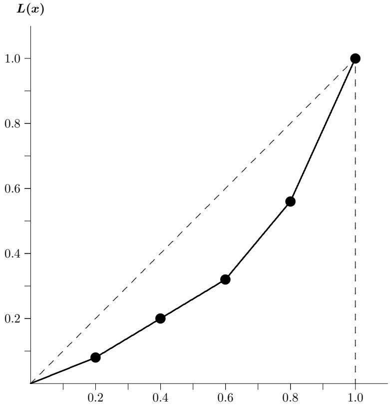
$$v_i = \text{fraction of the sum of } i \text{ smallest on the total sum} \;= \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}$$

Example I:

Five companies with customers: 6, 3, 11, 2, 3 (M)

$$\Rightarrow n = 5, \quad \sum_{k=1}^{5} x_k = 25$$

i	1	2	3	4	5
$x_{(i)}$	2	3	3	6	11
u_i	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$v_i = \dfrac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{5} x_{(j)}}$	$\frac{2}{25}$	$\frac{5}{25}$	$\frac{8}{25}$	$\frac{14}{25}$	1

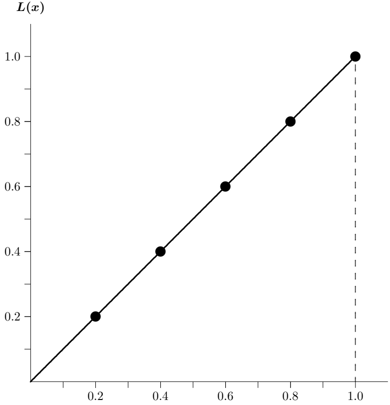


Example II:

Five companies with customers : 5, 5, 5, 5, 5 (M)

$$\Rightarrow n = 5, \quad \sum_{k=1}^{5} x_k = 25$$

i	1	2	3	4	5
$x_{(i)}$	5	5	5	5	5
u_i	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{5} x_{(j)}}$	$\frac{5}{25}$	$\frac{10}{25}$	$\frac{15}{25}$	$\frac{20}{25}$	$\frac{25}{25} = 1$



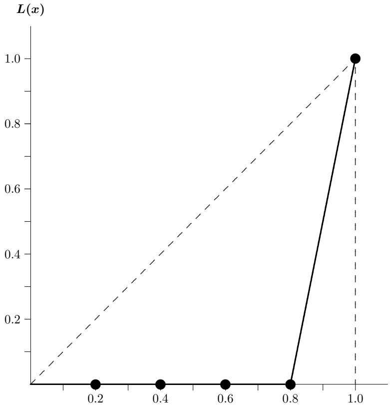
$\Rightarrow$ equal distribution

Example III:

Five companies with customers : 0, 0, 0, 0, 25 (M)

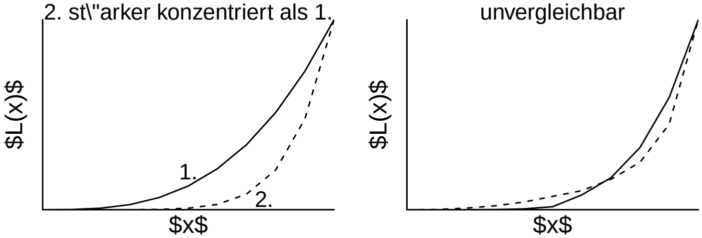
$$\Rightarrow n = 5, \quad \sum_{i=1}^{5} x_i = 25$$

i	1	2	3	4	5
$x_{(i)}$	0	0	0	0	25
u_i	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1
$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}$	0	0	0	0	$\frac{25}{25} = 1$



$\Rightarrow$ extreme concentration

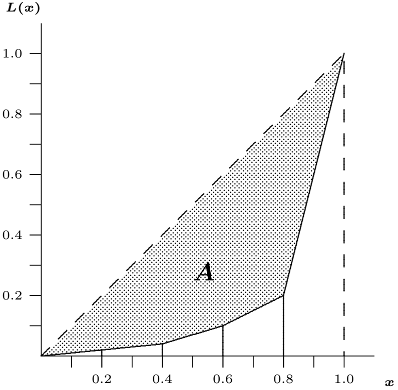
Comparison and properties of Lorenz curves:



- $0 \leq x \leq 1$
- $0 \leq L(x) \leq 1$ with $L(0) = 0$ and $L(1) = 1$
- $L(x) \leq x$
- $L(x)$ is convex
- $L(x)$ is a monotone non-decreasing function

Gini coefficient

Aim: measure of concentration



Use A, i.e. the square between the Lorenz curve and the bisector!

- Numerical measure of concentration:

$$G = \frac{2\sum_{i=1}^{n} i x_{(i)} - (n+1)\sum_{i=1}^{n} x_{(i)}}{n\sum_{i=1}^{n} x_{(i)}}$$

- **Problem:** $G_{\max} = \frac{n-1}{n}$
- **normalized Gini coefficient:**

$$G_* = \frac{n}{n-1} \cdot G \quad \in [0;1]$$

- Larger G_* implies stronger concentration.

Example:
Dour firms with revenues: 6, 3, 11, 2, 3 (Mio.) €

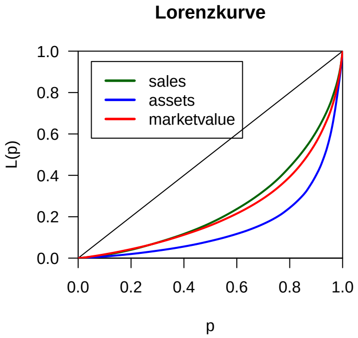
i	1	2	3	4	5	$\sum$
$x_{(i)}$	2	3	3	6	11	25

$$G = \frac{2 \cdot \left(1 \cdot 2 + 2 \cdot 3 + 3 \cdot 3 + 4 \cdot 6 + 5 \cdot 11\right) - 6 \cdot 25}{5 \cdot 25} = 0.336$$

With $G_{\max} = \frac{5-1}{5} = 0.8$ we have $G_{*} = \frac{5}{5-1} \cdot 0.336 = 0.42$

Example: Forbes

```
stetigeVar <- c("sales", "profits", "assets", "marketvalue")
apply(ForbesG7[, stetigeVar], 2, function(x) Gini(x, corr = TRUE))
##       sales     profits      assets marketvalue
##   0.5086780   0.9984003   0.6964531   0.5438243
plot(Lc(ForbesG7[, "sales"]), col = "darkgreen", main = "Lorenzkurve")
lines(Lc(ForbesG7[, "assets"]), col = "blue")
lines(Lc(ForbesG7[, "marketvalue"]), col = "red")
legend(0.05, 0.95, c("sales", "assets", "marketvalue"), lty = rep(1, 3), lwd = rep(
    3), col = c("darkgreen", "blue", "red"))
```



Further concentration measures

- Herfindahl index:

$$H = \sum_{i=1}^{n} p_i^2 \qquad (\in [\tfrac{1}{n}; 1])$$

- Exponential index:

$$E = \prod_{i=1}^{n} p_i^{p_i} \qquad (\in [\tfrac{1}{n}; 1]) \qquad \text{with} \qquad 0^0 = 1$$

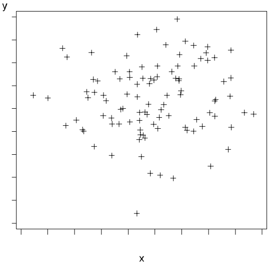
Characteristics of bivariate data sets

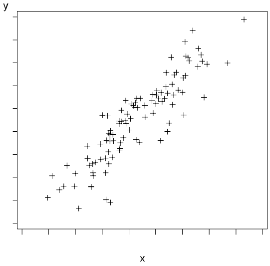
now: 2 variables/attributes X, Y, sample: $(x_1, y_1), \ldots, (x_n, y_n)$

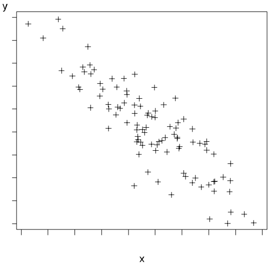
But: for each of the variables we can determine the individual measures of location and volatility as for univariate data sets.

For bivariate data sets we are particularly interested in the relationship between X and Y. This is the subject of the following discussion.

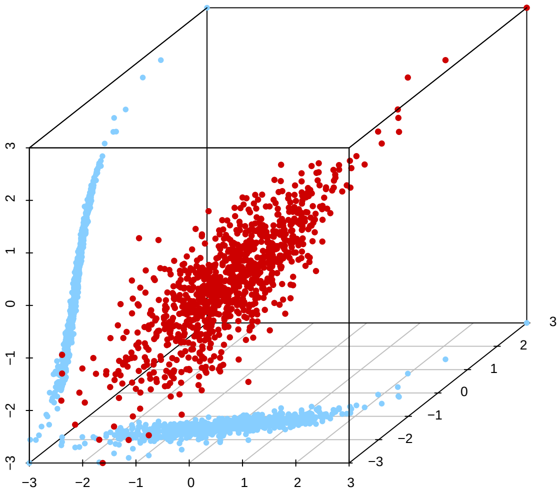
Scatterplots







3D-Scatterplot



Correlation measures for interval-scaled variables

Requirement: X and Y have interval scale

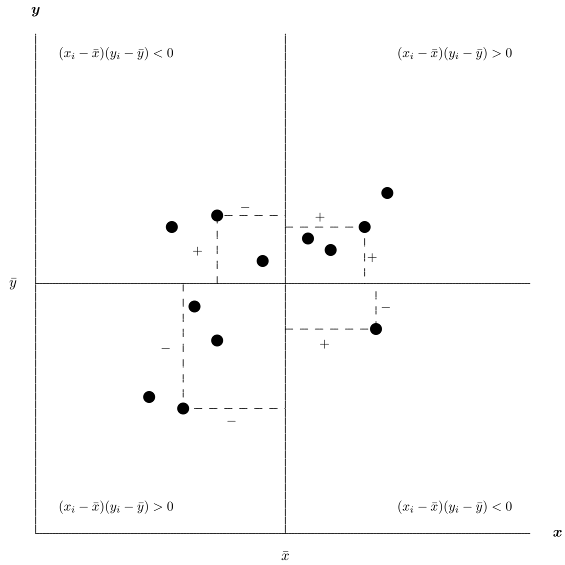
Aim: measure of correlation

positiv relationship: large (small) values of X with large (small) values of Y
negativ relationship: inverse tendency

empirical covariance

$$\tilde{s}_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

An alternativ measure is the sample covariance $s_{XY} = \frac{n}{n-1}\tilde{s}_{XY}$.



y
$(x_i - \bar{x})(y_i - \bar{y}) < 0$
$(x_i - \bar{x})(y_i - \bar{y}) > 0$
$-$
$+$
$+$
$+$
$\bar{y}$
$-$
$-$
$+$
$-$
$(x_i - \bar{x})(y_i - \bar{y}) > 0$
$(x_i - \bar{x})(y_i - \bar{y}) < 0$
x
$\bar{x}$

Properties:

- $\tilde{s}_{XY} = \tilde{s}_{YX}$
- Invariant to shifts in the location, i. e. for $x_i^* = a\,x_i + b$ and $y_i^* = c\,y_i + d$ it holds that $\tilde{s}_{X^*Y^*} = a\,c\,\tilde{s}_{XY}$.
- $|\tilde{s}_{XY}| \le \tilde{s}_X\,\tilde{s}_Y$
- It is sensitive to outliers.

Disadvantage: the empirical variance is not normalized and, therefore, depends on the scale

Sample correlation coefficient of Pearson:

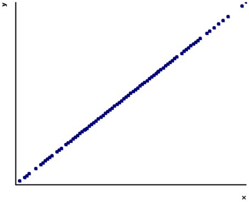
$$r_{XY} = \frac{s_{XY}}{s_X\, s_Y} = \frac{\tilde{s}_{XY}}{\tilde{s}_X\, \tilde{s}_Y}$$

Properties:

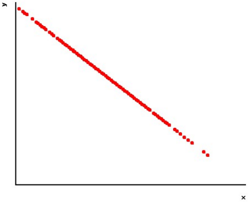
- $r_{XY} = r_{YX}$
- Invariant with respect to shifts in the location **and** in the scale
- $|r_{XY}| \leq 1$.
- If $r_{XY} = 1$ (or -1), then all observations (x_i, y_i) , $i = 1, \ldots, n$ lie on a single straight line with positive (negative) slope.
- The empirical correlation coefficient is a measure of linear dependence between two variables.
- We cannot conclude about casuality of the relationship!

Perfect correlation

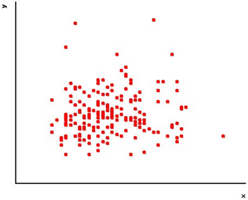
$r = +1$



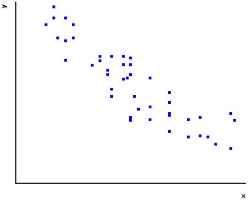
$r = -1$



weak correlation



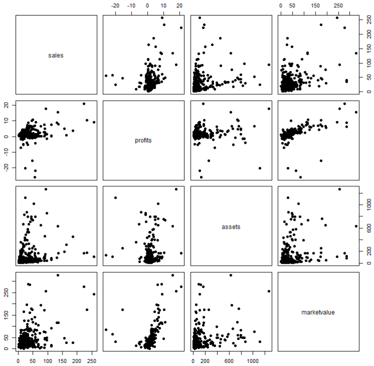
strong correlation



Universität Augsburg

Example: Forbes

```
pairs(~sales+profits+assets+marketvalue, data=ForbesG7, pch=1
```




```
> cor(ForbesG7[,stetigeVar])
                sales   profits    assets marketvalue
sales       1.0000000 0.3692856 0.3169091   0.5522812
profits     0.3692856 1.0000000 0.1555089   0.5308211
assets      0.3169091 0.1555089 1.0000000   0.3815484
marketvalue 0.5522812 0.5308211 0.3815484   1.0000000
```

Correlation measures for ordinal data

Requirement : X and Y are ordinal

Example: the relationship between the exam results (X, grade: $1, \ldots, 5$) and the participation in tutorials (Y, seldom, regularly, always)

Idea of the ranks: assign to each observation of the sample $x_1, \ldots, x_n$ its position in the ordered sample $x_{(1)},\ldots,x_{(n)}$:

$$R(x_j) = v \quad \Leftrightarrow \quad x_j = x_{(v)}$$

$R(x_j)$ is the rank of the observation x_j.

Example: $x_1 = 2, x_2 = 5, x_3 = 1, x_4 = 3.$ ordered sample: $x_3 < x_1 < x_4 < x_2.$ Thus $R(x_1) = 2,\ R(x_2) = 4,\ R(x_3) = 1,\ R(x_4) = 3.$

Given: sample $(x_1, y_1), \ldots, (x_n, y_n)$; assign to $x_1, .., x_n$ the ranks $R(x_1), .., R(x_n)$ and to $y_1, .., y_n$ the ranks $R(y_1), .., R(y_n)$.

Rank correlation coefficient of Spearman

$$R_{XY} = r_{R(X),R(Y)} = \frac{\sum_{i=1}^{n} \left(R(x_i) - \bar{R}\right) \left(R(y_i) - \bar{R}\right)}{\sqrt{\sum_{i=1}^{n} \left(R(x_i) - \bar{R}\right)^2 \sum_{i=1}^{n} \left(R(y_i) - \bar{R}\right)^2}}$$

with $\bar{R} = (n+1)/2$.

Example: quality management

i	x_i	y_i	$R(x_i)$	$R(y_i)$	$R(x_i)^2$	$R(y_i)^2$	$R(x_i)R(y_i)$
1	2	10	1	5	1	25	5
2	4	7	3	$\frac{1}{2}(3+4) = 3.5$	9	12.25	10.5
3	3	7	2	$\frac{1}{2}(3+4) = 3.5$	4	12.25	7
4	9	3	5	1	25	1	5
5	7	5	4	2	16	4	8
$\sum$			15	15	55	54.5	35.5

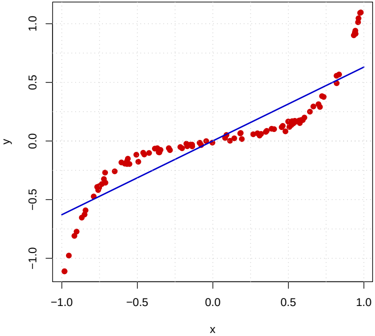
$$\bar{R} = \frac{5+1}{2} = 3$$

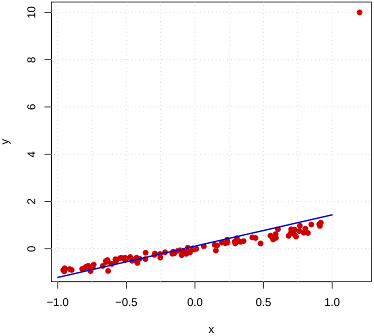
$$r_{SP} = \frac{35.5 - 5 \cdot 3^2}{\sqrt{55 - 5 \cdot 3^2}\sqrt{54.5 - 5 \cdot 3^2}} = -0.97$$

(strong negative monotone correlation)

$$\delta = 0.892, \quad \rho = 0.996$$

$$\delta = 0.659, \quad \rho = 0.982$$





Example: Forbes

```
cor(ForbesG7[,stetigeVar], method="spearman")
```

```
> cor(ForbesG7[,stetigeVar], method="spearman")
                sales   profits    assets marketvalue
sales       1.0000000 0.2602629 0.4738247   0.4856336
profits     0.2602629 1.0000000 0.2636245   0.6450604
assets      0.4738247 0.2636245 1.0000000   0.4716245
marketvalue 0.4856336 0.6450604 0.4716245   1.0000000
```

Correlation measures for nominal variables

Now: 2 nominal variables with realizations $a_1, \ldots, a_k$ for X and $b_1, \ldots, b_l$ for Y

Example: 156 graduates, 93 boys, 63 girls. 9 boys and 2 girls failed the exam.

Contingency table of absolute frequencies:

$\boldsymbol{X}$	$\boldsymbol{Y}$		
	passed	failed	Σ
B	84	9	93
G	61	2	63
Σ	145	11	156

Contingency table of relative frequencies :

$\boldsymbol{X}$	$\boldsymbol{Y}$		
	passed	failed	Σ
B	0.538	0.058	0.596
G	0.391	0.013	0.404
Σ	0.929	0.071	1.0

Bivariate frequency table

- absolute frequency for (a_i, b_j):

  $n_{ij} = n(X = a_i, Y = b_j) =$ the number of cases, where the pair (a_i, b_j) is observed in the sample
- absolute marginal frequency of a_i:
  $n_{i\cdot} =$ the number of cases, where the realization a_i is observed in $x_1, \ldots, x_n$
- the relative frequencies are $h_{ij} = n_{ij}/n$ and $h_{i\cdot} = n_{i\cdot}/n$ respectively.

on analogy: $n_{\cdot j}$, for Y

Example:

- relative marginal frequencies for gender: $(0.596, 0.404)$
- relative marginal frequencies for exam results: $(0.929, 0.071)$

contingency table for absolute frequencies

X	Y: b_1	b_2	$\cdots$	b_l	Σ
a_1	n_{11}	n_{12}	$\cdots$	n_{1l}	$n_{1\cdot}$
a_2	n_{21}	n_{22}	$\cdots$	n_{2l}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	$n_{k\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot l}$	n

Example: Dependence between the success of winning a new customer and the advertising channel

X	Y phone	email	direct mail	$n_{i\bullet}$
	$(= b_1)$	$(= b_2)$	$(= b_3)$	
yes $(= a_1)$	264 $(= n_{11})$	90 $(= n_{12})$	6 $(= n_{13})$	360 $(= n_{1\bullet})$
no $(= a_2)$	2 $(= n_{21})$	34 $(= n_{22})$	4 $(= n_{23})$	40 $(= n_{2\bullet})$
$n_{\bullet j}$	266 $(= n_{\bullet 1})$	124 $(= n_{\bullet 2})$	10 $(= n_{\bullet 3})$	400 $(= n)$

X	Y Phone	email	direct mail	$n_{i\bullet}$
	$(= b_1)$	$(= b_2)$	$(= b_3)$	
NK $(= a_1)$	0.66 $(= h_{11})$	0.225 $(= h_{12})$	0.015 $(= h_{13})$	0.90 $(= h_{1\bullet})$
kein NK $(= a_2)$	0.005 $(= h_{21})$	0.085 $(= h_{22})$	0.01 $(= h_{23})$	0.10 $(= h_{2\bullet})$
$h_{\bullet j}$	0.665 $(= h_{\bullet 1})$	0.31 $(= h_{\bullet 2})$	0.025 $(= h_{\bullet 3})$	1

Aim: a measure of dependency

Idea: weak dependency, if for all i, j

$$n_{ij} \approx \frac{n_{i.}\, n_{.j}}{n}$$

$$\rightsquigarrow \quad \chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.}\, n_{.j}/n}$$

χ^2 „large“ $\rightsquigarrow$ X and Y are dependent.

Since χ^2 increases with n, we consider

The contingency coefficient of Pearson

$$C = \sqrt{\chi^2/(\chi^2+n)}, \text{ with } C_{max} = \sqrt{\frac{\min\{k,l\}-1}{\min\{k,l\}}}$$

Thus

Corrected contingency coefficient of Pearson

$$C_{Corr} = C/C_{max} \in [0,1]$$

The smaller is C_{Corr}, the „weaker "is the dependence. $C_{Corr} = 0$ only if X and Y are independent .

Example: new customers

$$\begin{aligned}
\chi^2 &= n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{n_{i\bullet}\cdot n_{\bullet j}}-1\right)\\
&= 400\cdot\left(\frac{264^2}{360\cdot 266}+\frac{90^2}{360\cdot 124}+\frac{6^2}{360\cdot 10}\right.\\
&+ \left.\frac{2^2}{40\cdot 266}+\frac{34^2}{40\cdot 124}+\frac{4^2}{40\cdot 10}-1\right) = 77.085
\end{aligned}$$

We get:

$$\begin{aligned}
C &= \sqrt{\frac{\chi^2}{\chi^2+n}} = 0.402,\\
C_{\max} &= \sqrt{\frac{\min\{k,\ell\}-1}{\min\{k,\ell\}}} = \sqrt{\frac{\min\{2,3\}-1}{\min\{2,3\}}} = \sqrt{\frac{2-1}{2}} = 0.707
\end{aligned}$$

$\rightsquigarrow$ $C_* = C/C_{\max} = 0.402/0.707 = 0.569 \rightsquigarrow$ average correlation

```
> tab.CountryCategory <- table(ForbesG7$country, ForbesG7$category)

                 Aerospace & defense Banking Business services & supplies Capital g
  Canada                           0       6                            0
  France                           1       5                            0
  Germany                          0       3                            0
  Italy                            1       7                            0
  Japan                            0       5                            7
  United Kingdom                   1       9                            0
  United States                    6      31                            4

> assocstats(tab.CountryCategory)$cont
[1] 0.5632128
```

Chapter 2

Elements of Probability Theory

Probability of events

Origins of probability theory: Jakob Bernoulli (1655-1705), Pierre-Simon de Laplace (1749-1827)

The probability theory originated from the analysis of games of chance (gambling).

Aim: statements about probabilities of random events

- Subsets consisting of a single element of Ω are called elementary events: $\{\omega\} \in \Omega$
- Any subset of Ω is called an event: $A = \{\omega_1, \dots\} \in \Omega$.

Laplace probability

Starting point: All elementary events have the same probability!

If Ω is finite, then it holds

$$P(A) = \frac{\text{the number of for } A \text{ „favourable cases “}}{\text{the number of all possible cases}} = \frac{|A|}{|\Omega|},$$

where $|A|$ denotes the number of elements in A and similarly for $|\Omega|$.

Example: roulette game ($\Omega = \{0, .., 36\}$)

- $A =$ the set of numbers divisible by 3
- $B =$ the even numbers

It holds $P\big(\{0\}\big) = P\big(\{1\}\big) = \cdots = P\big(\{36\}\big) = 1/37$, i.e. it is a Laplace experiment. Then

$$P(A) = \frac{|A|}{|\Omega|} = \frac{12}{37}\ .$$

The probability, that we observe a number of pips, which is divisible by 3, but not divisible by 2, is

$$P(A \cap \bar{B}) = \frac{|\{3, 9, 15, 21, 27, 33\}|}{37} = \frac{6}{37}\ .$$

Statistical probability

Let $A \subset \Omega$. The experiment is repeated n times. $h(A)$ denotes the relative frequency of A.

Example: roulette ($\Omega = \{0, 1, \ldots, 36\}$)
Let A be the event "*we observe a number from the first dozen*", i.e. $A = \{1, 2, .., 12\}$.

16 replications produce the sample

23	34	13	11	28	9	8	21
16	33	31	15	3	13	23	32

Then $h(A) = \frac{4}{16} = 0.25$.

Example: We throw a coin n times. We obtain

n	$n(H)$	$h_n(H)$
10	7	0.700
20	11	0.550
100	47	0.470
400	204	0.510
1000	492	0.492
2000	1010	0.505

The coin is symmetric. Therefore the relative frequencies converge to the true probability of 0.5.

Richard von Mises (1931)

The probability of observing A:

$$P(A) := \lim_{n\to\infty} h_n(A)$$

Disadvantages: difficult to implement in practice

Axioms of the probability theory

Both the Laplace probability and the statistical probability have their pros and cons. A general approach to probability was suggested by Kolmogorov (1933).

A. N. Kolmogorov (1933)

The probability measure P is mapping, which assigns a number to (almost all) events $A \subseteq \Omega$ (namely $P(A)$) and fulfills the following properties:

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for all $A_i \subset \Omega$ with $A_v \cap A_j = \emptyset$ for $v \neq j$.

$P(A)$ is the probability of event A.

Rules for the probabilities

Let P be a probability measure on Ω. Then it holds:

- $P(\bar{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A) = P(A \cap B) + P(A \cap \bar{B})$
- If $B \subseteq A$, then $P(B) \leq P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If Ω is finite, then it holds for $A \subseteq \Omega$ that:

$$P(A) = \sum_{a \in A} P(\{a\}) \,.$$

Note: Both the Laplace probability and the statistical probability are probability measures.

Conditional probability and independence

Now: conditional probability of event A under the condition B (of A, if B is given or observed)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{for} \quad P(B) > 0$$

Note: $P(A \mid B) + P(\bar{A} \mid B) = 1$

Law of total probability

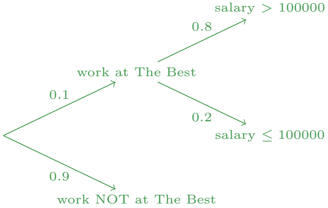
Let $A_1, \ldots, A_k$ be events, which are disjoint in pairs, with $A_1 \cup \ldots \cup A_k = \Omega$. Then for an arbitrary event B it holds

$$P(B) = \sum_{i=1}^{k} P(B \mid A_i) \cdot P(A_i)$$

Tree diagram:

$$\xrightarrow{\quad P(A) \quad} \boxed{A} \xrightarrow{\quad P(B|A) \quad} \boxed{B}$$

Example: job and salary
A: work at „The Best“
B: salary more than 100 000 Euro
$P(A) = 0.1$, $P(B \mid A) = 0.8$



$$\text{We get} \quad \begin{aligned} P(A \cap B) &= P(B|A) \cdot P(A) = 0.8 \cdot 0.1 = 0.08, \\ P(A \cap \bar{B}) &= P(\bar{B}|A) \cdot P(A) = 0.2 \cdot 0.1 = 0.02. \end{aligned}$$

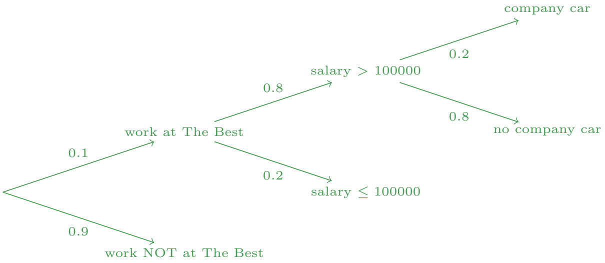
Aim: generalize for more events, e.g.

$$P(A_1 \cap A_2 \cdots \cap A_k)$$

Example: A: work at „The Best“ $\quad$ B : salary > 100000

C: company car in 3 years

$P(A) = 0.1,\ P(B \mid A) = 0.8,\ P(C \mid A \cap B) = 0.2$



We get

$$P(A \cap B \cap C) = \underbrace{P(A \cap B)}_{=P(A) \cdot P(B \mid A)} \cdot P(C \mid A \cap B)$$
$$= 0.1 \cdot 0.8 \cdot 0.2 = 0.016\,.$$

Chain rule

Let $A_1, \ldots, A_k$ be random events with $P(A_1 \cap \cdots \cap A_{k-1}) > 0$. Then for all$k \geq 2$

$$\begin{aligned} & P(A_1 \cap \cdots \cap A_k) \\ & = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \\ & \quad \cdot \ldots \cdot P(A_k \mid A_1 \cap \cdots \cap A_{k-1}). \end{aligned}$$

Example: Three machines produce 20%, 40% and 40% of the total output of a given product. We know from experience that the 1st machine manufactures in 5% of cases a faulty product, the 2nd - in 10% and the 3rd in 20%. We randomly pick up one product. What is the probability that it is defective?

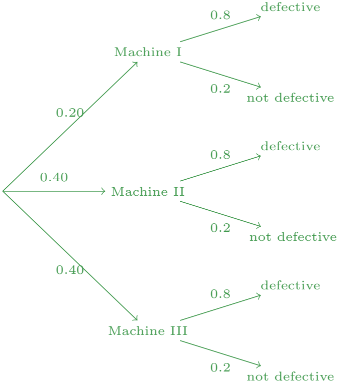
B : "defective product"

A_1 : "manufactured on the 1st machine"

A_2 : "manufactured on the 2nd machine"

A_3 : "manufactured on the 3rd machine"

$$P(A_1) = 0.2,\ P(A_2) = 0.4,\ P(A_3) = 0.4,$$
$$P(B|A_1) = 0.05,\ P(B|A_2) = 0.1,\ P(B|A_3) = 0.2.$$



The law of total probability provides:

$$P(B) = \sum_{i=1}^{3} P(B|A_i)P(A_i) = 0.05 \cdot 0.2 + 0.1 \cdot 0.4 + 0.2 \cdot 0.4 = 0.13$$

Bayes' rule (1702 – 1761)

Let $A_1, \dots, A_k$ be events, which are disjoint in pairs with $A_1 \cup \dots \cup A_k = \Omega$. Furthermore, let B be an arbitrary event. Then it holds for $i \in \{1, \dots, k\}$

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum\limits_{j=1}^{k} P(B \mid A_j) \cdot P(A_j)}.$$

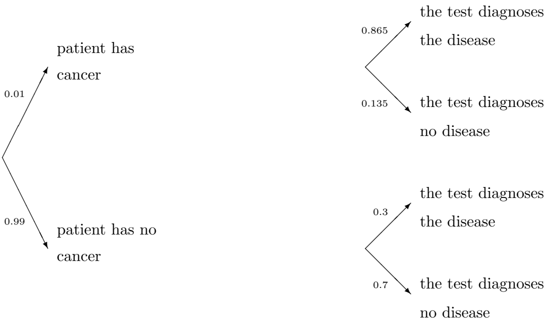
Bayes rule

Example: Extensive studies have shown that appr. 1.0% (a-priori probability) of all men between 40 and 50 have the cancer of prostate. A simple diagnostic test is the PSA test.

The PSA test has the property, that it makes the correct diagnosis with probability 0.7 for healthy patients (sensitivity) and with probability of 0.865 for ill patients.

What is the probability that a patient with negative (positive) test results is truly healthy (ill) (posteriori probability)?

Bayes rule II



Aim: P(patient is ill | test diagnoses the disease)

$$= \frac{\text{P(patient is ill AND the test diagnoses the disease)}}{\text{P(test diagnoses the disease)}}$$

$$= \frac{\text{P(test diagnoses the disease, if the patient is ill) P(patient is ill)}}{\text{P(test diagnoses the disease)}}$$

$$= \frac{0.865 \cdot 0.01}{\text{P(test diagnoses the disease)}}$$

Bayes rule III

Using the rule of total probability we obtain

$$\begin{aligned} &P(\text{the test diagnoses the disease}) \\ &\quad = \quad 0.865 \cdot 0.01 + 0.3 \cdot 0.99 = 0.297865. \end{aligned}$$

The probability that the patient is really ill, even if the test diagnosed it, equals

$$\frac{0.865 \cdot 0.01}{0.297865} \approx 0.029.$$

Independent events

Two events $A, B \subseteq \Omega$ are (stochastically) independent, if

$$P(A \cap B) = P(A) \cdot P(B)\,.$$

Note: If A and B are independent, then it holds that $P(B \mid A) = P(B)$ and $P(A \mid B) = P(A)$, since

$$P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B) = P(A) \cdot P(B)\,.$$

If two events are not (stochastically) independent, then we say, that they are (stochastically) dependent.

Random variables and distribution functions

A random variable (attribute) X is an appropriate mapping of the population Ω into the set S. In general $S \subset \mathbb{R}$.

Thus

$$X(\omega) = x,$$

where $\omega \in \Omega$ is a "state of the world" which causes the particular outcome $x \in S$ of the RV X.

If $S \subset \mathbb{R}^n$, then X is an n-dimensional random variable or a random vector.

The distribution function F_X of a random variable X is defined as

$$F_X(x) = P\Big(\{\omega \in \Omega : X(\omega) \leq x\}\Big)\ ,\ x \in \mathbb{R}.$$

Usually a short-hand form is used $F(x) = P(X \leq x)$ or $X \sim F$

- The distribution function is a mapping from a set of real numbers into the interval $[0, 1]$.
- The distribution function assigns to each event $\{X \leq x\}$ the corresponding probability.

Properties of distribution functions

Def: The distribution function F of a random variable X is a function with the following properties:

- $0 \leq F(x) \leq 1$ for all x
- $F(\infty) = \lim_{x \to \infty} F(x) = 1, \quad F(-\infty) = \lim_{x \to -\infty} F(x) = 0$
- $F(x)$ is monotone increasing in x
- F is right-side continuous

- Each function F which satisfies the above conditions is a distribution function.
- If there is a function, which satisfies the above properties, then we can construct a random variable and a probability measure, such that the distribution function of the random variable coincides with the given function.

Computation of the probabilities

The distribution function contains all the information relevant to a statistician. Using the distribution function we can compute all the probabilities related to the random variable.
Assuming $a < b$ it holds:

- $P(a < X \leq b) = F(b) - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a - 0)$
- $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$
- $P(X \geq a) = 1 - P(X < a) = 1 - F(a - 0)$.

where $F(a - 0)$ denotes the left-sided limit of F at a, i.e.
$F(a - 0) = \lim\limits_{\varepsilon \to 0} F(a - \varepsilon)$, with $\varepsilon > 0$.

Example: we toss a die till the first "6". Let X denote the number of tosses. Thus $\Omega = \mathbb{N}$.
Then it holds

$$f(i) = P(X = i) = \frac{1}{6}\left(\frac{5}{6}\right)^{i-1}.$$

$F(x) = P(\emptyset) = 0$ for $x < 1$.
For $n \in \mathbb{N}$, we obtain

$$\begin{aligned} F(n) &= P(X \leq n) = \sum_{i=1}^{n} f(i) \\ &= \,= \frac{1}{6}\sum_{i=0}^{n-1}\left(\frac{5}{6}\right)^{i} = 1 - \left(\frac{5}{6}\right)^{n}. \end{aligned}$$

For $n \leq x < n+1$ we obtain $F(x) = F(n)$.

- Probability of more than 10 tosses :

$$P(X > 10) = 1 - F(10) = \left(\frac{5}{6}\right)^{10} \approx 0.16$$

- Probability of more than 3 but less than 8 tosses:

$$\begin{aligned} P(3 < X < 8) \ &= \ P(3 < X \leq 7) \\ &= \ F(7) - F(3) = \left(\frac{5}{6}\right)^{3} - \left(\frac{5}{6}\right)^{7} \approx \ 0.3 \end{aligned}$$

Discrete random variables and discrete distribution functions

If X has a countable set of possible values, then X is a discrete random variable and F_X is a discrete distribution function.

- Let X take the values $x_1, x_2, \ldots$ and $p_i = P(X = x_i)$. Then

$$f(x) = \begin{cases} p_i & \text{if} \quad x = x_i \\ 0 & \text{if} \quad x \neq x_i \ \ \forall i \end{cases}$$

  is the probability function of X.

- Let $x_1 < x_2 < \ldots$. If $x_i \leq x < x_{i+1}$, then

$$F(x) \quad = \quad \sum_{v=1}^{i} f(x_v) = P(X = x_1) + \cdots + P(X = x_i)\,.$$

  Particularly $F(x) = 0$ for $x < x_1$, $F(x) = 1$ for $x > x_n$.

Examples for discrete distribution functions

a) Binomial distribution

- We repeat an experiment independently n times. The probability of observing the event A is $p = P(A)$.
- Define:

$$Z_i = \begin{cases} 1, & \text{if } A \text{ is observe in the } i\text{-th run} \\ 0, & \text{else} \end{cases}$$

- Then

$$X = \sum_{i=1}^{n} Z_i$$

  tells us how often A was observed in n experiments

- **Aim:** probability function of X, e.g. what is the probability that we observe A k times if we repeat the experiment n times?

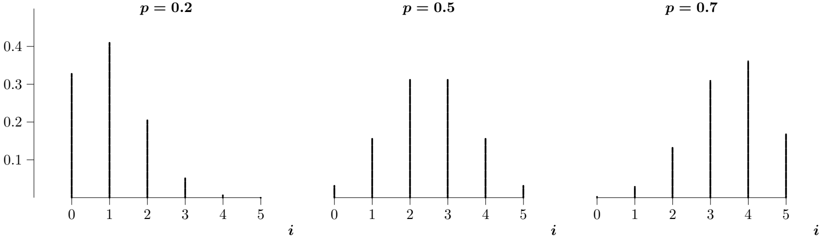
- Derivation:
  - $P(Z_i = 1) = P(A) = p$, $P(Z_i = 0) = P(\bar{A}) = 1 - p$
  - $\sum_{i=1}^{n} z_i = x$ corresponds to „x times event A and $n - x$ times event $\bar{A}$“
    probability (assuming independence): $p^x \cdot (1-p)^{n-x}$
  - But: the order is irrelevant! The number of possibilities: $\binom{n}{x}$

Probability function of the binomial distribution

$$f(x) = \begin{cases} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}, & \text{if } \ x \in \{0, 1, \ldots, n\} \\ 0, & \text{else} \end{cases}$$

- Short hand notation: $X \sim B(n; p)$
- $F(x)$ is determined using the general idea of CDFs for discrete RVs (e.g. $F(x) = \sum_{x_i \leq x} f(x_i)$)
- If $n = 1$, then we call this distribution Bernoulli distribution.

Probability function of the binomial distribution for $n = 5$



Example: cards

From a hand of 32 cards, three cards are drawn (with replacement). How likely is it to draw "hearts" twice?

$$
\begin{aligned}
X_i &= \begin{cases} 1, \text{ if the } i\text{-th card is "heart"} \\ 0, \text{ else} \end{cases} \\
X &= \sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 \\
X &\sim B(3; \tfrac{1}{4})
\end{aligned}
$$

Using the probability function

$$P(X = 2) = f(2) = \binom{3}{2} \cdot 0.25^2 \cdot 0.75^1 = 0.1406$$

Example: loans

From experience we know that a loan defaults with a probability of 0.1. What is the probability that exactly 48 out of 50 loans will not default?

$$
\begin{aligned}
P(X = 48) &= \binom{50}{48} \cdot 0.9^{48} \cdot 0.1^2 \\
&= 49 \cdot 25 \ \cdot\ 0.9^{48} \cdot 0.1^2 \\
&\approx 0.078
\end{aligned}
$$

b) Hypergeometric distribution

We consider a box with n balls. r of them are red, the rest are white. We draw k balls without replacement. Let the random variable X denote the number of drawn red balls.

$$P(X = i) \quad = \quad \frac{\binom{r}{i}\binom{n-r}{k-i}}{\binom{n}{k}}$$

This is the probability function of the hypergeometric distribution.

Example: from experience we know that the production of particular devices results in 20% of defective products. On a given day we produce 100 devices and randomly select arbitrary 10 of them. What is the probability, that the sample contains exactly 2 flaw products?

$$P(X = 2) \;=\; \frac{\binom{20}{2}\binom{80}{8}}{\binom{100}{10}} \approx 0.3181\,.$$

c) Poisson distribution

Let $X \sim B(n,p)$, i. e. $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$.

It is often the case that for the Binomial distribution n is large and p is small. Let p be a function of n, i.e. $p = p(n)$. If $\lim\limits_{n\to\infty} np(n) = \lambda > 0$, then

$$\lim_{n\to\infty} b(n,p(n))(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

We denote the limiting distribution by Poisson distribution and write $P(\lambda)$.

Example: A large insurance company computes the price of a vehicle insurance contract. On the basis of historical data the company assumes that the number of accidents X in a particular period follows the Poisson distribution with $\lambda = 3$.

Then

$$P(X = 2) = \exp(-3)\,\frac{3^2}{2!} \approx 0.224,$$

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \exp(-3) - 3\exp(-3) \approx 0.8009.$$

Note: The assumption of Poisson distribution is suitable here, because there are very many contracts and relatively few accidents.

Discrete distributions

	probability-function $f(m)$	Parameter space	Expected value $\mu = E(X)$	Variance $\sigma^2 = E\big([X-\mu]^2\big)$
Binomial $B(n,p)$	$\binom{n}{m} p^m (1-p)^{n-m}$ $m \in \{0,1,\ldots,n\}$	$0 < p < 1$ $n \in \{1,2,\ldots\}$	$n\,p$	$n\,p\,(1-p)$
Hyper-geometric $H(N,M,n)$	$\dfrac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$ $m \in \{m_{min}, m_{min}+1, \ldots, m_{max}\}$, $m_{min} := \max\{0, n-(N-M)\}$, $m_{max} := \min\{n, M\}$	$N \in \{1,2,\ldots\}$, $M \in \{0,1,\ldots,N\}$, $n \in \{1,2,\ldots N\}$	$n\,\dfrac{M}{N}$	$n\,\dfrac{M}{N}\,\dfrac{N-M}{N}\,\dfrac{N-n}{N-1}$ (for $N > 1$)
Poisson $P(\lambda)$	$\dfrac{\lambda^m}{m!}\,e^{-\lambda}$ $m \in \{0,1,\ldots\}$	$\lambda > 0$	λ	λ
Geometric $G(p)$	$p\,(1-p)^{m-1}$ $m \in \{1,2,\ldots\}$	$0 < p < 1$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$

Continuous random variables

X is a continuous random variable, if there exists a non-negative function f, such that:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \quad \text{for all} \quad x \in \mathbb{R}.$$

The function f is called the density (probability density) function of X.

Properties:

- $P(a < X \leq b) = \int_a^b f(t)\,dt$
- It holds $P(X = x) = 0$ for all $x \rightsquigarrow P(a < X < b) = P(a \leq X \leq b)$.
- If F is a continuous function, then $F' = f$.
- $\int_{-\infty}^{\infty} f(t)\,dt = 1$.
- The inverse CDF $F^{-1}(\beta)$ is called the quantile function.

$$F^{-1}(\beta) = inf\{x : F(x) > \beta\} \;\rightsquigarrow\; P(X \leq F^{-1}(\beta)) \geq \beta$$

Continuous distributions I

	Density f	Parameter space	Expected value $\mu = E(X)$	Variance $\sigma^2 = E(X-\mu)^2$
Uniform $U(a,b)$	$\frac{1}{b-a}$, $x \in [a,b]$	$-\infty < a < b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal $N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}$, $x \in \mathbb{R}$	$\mu \in \mathbb{R}, \sigma > 0$	μ	σ^2
Exponential $E(\lambda)$	$f(x) = \lambda\, e^{-\lambda x}$, $x \geq 0$	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
χ^2_n χ^2_n	$\frac{1}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\frac{n}{2}-1}\, e^{-\frac{x}{2}}$, $x > 0, n \in \mathbb{N}$		n	$2\,n$
t-distr. (Student) t_n	$\frac{\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}}{B(n/2,1/2)\sqrt{n}}$, $x \in \mathbb{R}, n \in \mathbb{N}$		$0 \quad (n > 1)$	$\frac{n}{n-2} \quad (n > 2)$

Continuous distributions II

	Dichte f	Parameter space	Expected value $\mu = E(X)$	Variance $\sigma^2 = E(X-\mu)^2$
F-distr. $F_{m,n}$	$\frac{(m/n)^{m/2}}{B(\frac{m}{2},\frac{n}{2})}\, x^{\frac{m}{2}-1} \left(1+\frac{m}{n}\,x\right)^{-\frac{m+n}{2}}$, $x \geq 0, m, n \in \mathbb{N}$		$\frac{n}{n-2}$ $(n > 2)$	$\frac{2\,n^2\,(m+n-2)}{m\,(n-2)^2\,(n-4)}$ $(n > 4)$
Gamma-distr.	$\frac{\lambda^r}{\Gamma(r)}\, x^{r-1}\, e^{-\lambda x}$, $x \geq 0$	$\lambda > 0,\ r > 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
Cauchy distr.	$\frac{1}{\pi\,\beta\,\{1+[(x-\alpha)/\beta]^2\}}$, $x \in \mathbb{R}$	$\beta > 0, \alpha \in \mathbb{R}$	-	-

with $\Gamma(x) = \int_0^\infty \exp(-t) t^{x-1} dt$ and $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.

Normal (Gaussian) distribution

The Normal distribution is the most important continuous distribution. It depends on 2 parameters, $\mu \in \mathbb{R}$ and $\sigma > 0$. Its density is given by

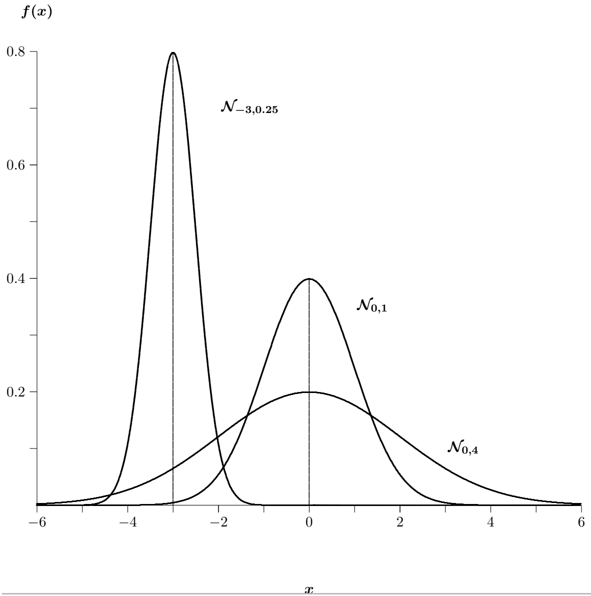
$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

For the distribution function of normal distribution we use the symbol $N(\mu, \sigma^2)$ or N_{μ,σ^2}.

Properties:

- f symmetric w.r.t. $x = \mu$, i.e. it holds $f(\mu + x) = f(\mu - x)$ for all x.
- The maximum of f is attained at μ.
- f has two turning points at $\mu \pm \sigma$.

Density functions of normal distribution for different parameters



Standard normal distribution

By standard normal distribution we denote the normal distribution with $\mu = 0$ and $\sigma = 1$. We write Φ for the distribution function and ϕ for the density.

Properties:

- Since $\phi(x) = \phi(-x)$, it follows that $\Phi(x) = 1 - \Phi(-x)$.
- If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \Phi$. This implies

$$F_X(x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

- If $X \sim \Phi$, then $\mu + \sigma\, X \sim N(\mu, \sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$, then $a\, X + b \sim N(a\,\mu + b, a^2\sigma^2)$.

Further properties

- Probability for deviation from the mean for at most c:

$$\begin{aligned}
P(\mu - c \leq X \leq \mu + c) &= F(\mu + c) - F(\mu - c) \\
&= \Phi\left(\tfrac{\mu + c - \mu}{\sigma}\right) - \Phi\left(\tfrac{\mu - c - \mu}{\sigma}\right) \\
&= \Phi\left(\tfrac{c}{\sigma}\right) - \Phi\left(-\tfrac{c}{\sigma}\right) \\
&= \Phi\left(\tfrac{c}{\sigma}\right) - \left[1 - \Phi\left(\tfrac{c}{\sigma}\right)\right] \\
&= 2 \cdot \Phi\left(\tfrac{c}{\sigma}\right) - 1
\end{aligned}$$

$k\sigma$-intervals $[\mu - k\sigma, \mu + k\sigma]$:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = 2\Phi(k) - 1 = \begin{cases} 0{,}683, & \text{for} \quad k = 1 \\ 0{,}954, & \text{for} \quad k = 2 \\ 0{,}997, & \text{for} \quad k = 3 \end{cases}$$

Exponential distribution

Exponential distribution arises in the analysis of life expectancy. Its density is given by

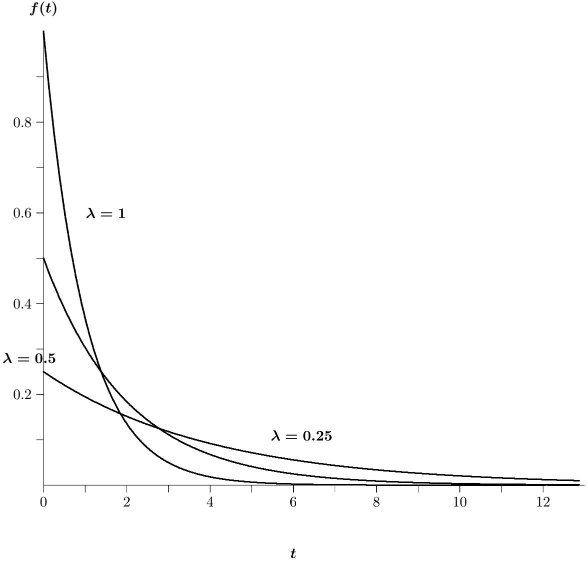
$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{for} \quad x \geq 0 \\ 0 & \text{for} \quad x < 0 \end{cases}$$

with $\lambda > 0$. Therefore $F(x) = 1 - \exp(-\lambda x)$. We write $E(\lambda)$.

Example: The life-span of TV-sets follows exponential distribution with $\lambda = 0.08$. What is the probability that the TV-set would have a life-span of more than 10 years?
It holds

$$P(X > 10) = 1 - F(10) = \exp(-0.08 \cdot 10) = \exp(-0.8) \approx \ldots.$$

Density of the exponential distribution for different parameters λ



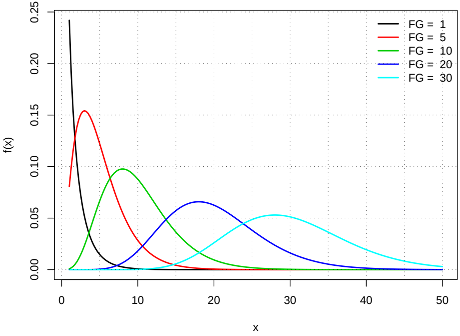
Chi-Square-Distribution (χ^2_f)

Assume that n RV's $Z_1, \dots, Z_n$

- are independent and
- follow standard normal distribution $Z_i \sim N(0;1)$ for $i = 1, \dots, n$

Then the sum of squares follows χ^2 distribution with n degrees of freedom

$$Z_1^2 + \ldots + Z_n^2 \sim \chi^2_n$$



t-distribution (Student-Distribution)

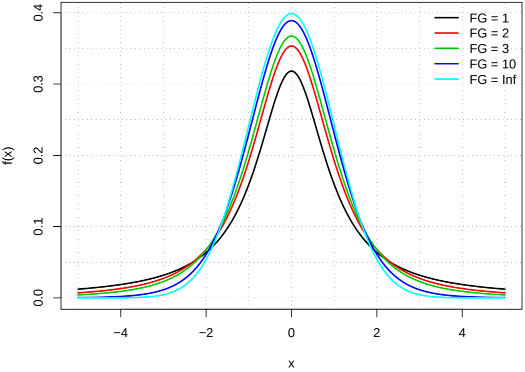
- Z follows the standard normal distribution: $Z \sim N(0,1)$
- Y is independent from Z and follows the chi-square distributed with df d: $Y \sim \chi^2_d$

Then the random variable

$$T = \frac{Z}{\sqrt{Y/d}}$$

follows the t distribution with degrees of freedom d.

- the density of the t-distribution is a symmetric bell-shaped curve
- the density of the t-distribution has heavier tails compared to the density of the normal distribution
- as $d \to \infty$ the density function of the t_d-distribution converges to the density of the standard normal distribution.



F-distribution

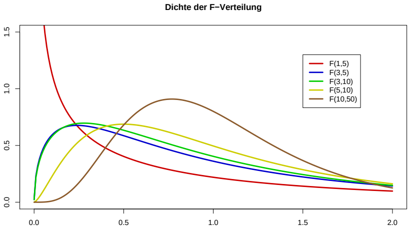
- Having two independent random variables Y_1 and Y_2, both following the chi-square-distributions with f_1 and f_2 df respectively:

$$Y_1 \sim \chi^2(d_1) \qquad Y_2 \sim \chi^2(d_2)$$

Then the distribution of the random variable

$$F = \frac{Y_1/d_1}{Y_2/d_2}$$

is called F-distribution with parameters d_1 and d_2



Characteristics of random variables

- In the descriptive statistics we discussed the location and dispersion measures of random samples.
- Here we discuss the measures of location and dispersion for random variables.
- The aim of the discussion is make statements about the center (central tendency) of the distribution.

The value x_{Med}, for which

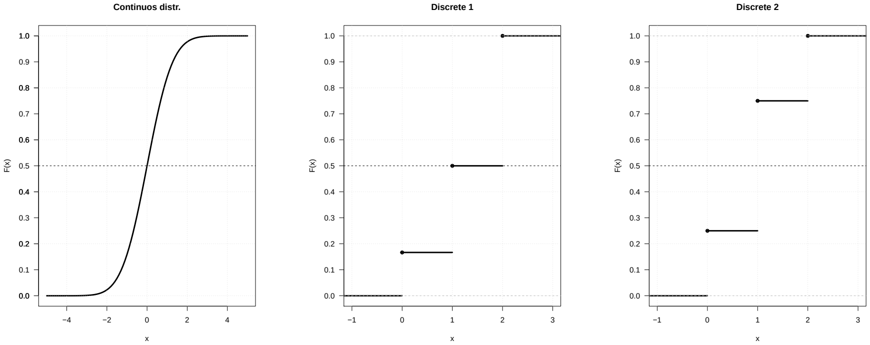
$$P(X \geq x_{\text{Med}}) = 1 - F(x_{\text{Med}}-) \geq \frac{1}{2} \quad \text{und} \quad P(X \leq x_{\text{Med}}) = F(x_{\text{Med}}) \geq \frac{1}{2}$$

is called **Median**.

- X is with at least 50% prob. larger or smaller than x_{Med}.
- **Note:** Median is not always unique.
- Every point for which $F(x) = \frac{1}{2}$ is a median.
- If there is no such point that $F(x) = \frac{1}{2}$ (for example, for discrete RV), then the median is the smallest such value that $F(x) > \frac{1}{2}$.

Example:

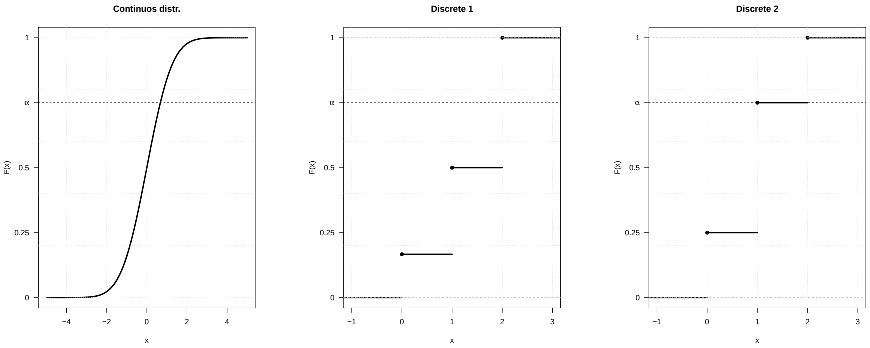
- Normal distribution: $x_{\text{Med}} = \mu$



If X is continuous, then there is for each $\alpha \in (0, 1)$ (at least) one x_α, such that $X \leq x_\alpha$ with prob. α.

The x-value, that satisfies the condition $F(x) = \alpha$, is **α-quantile** of the cdf F.

Interpretation: X is with at least $100 \cdot \alpha\%$ pron. less or equal than x_α and with at least $100 \cdot (1 - \alpha)\%$ prob. larger or equal than x_α.



Note

Beispiel:

- Quantiles of $X \sim N(0;1)$ are frequently denoted by z_α.

$$
\begin{aligned}
z_{0.975} &= 1.96 && \left(\Leftarrow \; P(X \leq z_{0.975}) = \Phi(z_{0.975}) = 0.975 \right) \\
z_{0.025} &= -z_{0.975} = -1.96 && \left(\Leftarrow \quad \text{symmetric distribution} \right)
\end{aligned}
$$

```
> qnorm(p = 0.975, mean = 0, sd = 1)
[1] 1.959964
> qnorm(p = 0.025)
[1] -1.959964
> qnorm(p = 0.025, lower.tail = FALSE)
[1] 1.959964
```

- Quantile of $Y \sim N(39; 4)$: the duration of the project that will not be exceeded with prob. of 97.5%

$$y_{0.975} = 39 + 2 \cdot z_{0.975} = 42.92$$

$$\left(\Leftarrow \; P(Y \leq y_{0.975}) = P\Big(\frac{Y - \mu}{\sigma} \leq \frac{y_{0.975} - \mu}{\sigma}\Big) = P(X \leq z_{0.975})\right)$$

$$y_{0.025} = 39 + 2 \cdot z_{0.025} = 35.08$$

```
qnorm(p = 0.975, mean = 39, sd = 2)
39 + 2*qnorm(p = 0.975, mean = 0, sd = 1)
qnorm(p = 0.025, mean = 39, sd = 2)
39 + 2*qnorm(p = 0.025)
```

Expectation (Mean)

Let X be a discrete RV und take values $x_1, x_2, \ldots$. Then the expectation of X (or equivalently of F) is given by

$$E(X) = \sum_i x_i \, P(X = x_i).$$

Examples:

- You win 4 Euro, if you throw “6” on a die and loose 1 Euro if you throw another number of pips. Then

$$E(X) \quad = \quad -1 \cdot \frac{5}{6} + 4 \cdot \frac{1}{6} = -\frac{1}{6}\,.$$

- Poisson distribution, i.e. $P(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}$ for $k \geq 0$

$$\begin{aligned}
E(X) &= \sum_{k=0}^{\infty} k \, P(X = k) = \sum_{k=0}^{\infty} k \, \frac{\lambda^k}{k!} \, e^{-\lambda} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda \, e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda \, e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda \, e^{-\lambda} \, e^{\lambda} = \lambda\,.
\end{aligned}$$

Let X be a continuous RV with the density function f. Then the expectation of X (or of F) is given by

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

The integral exists if $\int\limits_{-\infty}^{\infty} |x| f(x) dx < \infty$.

Example: number of clients arriving per unit of time (Exp with $\lambda = 1$)
It holds $f(x) = exp(-x)$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. Thus

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_{0}^{\infty} x\, e^{-x}\, dx \\ &= -x\, e^{-x}\Big|_0^{\infty} + \int_0^{\infty} e^{-x}\, dx = -e^{-x}\Big|_0^{\infty} = 1\,. \end{aligned}$$

```
> xfx <- function(x){x * exp(-x)}
> integrate(f = xfx, lower = 0, upper = Inf)
1 with absolute error < 6.4e-06
```

Note:

- If f is symmetric with respect to m, i.e.

$$f(m+x) = f(m-x) \quad \text{for all} \quad x$$

  then $E(X) = m$, if it exists.
- This implies that for $X \sim N(\mu, \sigma^2)$ it holds that $E(X) = \mu$.
- The expectation of the Cauchy distribution does not exist.

Rules for computing the expectations

Aim: computation of the expectation of $Y = g(X)$

If X is discrete, then it holds

$$E(Y) \quad = \quad \sum_i g(x_i)\, P(X = x_i)\,.$$

If X is continuous, then it holds

$$E(Y) \quad = \quad \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

(if the integral exists)

Examples

- If $Y = X^2$ and $X \sim \Phi$, then

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \\ &= -x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx = 1 \, . \end{aligned}$$

- Linear transformation (special case $g(X) = a + b \cdot X$)

$$E\big(a + b \cdot X\big) = a + b \cdot E(X)$$

$$\begin{aligned} E\big(a + b \cdot X\big) &= \int_{-\infty}^{\infty} (a + b \cdot x) \, f(x) \, dx \\ &= a \int_{-\infty}^{\infty} f(x) \, dx \, + \, b \int_{-\infty}^{\infty} x \, f(x) \, dx \\ &= a + b \cdot E(X) \end{aligned}$$

Sums and products of random variables

- Let $X_1, \ldots, X_n$ be random variables with existing expectations. Then it holds that

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)\,.$$

- If the RVs $X_1, \ldots, X_n$ are additionally independent, then it holds that

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)\,.$$

- Consider the portfolio consisting of n assets and its return R
  Let P_t denote the price of an asset at time point t. The simple return of the asset at time point t is given by

$$R_t = 100\,(P_t - P_{t-1})/P_{t-1}.$$

We consider now the returns of n assets at a given time point t. We denote them by $R_1, ..., R_n$. Let the relative fraction of the ith asset in the portfolio be given by w_i. This implies $\sum\limits_{i=1}^{n} w_i = 1$. Then the portfolio return equals $R = \sum_{i=1}^{n} w_i R_i$. Thus it follows:

$$E(R) = E(\sum_{i=1}^{n} w_i\, R_i) = \sum_{i=1}^{n} E(w_i\, R_i) = \sum_{i=1}^{n} w_i\, E(R_i).$$

If $E(R_i) = \mu$ for all $i = 1, \ldots, n$, then $E(R) = \mu$ too.

Dispersion measures of distribution functions

The dispersion (variability) measures for the distribution function measure the concentration of the probability around the center of symmetry.

The most popular dispersion measure is the variance. It is measured as the expected quadratic deviation from the expectation $\mu = E(X)$:

$$Var(X) = E\big([X - \mu]^2\big).$$

The variance exists if $E(X^2) < \infty$. Often it is denoted by $\sigma^2 = Var(X)$.

The quantity σ is called the standard deviation.

Let X be a discrete RV with the realizations $x_1, x_2, ...$ Then it holds that

$$Var(X) = \sum_i \left(x_i - \mu\right)^2 P(X = x_i).$$

If X is a continuous RV with the density function f, then

$$Var(X) = \int_{-\infty}^{\infty} \left(x - \mu\right)^2 f(x)\, dx.$$

Note:

- If $Var(X) = 0$, then $X = E(X)$. For continuous RVs it holds "*almost everywhere*".
- For all $a, b \in \mathbb{R}$ it holds that

$$Var(a\,X + b) = a^2\,Var(X)\,.$$

- If $X \sim N(\mu, \sigma^2)$, then X has the same distribution as $\mu + \sigma Y$ with $Y \sim \Phi$. This implies

$$Var(X) = Var(\mu + \sigma Y) = \sigma^2 Var(Y) = \sigma^2.$$

  Note: the parameter σ^2 of the normal distribution equals the variance!

- If the RVs $X_1, \ldots, X_n$ are independent (!) and the respective variances exist, then

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)\,.$$

Important statements about expectation and variance

-

If X is a RV with $E(X) = \mu$ and $Var(X) = \sigma^2$, then

$$Y = \frac{X - \mu}{\sigma} \qquad \text{(standardised ZV)}$$

has the expectation 0 and variance 1.

$$E(Y) = E\left(\frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma} \cdot E(X) - \frac{\mu}{\sigma} = \frac{1}{\sigma} \cdot \mu - \frac{\mu}{\sigma} = 0$$

$$Var(Y) = Var\left(\frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2} \cdot Var(X) = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$$

- Let $X_1, \ldots, X_n$ be independent with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, then

$$\bar{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \qquad \text{(sample mean)}$$

has the expectation $\boldsymbol{\mu}$ and the variance $\boldsymbol{\frac{\sigma^2}{n}}$.

For independent RVs $X_1, \ldots, X_n$ gilt, it holds

$$E\left(w_1 \cdot X_1 + \cdots + w_n \cdot X_n\right) = w_1 \cdot E(X_1) + \cdots + w_n \cdot E(X_n),$$

$$Var\left(w_1 \cdot X_1 + \cdots + w_n \cdot X_n\right) = w_1^2 \cdot Var(X_1) + \cdots + w_n^2 \cdot Var(X_n).$$

Characteristics of 2D distributions

The most popular measures of comovement are the covariance and the correlation.

- The covariance between X and Y is given by:

$$Cov(X, Y) = E\big([X - E(X)]\,[Y - E(Y)]\big).$$

The covariance exists if $E(|XY|) < \infty$.

- If $Var(X) > 0$ and $Var(Y) > 0$, then

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)\,Var(Y)}}$$

is called the correlation coefficient of Pearson.

- X und Y are uncorrelated if $Corr(X, Y) = 0$.

If X and Y are discrete random variables with realizations $x_1, x_2, \dots, y_1, y_2 \dots$, then

$$Cov(X,Y) = \sum_i \sum_j \big(x_i - E(X)\big)\big(y_j - E(Y)\big) \cdot P(X = x_i, Y = y_j)$$

$$= \sum_i \sum_j x_i\, y_j\, P(X = x_i, Y = y_j) - E(X)\, E(Y)\,.$$

If (X, Y) is a continuous random vector with the density function f, then

$$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(x - E(X)\big)\big(y - E(Y)\big) \cdot f(x,y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, y\, f(x,y)\, dx\, dy - E(X)\, E(Y) = E(XY) - E(X)E(Y)\,.$$

Rules for covariances and correlations:

- $Corr(a\,X + b, c\,Y + d) = Corr(X, Y)$ (if a and c have the same sign)
  (invariance w.r.t. to location and scale shifts)

- $|Corr(X, Y)| \leq 1$,
  $|Corr(X, Y)| = 1$, if X and Y lie on a straight line , i. e.
  $Y = \alpha + \beta\, X$.

- If X and Y are independent, then $Cov(X, Y) = 0$.
  The inverse statement is not correct in general!!!

- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$, since

$$Var(aX + bY) = E[(aX + bY^{-}E(aX + bY))^2]$$

$$= E(\,a(X - E(X)) + b(Y - E(Y))\,)^2$$

$$= a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

Two dimensional distribution functions

Let $X = (X_1, X_2)'$ (for example, the returns of Daimler and BMW, exchange rates Euro/$ and Euro/CHF). Then

$$F_X(x_1, x_2) = P\Big(\{\omega \in \Omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2\}\Big), \quad x_1, x_2 \in \mathbb{R}$$

is a (2–dimensional) distribution function of the random vector X. The short-hand notation is $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$.

$F_X(x_1, \infty)$ is the marginal distribution of X_1 and
$F_X(\infty, x_2)$ is the marginal distribution X_2.

Note: it holds

$$\begin{aligned} F_X(x_1, \infty) &= P(X_1 \leq x_1) =: F_1(x_1) \quad \text{and} \\ F_X(\infty, x_2) &= P(X_2 \leq x_2) =: F_2(x_2). \end{aligned}$$

Discrete and continuous random vectors

If the set of possible values of X is countable, then X is discrete and

$$f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

is the (joint) probability function of (X_1, X_2).
If X is continuous, then the distribution function F of X is given by

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2)\, dt_2\, dt_1, \quad x_1, x_2 \in \mathbb{R}$$

with $f(t_1, t_2) \geq 0$ for all t_1, t_2. The function f is a (2-dimensional) probability density function (pdf) of (X_1, X_2).

Note: If f is given, then the density function of f_1 (f_2) of X_1 (X_2) can be obtained in the following way

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2, \quad f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1.$$

Multivariate normal distribution

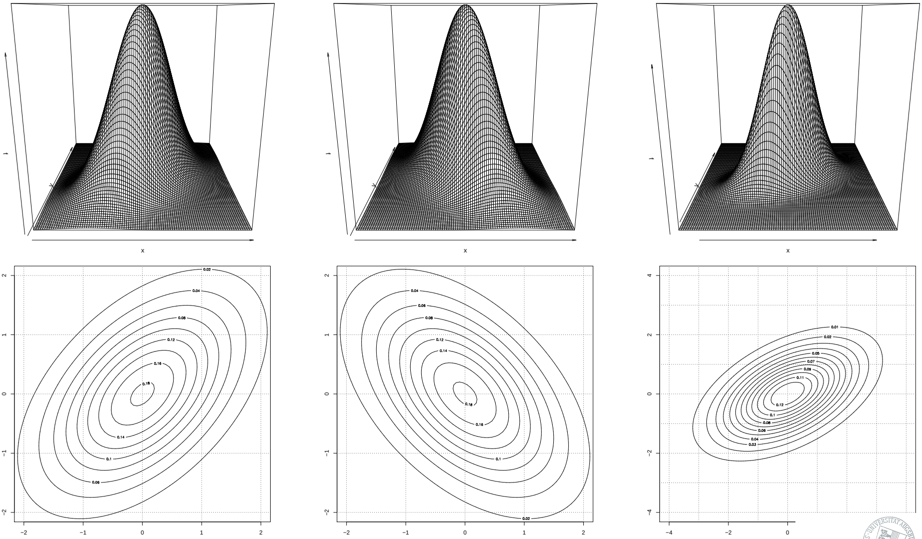
Def: The random vector $\boldsymbol{X}$ follows a p-dimensional multivariate normal distribution ($\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), if its density is given by

$$f(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} exp\Big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big].$$

Other multivariate distributions known in explicit form: t, Laplace, Wishart, and very few others.

Example (2-dimensional normal distribution)

$\sigma_1^2 = 1,\ \sigma_2^2 = 1,\ \rho = 0.5$ $\quad$ $\sigma_1^2 = 1,\ \sigma_2^2 = 1,\ \rho = -0.5$ $\quad$ $\sigma_1^2 = 2,\ \sigma_2^2 = 1,\ \rho = 0.5$



Multivariate RV

Def: $\boldsymbol{X}$ is a p-dimensional random vector, if the components $X_1, \ldots, X_p$ are scalar RVs.

The joint CDF is given by

$$F(\boldsymbol{x}) = P(X_1 \le x_1, \ldots, X_p \le x_p)$$

For a continuous random vector $\boldsymbol{X}$ it holds:

$$F(x_1, \ldots, x_{i-1}, -\infty, x_{i+1}, \ldots, x_p) = 0$$
$$F(+\infty, \ldots, +\infty) = 1$$
$$F(\boldsymbol{x}) = \int_{-\infty}^{x_p} \ldots \int_{-\infty}^{x_1} f(\boldsymbol{u}) d\boldsymbol{u}$$

Expectation and covariance matrix

Def: For a random vector $\boldsymbol{X}$ the expectation is defined by

$$E(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)' = (EX_1, \ldots, EX_p)'$$

and the covariance matrix by

$$Cov(\boldsymbol{X}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_p^2 \end{pmatrix}$$

$$= E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \ldots & Var(X_p) \end{pmatrix}$$

The correlation matrix is given by $\boldsymbol{R} = (\rho_{ij})_{i,j=1,\ldots p}$ with $\rho_{ii} = 1$ and $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$.

Rules

$$
\begin{aligned}
E(\boldsymbol{X}+\boldsymbol{Y}) &= E(\boldsymbol{X})+E(\boldsymbol{Y})\\
E(a\boldsymbol{X}+b) &= aE(\boldsymbol{X})+b\\
Cov(\boldsymbol{X}) &= E(\boldsymbol{X}\boldsymbol{X}')-\boldsymbol{\mu}\boldsymbol{\mu}'\\
Var(\boldsymbol{a}'\boldsymbol{X}) &= \boldsymbol{a}'Cov(\boldsymbol{X})\boldsymbol{a}=\sum_{i,j=1}^{p} a_i a_j \sigma_{ij}\\
Cov(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}) &= \boldsymbol{A}Cov(\boldsymbol{X})\boldsymbol{A}'
\end{aligned}
$$

$Cov(\boldsymbol{X})=\boldsymbol{\Sigma}$ and $\boldsymbol{R}$ is symmetric and positive semidefinite.

Let $\boldsymbol{Z}=(\boldsymbol{X}',\boldsymbol{Y}')'$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are p and q-dim. Then it holds

$$
\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{Z}} &= (\boldsymbol{\mu}'_{\boldsymbol{X}},\boldsymbol{\mu}'_{\boldsymbol{Y}})'\\
\boldsymbol{\Sigma}_{zz} &= \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{pmatrix} = \begin{pmatrix} Cov(\boldsymbol{X}) & Cov(\boldsymbol{X},\boldsymbol{Y})\\ Cov(\boldsymbol{Y},\boldsymbol{X}) & Cov(\boldsymbol{Y})\end{pmatrix}.
\end{aligned}
$$

Note: $\boldsymbol{\Sigma}_{xy}=\boldsymbol{\Sigma}_{yx}$.

Independent random vectors

up to now: independence of events

Recall: two events A_1 and A_2 are independent, if $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$. Then it holds $P(A_1 \mid A_2) = P(A_1)$.

Example: A_1 = „success of a therapy“, A_2 = „a drug was given“.

$X_1, \ldots, X_n$ are (stochastically) independent, if it holds for all $x_1, \ldots, x_n \in \mathbb{R}$

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i)\,.$$

Note:

- If $X_1, \ldots, X_n$ are independent and $g_1, \ldots, g_n$ are function, then $g_1(X_1), \ldots, g_n(X_n)$ are also independent.
- Let f be the probability function (density) of $(X_1, \ldots, X_n)$ and let f_i denote the probability function (density) of X_i.

$X_1, \ldots, X_n$ are independent if and only if

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

holds for all $x_1, \ldots, x_n \in \mathbb{R}$

Example: toss two symmetric dice: X_1 = number on the first die, X_2 = number on the second die

$$P(X_1 = i, X_2 = j) \;=\; 1/36 \;=\; P(X_1 = i)\, P(X_2 = j)$$

The random variables X_1 and X_2 are independent .

Marginal distributions

Let a $p+q$-dim. vector $\boldsymbol{Z}$ be partitioned into $\boldsymbol{Z} = (\boldsymbol{X}', \boldsymbol{Y}')'$, such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are p and q dim. respectively.

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x}) = F_{\boldsymbol{Z}}(x_1, \ldots, x_p, +\infty, \ldots, +\infty)$$

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{+\infty} f(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y}$$

Independency

Def: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent iff

$$F_{\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y}) = F_{\boldsymbol{X}}(\boldsymbol{x}) F_{\boldsymbol{Y}}(\boldsymbol{y}) \quad \text{or} \quad f_{\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}}(\boldsymbol{y}).$$

Conditional distributions

We consider the distribution of the explained variables $\boldsymbol{y}$ conditional on a set of explanatory variables $\boldsymbol{x}$.

$$f(\boldsymbol{y}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, \boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

$\rightsquigarrow$ The conditional expectation plays a key role in econometrics and a large portion of research is aimed to estimate it.

$$E(\boldsymbol{y}|\boldsymbol{x}) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \boldsymbol{y} f(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y}.$$

For $\boldsymbol{x} = f(\boldsymbol{w})$ it holds that:

$$\begin{aligned}
E(\boldsymbol{y}|\boldsymbol{x}) &= E[E(\boldsymbol{y}|\boldsymbol{w})|\boldsymbol{x}] \\
E(\boldsymbol{y}|\boldsymbol{x}) &= E[E(\boldsymbol{y}|\boldsymbol{x})|\boldsymbol{w}] \\
E(\boldsymbol{y}|\boldsymbol{x}) &= E[E(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})|\boldsymbol{x}] \\
E[E(\boldsymbol{y}|\boldsymbol{x})] &= E(\boldsymbol{y})
\end{aligned}$$

Transformation of random variables

Requirement: X_1 and X_2 are independent.

- If X_1 and X_2 are discrete, then

$$P(X_1 + X_2 = x) = \sum_{\substack{u,\,t \\ u+t=x}} P(X_1 = u, X_2 = t) \underset{indep.}{=} \sum_t P(X_1 = x - t) P(X_2 = t).$$

- Let f_1 and f_2 be the densities of X_1 and X_2. Then the density of $X_1 + X_2$ is given by

$$f_{X_1+X_2}(x) \underset{indep.}{=} \int_{-\infty}^{\infty} f_1(x-t)\, f_2(t)\, dt.$$

Implications:

If $X_1, \ldots, X_n$ are independent with

- $X_i \sim N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

- $X_i \sim N(\mu, \sigma^2)$, then

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

- $X_i \sim B(1, p)$, then

$$\sum_{i=1}^{n} X_i \sim B(n, p).$$

Lemma 1: Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$, where $\boldsymbol{A}$ is a $q \times p$-matrix with $rg(\boldsymbol{A}) = q \leq p$. Then $\boldsymbol{Y} \sim \mathcal{N}_q(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$.

Lemma 2: Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$, where $\Sigma^{-1/2}$ is the Cholesky decomposition of matrix $\boldsymbol{\Sigma}$. Then $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I})$.