Statistical Basics

SS 2016

Dr. Holger Fröhlich
Global Statistical Sciences
UCB Biosciences GmbH, Monheim

Content

1. Motivation, Descriptive Statistics
2. Random Variables and Statistical Distributions
3. Statistical Modeling and Inference

Motivation: Why Statistics?

- Increasing role of large scale –omics data in molecular biology and biomedicine, e.g.
  - Transcriptome
  - Methylome
  - Genomic variations
  - ...
- We want to detect patterns in these data and to make predictions!
- All measured data exhibit a certain variability

tumor	normal
5.423	1.234
6.239	0.283
8.288	1.488
4.999	1.048
5.399	0.599

- **Example:** Expression of gene p53 in tumor and normal tissues

Mean, Variance, Standard Deviation, Median, MAD

- **Question:** Is p53 differentially expressed?
- Compute some basic statistics:

$$\text{mean}: \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

mean (tumor) = 6.070
mean (normal) = 1.257

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

std(tumor) = 1.319
std(normal) = 0.486

- s^2 is called **variance**.
- Other basic statistics:
  - median: data point, for which 50% are larger and 50% smaller
    - median (tumor) = 5.423; median (normal) = 1.048
  - Median absolute deviation (mad):

$$mad = median(|x_i - median|)$$

mad (tumor) = 4.985
mad (normal) = 0.721

Median vs. Mean

- Assume the following data:

  mean = 4.992

  median = 5.423

- Comparison: mean without outlier:

  mean = 6.237

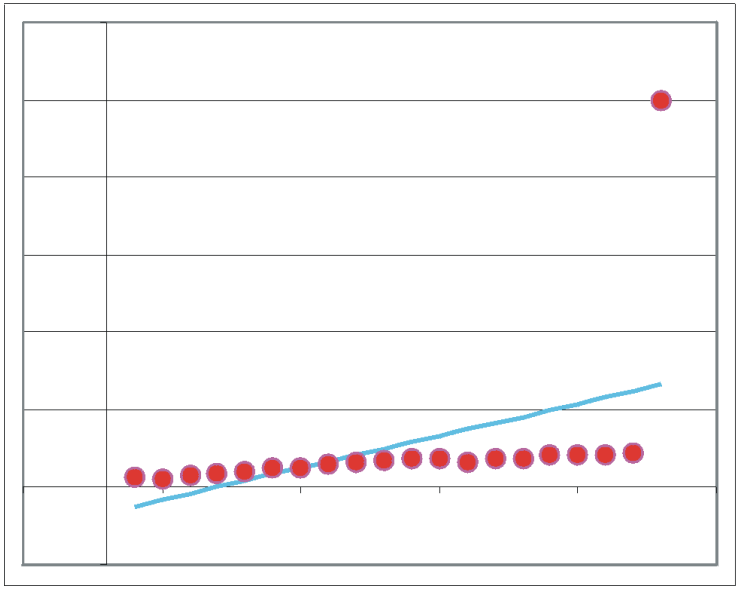
  median = 0.5*(5.423 + 6.239) = 5.831

tumor	normal
5.423	1.234
6.239	0.283
8.288	1.488
4.999	1.048
0.01	0.599

- Mean and std. dev. are heavily influenced by the outlier measurement!
- Median and MAD are more robust against outliers!

Outliers

- An **outlier** is a value, which does not fit to the rest of the data, i.e. does not meet our expectations
- Possible reasons:
  - Measurement errors
  - A real, but rare event
    - Ozone depletion was detected some years, before it was described, but was supposed to be a measurement error
- Outliers can spoil our estimations (e.g. mean and variance)
- Possible solution: **robust statistics**
  - Median instead of mean
  - MAD instead of variance



Quantiles

- The median is the 50% **quantile** of the data.
- p% quantile: Value in the dataset, for which only p% are smaller.
- In R:

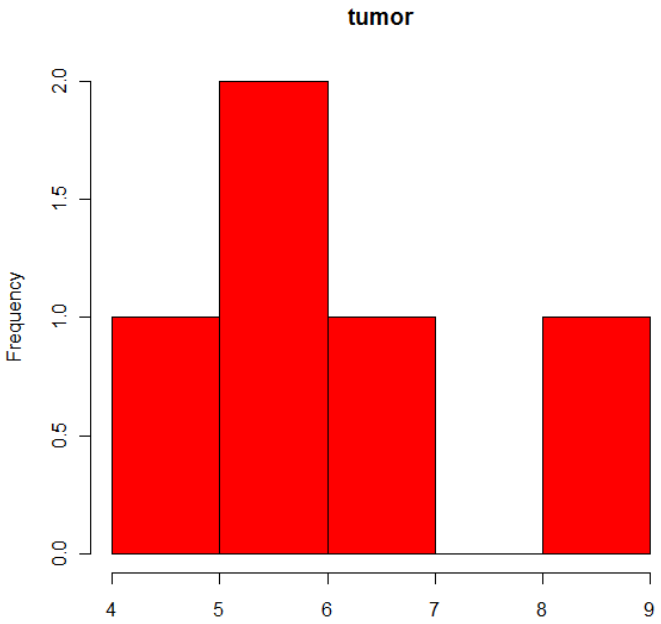
```
> x=c(5.423, 6.239, 8.288, 4.999, 5.399)
> quantile(x)
    0%    25%    50%    75%   100%
4.999  5.399  5.423  6.239  8.288
```

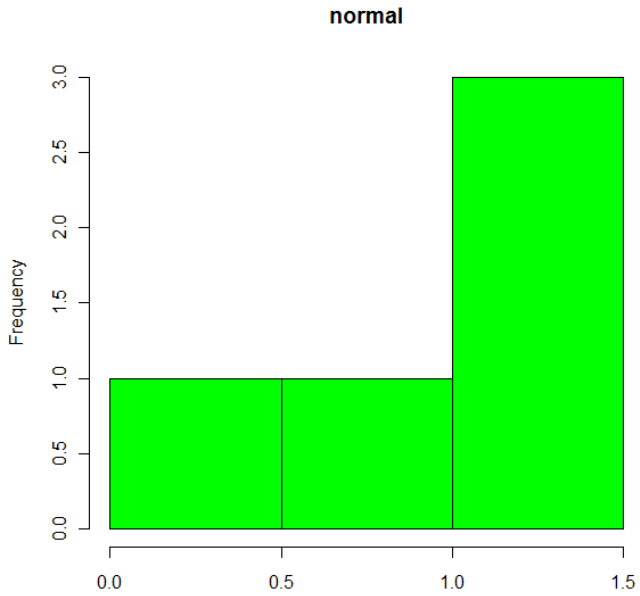
Lower quartile | median | Upper quartile

- The difference between the upper and lower quartile is called **interquartile range** (IQR)

Histograms: Are tumor and normal really different?

- A histogram is a partitioning of the input space into disjoint *bins*, shown on the x-axis
- On the y-axis of the plot the number of observations falling into a specific bin is drawn



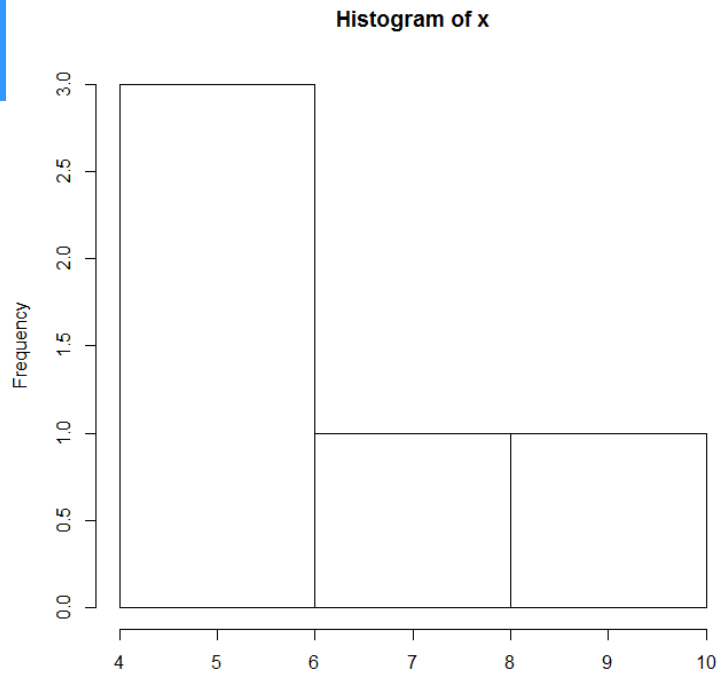


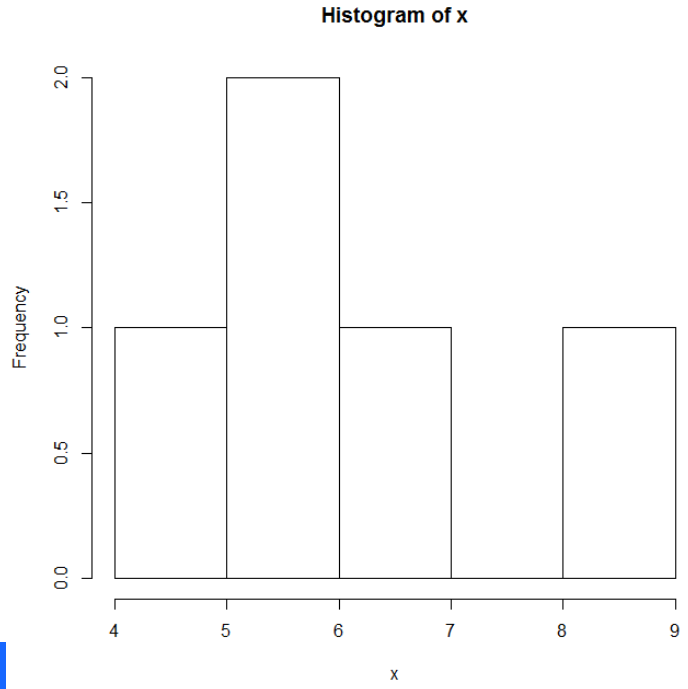

```
> x=c(5.423, 6.239,
   8.288, 4.999, 5.399)
> hist(x, col="red",
   main="tumor")
```

Binning

- Histogram depends on used binning
  - How many bins?
  - How to define break points?

```
> x=c(5.423, 6.239,
  8.288, 4.999, 5.399)
> hist(x, col="red",
  main="tumor",
  breaks="Scott")
> hist(x, col="red",
  main="tumor",
  breaks=4
```

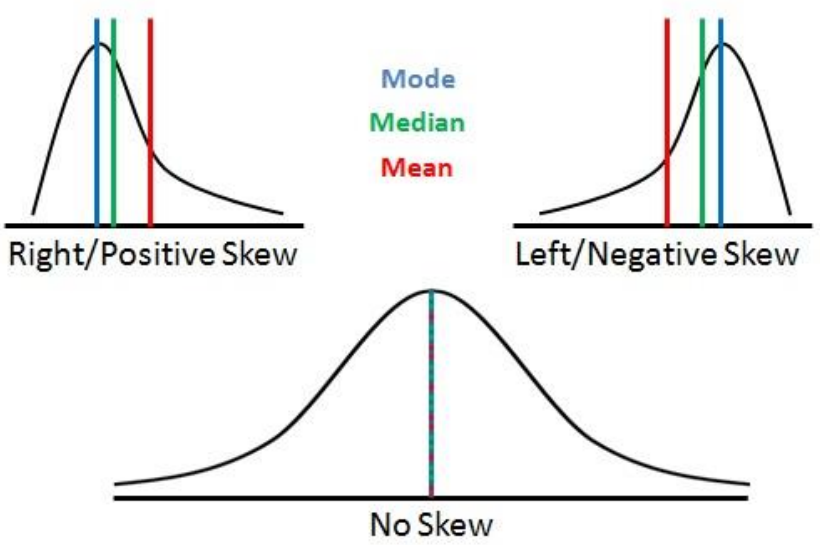




Mode

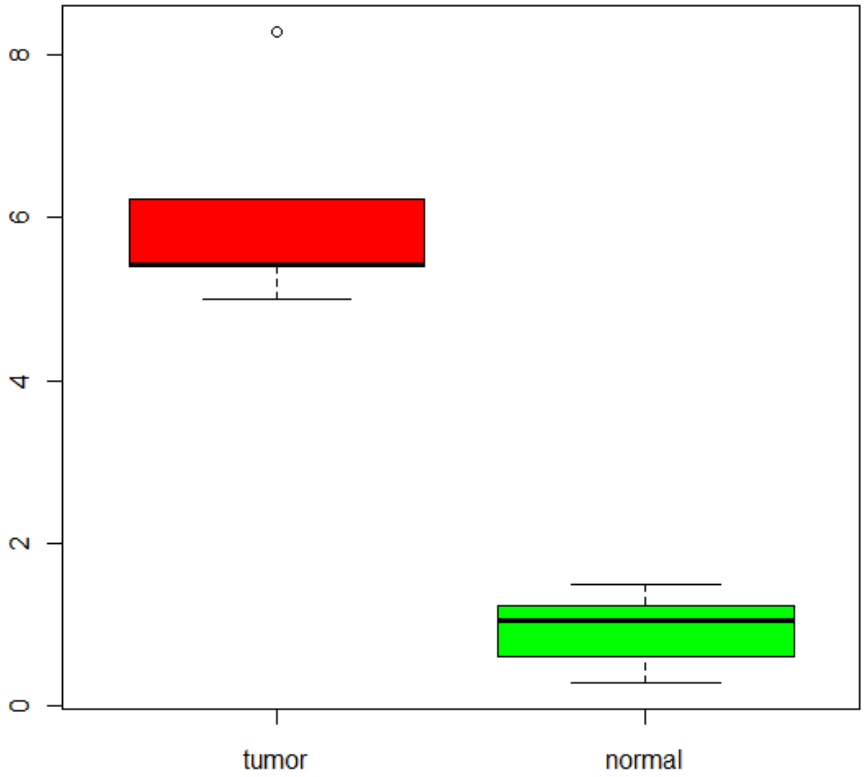
- The mode of a set of data is the most frequent value
- **Example**: The mode of values 8, 11, 5, 11, 14, 8, 11 is 11.

- It is hence identical to the maximum of observed frequencies in the histogram
- There can be exactly one mode (unimodel distribution) or several once (e.g. bimodel distribition)
- Mean, mode and median are not necessarily identical



Boxplots

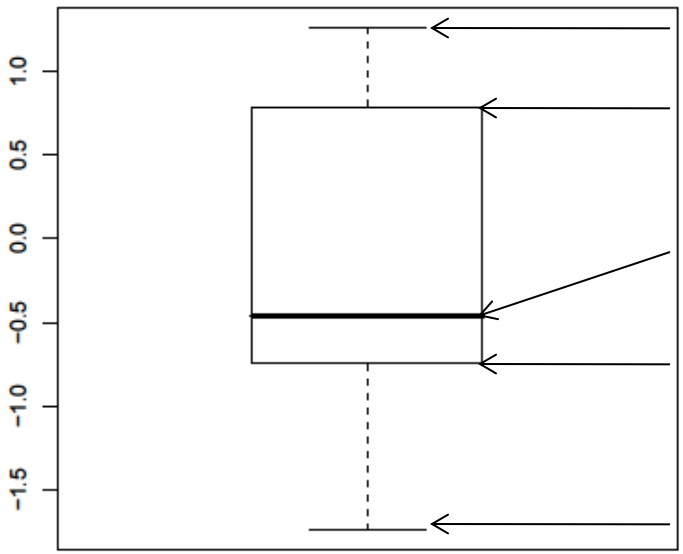
- Histograms offer a view on the distribution of the data
- Disadvantage: summary statistics are hard to see
- Solution: boxplots




```
>X = cbind(x,y)

>boxplot(X,col=c("red","green"),
names=c("tumor","normal"))
```

Boxplots

- Boxplot contains information about quantiles

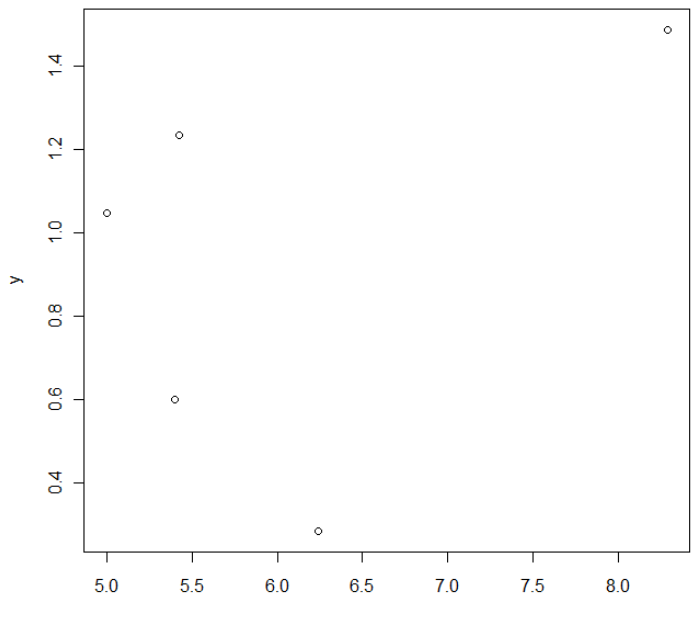


- Advantage of boxplots vs. barplots
  - View on whole distribution of values
  - Not only mean and SD
  - Mean may not be robust

- Barplots with error bars are not recommendable

Alternative visualization

- A plot of x against y (scatter plot)
- Are tumor and normal different?




```
> plot(x,y)
```

Correlation

- We can look for the **correlation** between tumor and normal samples.
- There are many correlation measures. Most popular one is **Pearson's correlation** (min: -1, max: 1)

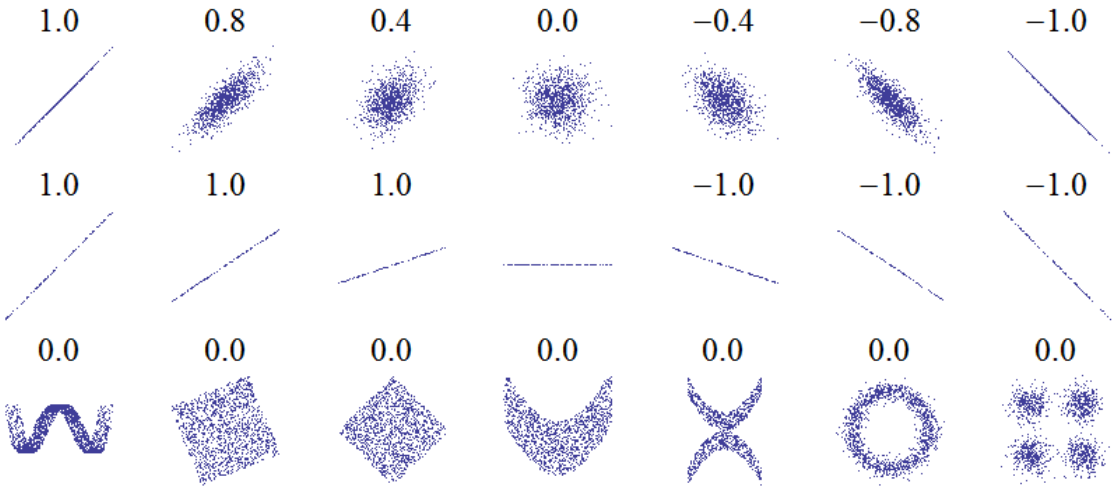
$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

- Correlation between tumor and normal in R:

```
> x=c(5.423, 6.239, 8.288, 4.999, 5.399)
> y=c(1.234, 0.283, 1.488, 1.048, 0.599)
> cor(x,y)
[1] 0.4002836
```

Correlation

- Pearson‘s correlation measures a **linear dependency** of x and y.
- It is scaled between -1 (perfect negative correlation) and 1 (perfect positive correlation)
- If the correlation is 0, x and y are independent / not at all correlated.
- **BUT**: The opposite is **NOT** true (because Pearson correlation only measures linear dependency)



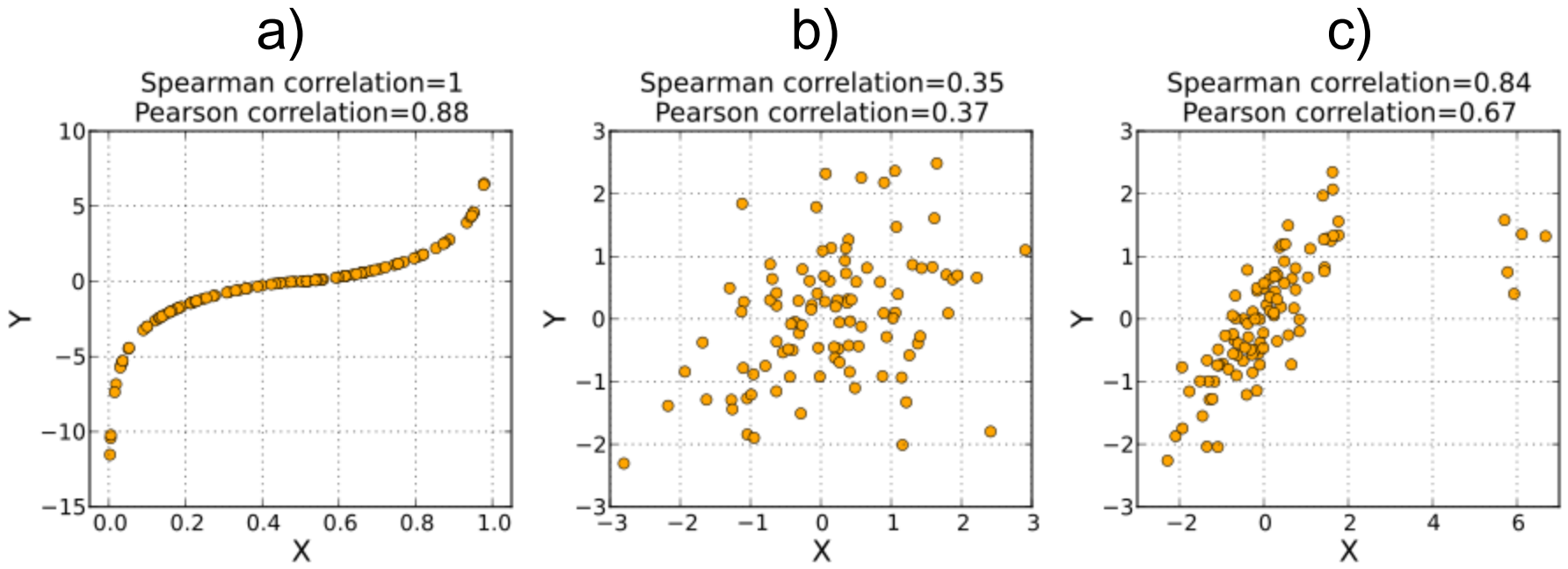
Rank Correlations

- ***Spearman's rank correlation coefficient*** and ***Kendall's rank correlation coefficient (τ)*** measure the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship.
- Spearman: convert raw measures into *ranks*, then compute Pearson correlation coefficient between ranks

Variable	Position	Rank
0.8	5	5
1.2	4	(4+3)/2
1.2	3	(4+3)/2
2.3	2	2
18	1	1

```
> cor(x,y,
method="spearman")

[1] 0.3

> cor(x,y,
method="kendall")

[1] 0.2
```

Examples: Spearman‘s rank vs. Pearson‘s correlation



- Spearman‘s rank correlation can capture non-linear dependencies (a)
- If the data data are roughly elliptically distributed and there are no strong outliers, Pearson‘s and Spearman‘s corr. are similiar (b)
- Spearman‘s rank corr. is less sensitivite to strong outliers at the tails of the distribution (c)

Covariance and correlation

- Let us reconsider the variance of variables x and y:

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

$$s_y^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- The covariance of x and y is defined as:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- ... which is related to Pearson‘s correlation:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

Summary: Descriptive Statistics

- Basic statistics for summarizing data
  - Central moments: Mean, median
  - Mode
  - Variability: variance, MAD, IQR
  - Quantiles
  - Covariance, correlation

- Statistical plots:
  - Scatter plot
  - Histogram
  - Boxplot

Going a bit more into depth

- A second measurement of p53:

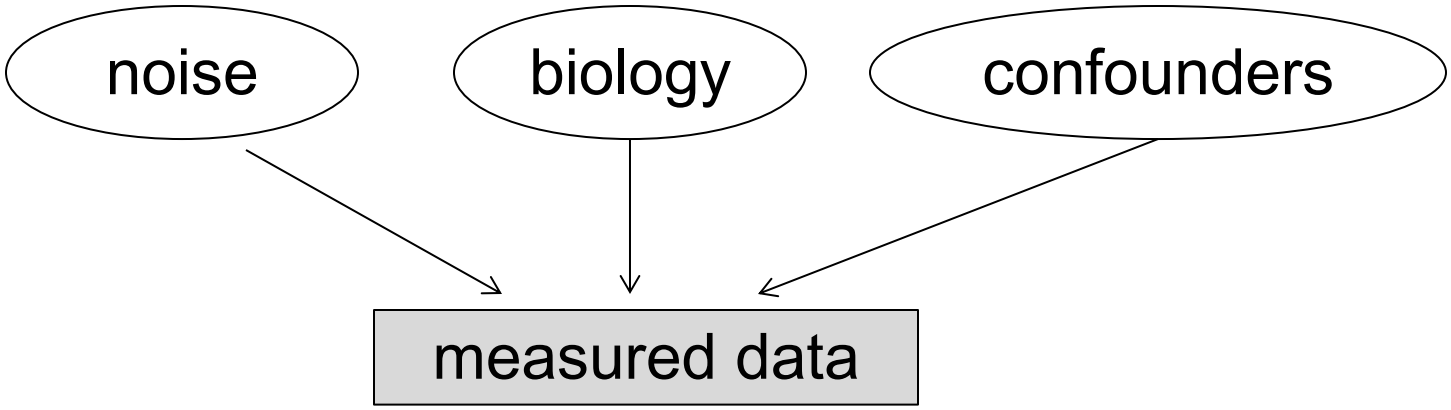
tumor	normal
4.299	2.398
6.382	1.209
7.397	2.087
5.399	1.287
6.388	1.297

mean (tumor) = 5.973
mean (normal) = 1.656
std(tumor) = 1.172
std(normal) = 1.297

- Obviously these numbers look differently than before.

Where does the data variability come from?

- Sources of data variability:
  - Noise: all data are impaired by measurement noise
  - Biological variability
  - Confounding factors (e.g. different measurement protocols, different handling of cells, etc.)
- In principle only systematic confounders can be controlled / removed, if they are known.



Necessity for a probabilistic treatment

- Assume we can control for confounding factors.
- How can we know, whether observed differences are due to noise are due to true biological differences?
- Idea: noise is something unsystematic, whereas true biological differences should be reproducible.
  - Take many measurements
  - Look, how likely it is that observed differences are just due noise, i.e. can be expected by chance
- ➔ **We need to compute probabilities**

Probability

- Probability is a way of expressing knowledge or belief that an event will occur or has occurred.
- In mathematics the concept has been given an exact meaning in probability theory.
  - ➔ computing with probabilities
- The word probability itself does not have a consistent direct interpretation
  - **Frequentist**: probability = limit of relative frequency of occurrence of an event
  - Example: throw a dye n times and observe the fraction that "6" occured

$$\Pr(\text{"6"}\,observed) = \lim_{n\to\infty} \frac{\text{number of times we observe 6}}{n}$$

Probability

- **Bayesian**: probability = individual's degree of belief
- Bayesian and frequentist probability are fundamentally different
  - Bayesians assign a probability to everything

  - Example: *a priori* Pr(„6“ observed) = 0.1
  - →Maybe we know that the dye is biased

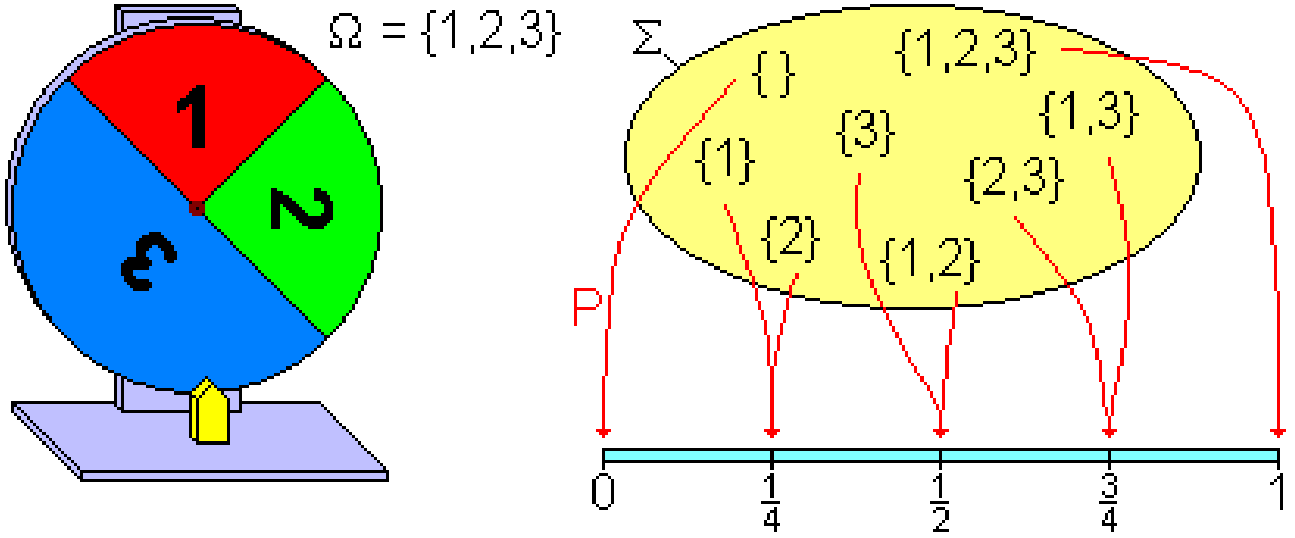
Mathematical probability theory

- A probability of an event A is represented by a real number in the range from 0 to 1 and written as P(A), p(A) or Pr(A).
- An impossible event has a probability of 0, and a certain event has a probability of 1.
  - **Note**: This does not say anything about where these probabilities come from
- Classical definition of probability (Laplace probability) for events:

$$\Pr(A) = \frac{\#\,\text{possibilities for A}}{\#\,\text{all possibilities}}$$

- Modern definition: Kolmogorov axioms
  - Basis: set theory
  - events = subsets of a sample space, to which probabilities are assigned
  - Probabilities = measures fullfilling three axioms

Idea of a probability space



- Rotating the wheel results in events (Σ), which are subsets of the set of all possible events (Ω).
- Each event in Σ has a defined probability, which depends on the area on the wheel.

Kolmogorov Axioms

Ω : sample space

Σ : event space (σ - algebra - set of subsets of Ω fullfilling certain conditions)

$P : \Sigma \to [0,1]$ probability measure

(Ω, Σ, P) = probability space

Nothing said where this measure comes from ➔ consistent with Bayesian as well as frequentist interpretation

First axiom : $0 \le P(A) \le 1 \; \forall A \in \Sigma$

Second axiom : $P(\Omega) = 1$ and $P(\{\}) = 0$

Third axiom : For each countable set of pairwise disjoint (i.e. $A_i \cap A_j = \{\} \; \forall i \neq j$) events

$A_1, A_2, ..., A_n : P(\bigcup_i A_i) = \sum_i P(A_i)$

Consequences :

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. $P(\Omega \setminus A) = 1 - P(A)$

Example

- Throwing two fair dies in parallel: What is the probability that either the first or the second one shows a 6?

A = first die shows a 6

B = second die shows a 6

P(A) = 1/6

P(B) = 1/6

$$P(A \cap B) = \frac{\#\,\text{times both dies show 6}}{\#\,\text{all possibilities}}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

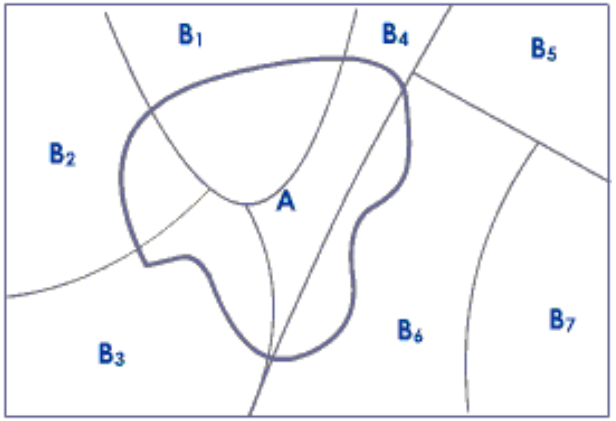
$$= 1/6 + 1/6 - 1/36$$

$$= 11/36$$

	1	2	3	4	5	6
1	1,1	1,2	1,3	...		
2	2,1	...				
3	...					
4						
5						
6	6,1					6,6

Law of Total Probability

- From the third Kolmogorov axiom it further follows (**law of total probability**):



$$\text{For pairwise disjoint } B_1,\ldots,B_n:$$

$$P(A)=\sum_i P(A\cap B_i)$$

- This is called **marginalization**.
- P(A) is called **marginal probability** of event A. P(A) is regardless of whether $B_1,\ldots,B_n$ occured or not.

Example

A = lawn is wet

B1 = It‘s raining

B2 = The sprinkler is on

$$P(A \cap B1) = 0.5$$

$$P(A \cap B2) = 0.2$$

$$P(A) = P(A \cap B1) + P(A \cap B2) = 0.7$$

Conditional probability and statistical independence

- **Conditional probability** is the probability of some event A, given the occurrence of some other event B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \longleftarrow \text{joint probability}$$

- Consequence: $P(A \cap B) = P(A \mid B)P(B)$

- Two random events A and B are **statistically independent**, if

$$P(A \cap B) = P(A)P(B)$$

  - Equivalently: $P(A \mid B) = P(A)$

- They are **conditionally independent** on Z, if

$$P(A \cap B \mid Z) = P(A \mid Z)P(B \mid Z)$$

Example

- Suppose 1% of a population suffer from a disease and the rest are well, i.e.

  P(ill) = 0.01

  P(well) = 0.99
- Suppose that when a screening test is applied to a person not having the disease, there is a 1% chance of getting a false positive result, i.e.

  P(positive | well) = 0.01

  P(negative | well) = 0.99
- Finally, suppose that when the test is applied to a person having the disease, there is a 1% chance of a false negative result, i.e.

  P(negative | ill) = 0.01

  P(positive | ill) = 0.99
- Probability of being well and **test negative**:

  P(well and negative) = P(well) * P(negative | well) = 99% * 99% = 98.01%

Example (cont'd)

- Probability of being ill and test positive:
  P(ill and positive) = P(ill) * P(positive | ill) = 1% * 99% = 0.99%

- Probability of being well and test positive:
  P(well and positive) = P(well) * P(positive | well) = 99% * 1% = 0.99%

- Probability of being ill and **test negative**:
  P(ill and negative) = P(ill) * P(negative | ill) = 1% * 1% = 0.01%

- Finally, the probability that an individual actually has the disease, given that the test result is positive:

$$P(ill \mid positive) = \frac{P(ill \cap positive)}{P(positive)} = \frac{P(ill \cap positive)}{P(positive \cap ill) + P(positive \cap well)}$$

$$= \frac{0.99\%}{0.99\% + 0.99\%} = 50\%$$

Rule of Bayes

- From the definition of conditional probability it directly follows the so-called **Bayes theorem**:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

- Given a partition of the event space in $B_1,\ldots,B_n$ according to the law of total probability we get:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_i P(A \mid B_i)P(B_i)}$$

$P(A \mid B_i)$ is called likelihood

$P(B_i)$ is called prior

The denominator is called partition function

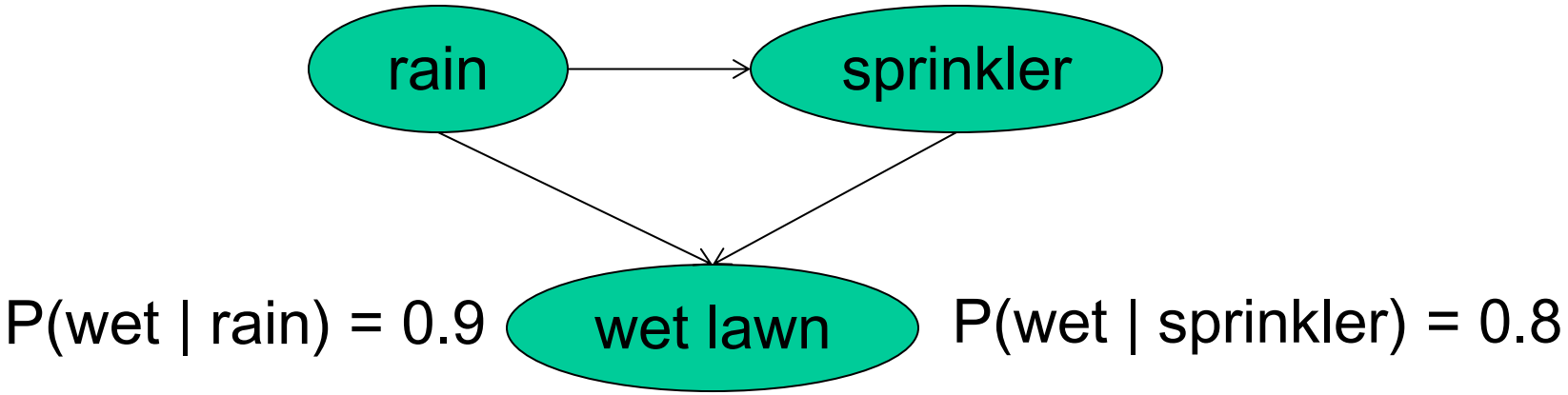
$P(B_i \mid A)$ is called posterior

Rule of Bayes

- The rule of Bayes provides a mechanism for **inference**.
  - **Likelihood & prior → posterior**

Example:

P(sprinkler | rain) = 0.1
P(sprinkler | sun) = 0.9



P(wet | rain) = 0.9

P(wet | sprinkler) = 0.8

- The weather forecast tells us: P(rain) = P(sun) = 0.5
- Given we see the lawn is wet. What is the probability that it rains? In other words what is P(rain | wet lawn)?

Example (cont.'d)

$$P(rain \mid wet) = \frac{P(wet \mid rain)P(rain)}{P(wet)}$$

$$= \frac{P(wet \mid rain)P(rain)}{P(wet \mid rain)P(rain) + P(wet \mid sprinkler)P(sprinkler)}$$

$$P(sprinkler) = P(sprinkler \mid rain)P(rain) + P(sprinkler \mid sun)P(sun)$$

$$= 0.1 * 0.5 + 0.9 * 0.5$$

$$= 0.5$$

Hence :

$$P(rain \mid wet) = \frac{0.9 * 0.5}{0.9 * 0.5 + 0.8 * 0.5} \approx 0.53$$

There is only a 53% chance that it is raining, if the lawn is wet, even if the conditional probability P(wet | rain) is 90%.

Rule of Bayes in Broader Context

- The rule of Bayes is a tool for **probabilistic** inference
- Comparison: modus ponens in propositional logic

$$\text{modus ponens:}$$

$$\frac{A \Rightarrow B, A}{B}$$

$$\text{Bayes:}$$

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

- Rule of Bayes is the backbone of the Bayesian approach to
  - Inference / inferential statistics (next)
  - Predictive modeling / machine learning (later)

Random Variables

- A random variable describes the possible outcome of a random experiment (e.g. throwing a coin).
- Thus, a random variable can be thought of as a function mapping the sample space of a random process to the real numbers.
- **Example**: For a coin toss, the possible events are heads or tails. The number of tails appearing in one fair coin toss can be described using the following random variable:

$$X = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}$$

- Formal definition: A random variable X is a function Ω → **R**

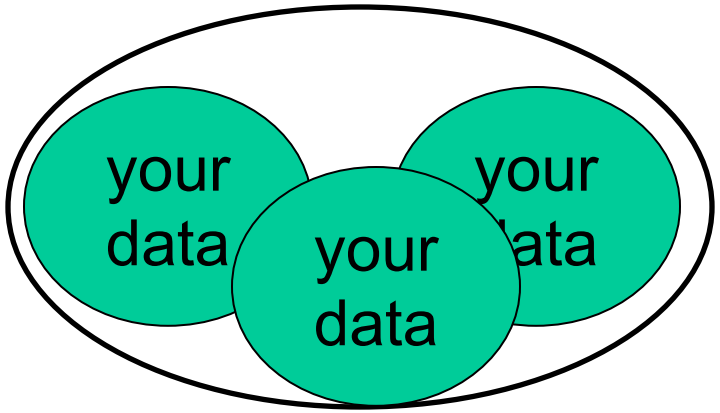
- Pr(X = 1) means “probability to observe tails” = “probability that random variable X takes value 1”

Random Variables

- A possible realization (*sample*) of X: 1, 0, 1, 0, 0, 1, 1
- Another possible sample of X: 0, 0, 0, 0, 1, 1, 0, 0

- Observed data are realizations of a random variable, i.e. measured data are sampled from a random variable

- **Example:** SNPs measured in 100 MS patients are drawn from a random variable describing SNP data



population

whole universe of data

random variable

Example

- Rolling of a die:

Sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$

$X: \Omega \rightarrow \mathbf{R}$ with $X(\omega) = \omega$ for all $\omega \in \Omega$, i.e.:

$X(1) = 1$

$X(2) = 2$

...

- We can assign a probability to each of these events, e.g. for a fair die

$\Pr(X = \omega) = 1/6$

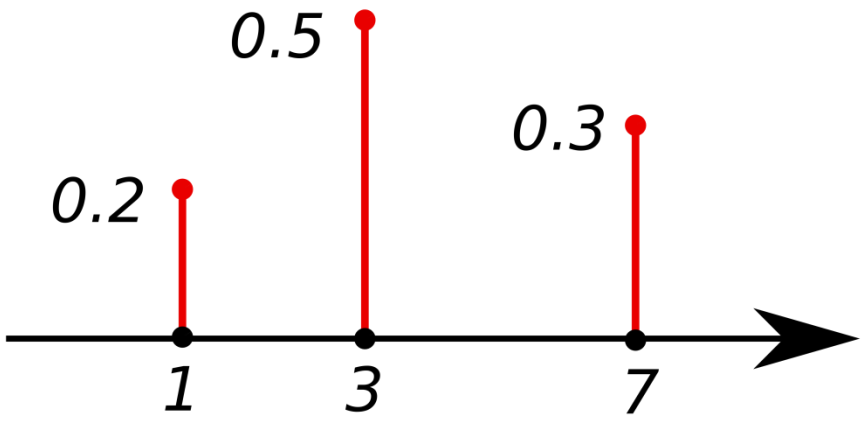
- X is an example of discrete random variable, because X can only take on a finite set of values (e.g. 1, 2, ..., 6)

Discrete Random Variables

- Let X be a discrete random variable that can take on m different values in the set D = $\{v_1, v_2, \ldots, v_m\}$
- A **probability mass function** (pmf) is a function P that gives the probability that a discrete random variable is exactly equal to some value.
- P must satisfy:

$$P(v) \geq 0$$

$$\sum_{v \in D} P(v) = 1$$



Summary statistics for random variables

- **Descriptive statistics:** Describe features (mean, standard dev., ...) of our observed data

- **BUT**: data is just a sample taken from a larger population and is subject to randomness

- How to describe the whole population in terms of summary statistics?
- Idea: define mean, standard deviation, etc. for random variables

Expected value

- The **expected value**, mean or average of a random variable X is defined by

$$\mu = E[X] = \sum_{v \in \text{all possible values of } X} vP(v)$$

- **Example:** Expected value for X = rolling a fair die:

$$E[X] = 1*\frac{1}{6} + 2*\frac{1}{6} + 3*\frac{1}{6} + 4*\frac{1}{6} + 5*\frac{1}{6} + 6*\frac{1}{6} = 3.5$$

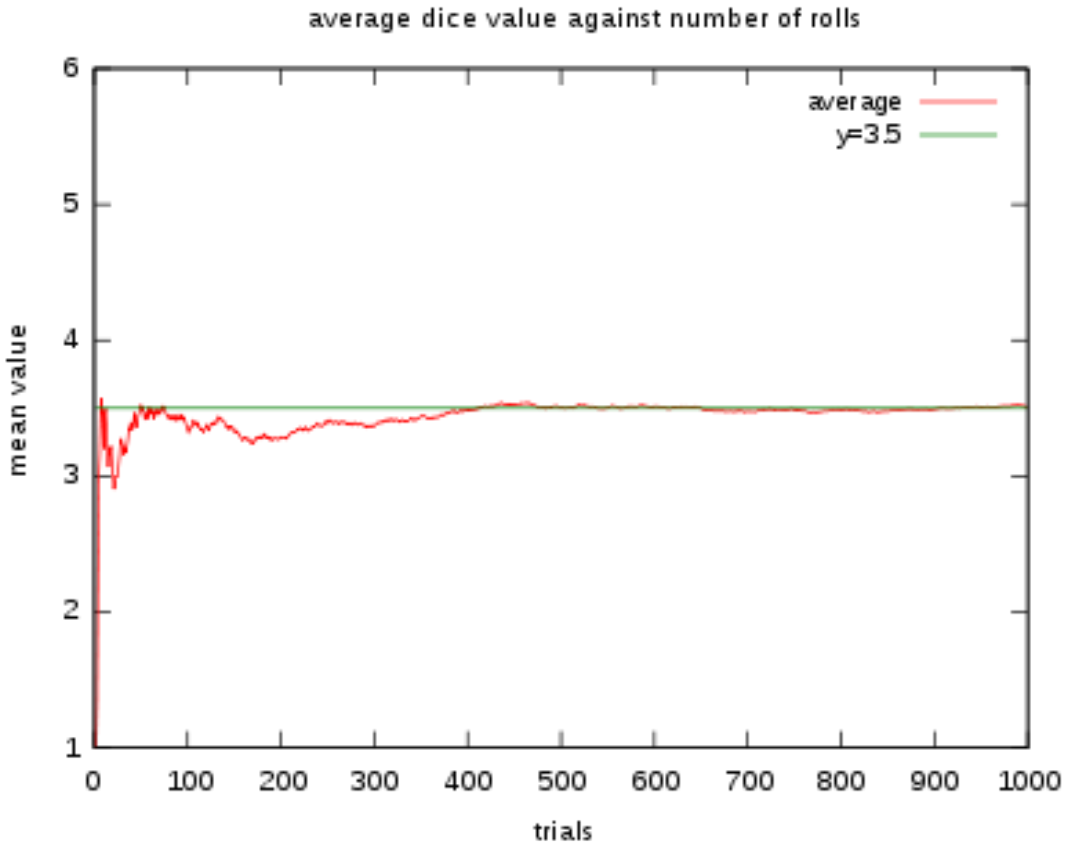
- Remark: The empirical average over n die roles converges to the expectation value (almost surely)

Convergence of the mean towards the expectation value

- Empirical mean is **not** necessarily identical to the expectation value
- But: It converges towards it.

- Empirical mean is an unbiased estimator of the expectation value

$$E[\bar{x}_n - \mu] = 0$$



Almost sure convergence:

$$\Pr\left(\lim_{n\to\infty} \bar{x}_n = E[X]\right) = 1$$

$\bar{x}_n$ = mean after n die rolls

Variance

- The **variance** of random variable X is defined as the mean squared difference from its expectation value:

$$\sigma^2 = Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

- Comparison: sample or empirical variance (= **descriptive statistics**):

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

- **Example:** Variance of X = rolling a fair die

$$Var[X] = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - 3.5^2$$

$$\approx 2.9$$

Covariance

- The **covariance** of two random variables X and Y is defined as:

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

- Comparison: sample or empirical covariance (= **descriptive statistics**):

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- **Example:** Two dice thrown at once. X = score of die 1, Y = sum of dice scores

$$E[Y] = \frac{1}{36}*2 + \frac{2}{36}*3 + \frac{3}{36}*4 + \frac{4}{36}*5 + \frac{5}{36}*6$$
$$+ \frac{6}{36}*7 + \frac{5}{36}*8 + \frac{4}{36}*9 + \frac{3}{36}*10 + \frac{2}{36}*11 + \frac{12}{36}*12$$
$$= 7$$

Y

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

Covariance (cont.‘d)

Y

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

$$A := X - E[X] = 1 - 3.5, 2 - 3.5, \ldots 6 - 3.5$$

$$B := Y - E[Y] = Y - 7$$

$$Cov[X, Y] = E[A * B] = \frac{1}{36}\left((2-7)*(1-3.5) + (2-7)*(2-3.5) + \ldots + (12-7)*(6-3.5)\right)$$

$$= \frac{35}{12} \approx 2.92$$

Properties of Expectation Value and Variance

- The expected value is a linear operator:

$$E[aX + bY] = aE[X] + bE[Y]$$

- The variance is **not** a linear operator:

$$Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] + 2abCov[X, Y]$$

In general :

$$Var\left[\sum_i X_i\right] = \sum_{i,j} Cov[X_i, X_j]$$

Please note:

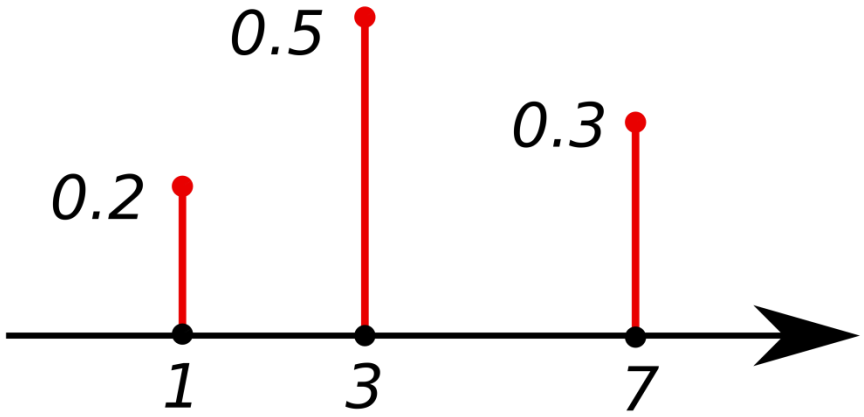
$$Cov[X, X] = Var[X]$$

Once again: Discrete Random Variables

- Let X be a discrete random variable that can take on m different values in the set D = $\{v_1, v_2, \ldots, v_m\}$
- A **probability mass function** (pmf) is a function P that gives the probability that a discrete random variable is exactly equal to some value.
- P must satisfy:

$$P(v) \geq 0$$

$$\sum_{v \in D} P(v) = 1$$



Expectation value:

$$E[X] = \sum_{v \in D} vP(v)$$

Variance:

$$Var[X] = \sum_{v \in D} P(v)(v - E[v])^2$$

Law of total probability:

$$P(X) = \sum_{b \in B} P(X \mid B = b)P(B = b)$$

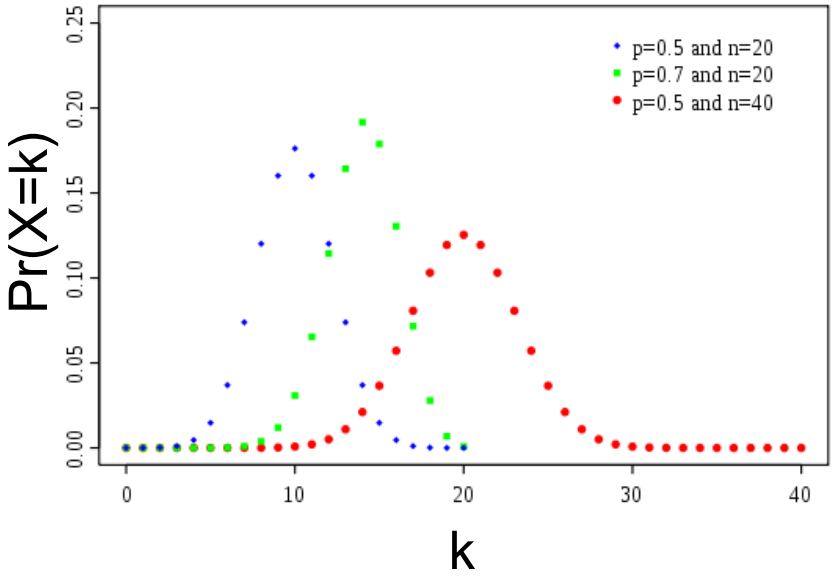
PMFs: binomial distribution

- **Idea**: number of successes k in a sequence of n independent yes/no experiments, each of which yields success with probability p.
- **Example:** Throwing (biased) a coin n times and counting the number k of heads
- Probability mass function:

$$Bin(n \mid k, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

- Expected value: np
- Variance: n*p(1 - p)



Application

- Compute the probability to observe 4 heads in a series of 10 tosses with an unfair coin, which produces heads with probability 0.3:

$$\Pr(X = 4 \mid n = 10, p = 0.3) = \binom{10}{4} 0.3^4 (1 - 0.3)^{10-4}$$

$$\approx 0.2$$

- Expectation value and variance for unfair coin:

E[X | n=10, p] = n*p = 10 * 0.3 = 3

➔ we expect 3 heads in 10 tosses

Var[X | n=10, p] = n*p*(1-p) = 10*0.3 *0.7 = 2.1

➔ The expected squared difference to the expecation value is 2.1 head counts.

PMFs: multinomial distribution

- Generalization of the binomial distribution to a situation, where each trial has 1 out of k possible outcomes (multivariate distribution)
  - Example: Throw a (biased) die 10 times and compute the probability for the result 1: 2 times; 2: 0 times; 3: 3 times; 4: 0 times; 5: 1 time; 6: 4 times
- Success probabilities: **p**= (p_1, p_2, ..., p_k) with $\sum p_i = 1$
- n = number of trials
- Probability mass function for observing non-negative integers $(x_1,...,x_k)$:

$$Mult(x_1,...,x_k \mid n; p_1,...,p_k) = \begin{cases} \dfrac{n!}{\prod_i x_i!} \prod_i p_i^{x_i} & \text{if } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise} \end{cases}$$

- Expected value: $E[x_i] = n*p_i$
- Variance: $n*p_i(1 - p_i)$
- Covariance: $Cov(x_i, x_j) = -np_i p_j$ for $i \neq j$

Application

- Assuming a fair die, what is the probability to observe the above mentioned result?

$$\Pr(X = (2,0,3,0,1,4)^T \mid n = 10,\ p = (1/6,1/6,1/6,1/6,1/6,1/6)^T)$$

$$= \frac{10!}{(2!0!3!0!1!4!)} \frac{1}{6}^2 \frac{1}{6}^0 \frac{1}{6}^3 \frac{1}{6}^0 \frac{1}{6} \frac{1}{6}^4$$

$$\approx 0.000208381$$

- Expected number of 6s in 10 die throws:

E[X=6| n=10, p] = 10 * 1/6 = 1.67

- Covariance between seeing 1 and 6 in 10 die throws:

Cov[X=1, X=6 | n=10, p] = -10 * 1/6 * 1/6 = -10/36 = -0.278

➔ The expected number of 1 and 6 is not independent!

Continous Random Variables

- Up to now: random variable X can only take on certain discrete values
- Now: X can take values in the continuum.
- **Problem:** P(X = x) does not make sense (is always zero)
- Instead: introduce a **probability density function** (PDF) *p* according to:

$$P(a \le x \le b) = \int_a^b p(x)dx$$

with

$$p(x) \ge 0 \; \forall x \text{ and } \int_{-\infty}^{\infty} p(x)dx = 1$$

- Most definitions and formulas for discrete random variables carry over to continous random variable with sums replaced by integrals.
- Expectation value: $E[X] = \int_{-\infty}^{\infty} x p(x)dx$

  Variance: $Var[X] = \int_{-\infty}^{\infty} (x - E[x])^2 \, p(x)dx$

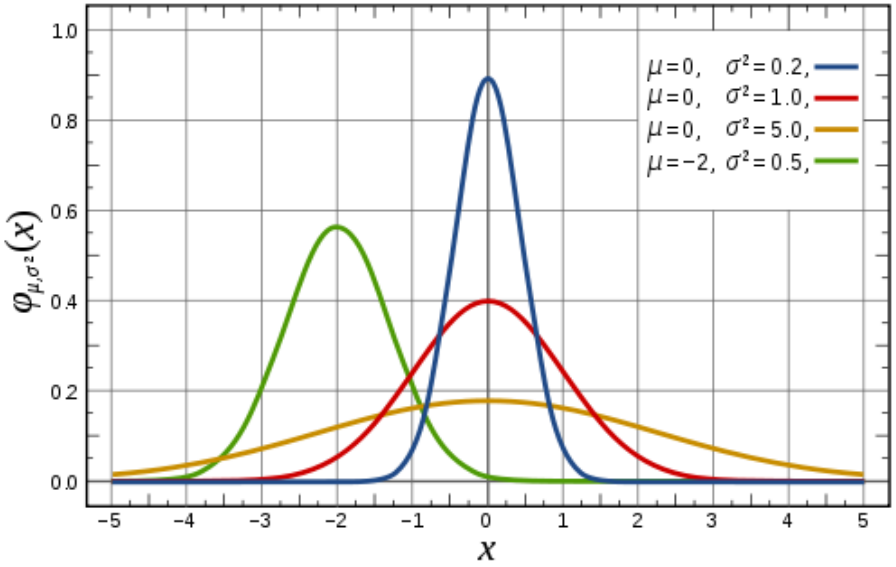
  Law of total density: $p(x) = \int_{-\infty}^{\infty} p(x \mid b) \, p(b)db$

PDFs: Normal / Gaussian distribution

- Most important statistical distribution
- Normal distribution is often used to describe any variable that tends to cluster around the expected value.
- **Example**: Heights of adult males in the US are roughly normally distributed, with a mean of about 1.8 m. Most men have a height close to the mean, though a small number of outliers have a height significantly above or below the mean.
- PDF:

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

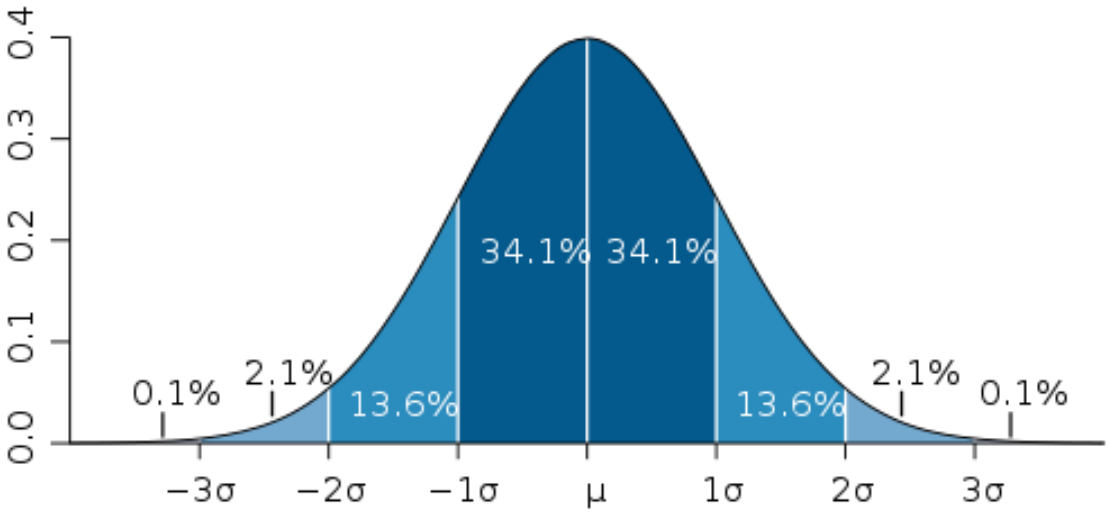
- PDF with μ=0 and σ=1 is called **standard normal distribution**



$$\text{expectation value} : \mu$$

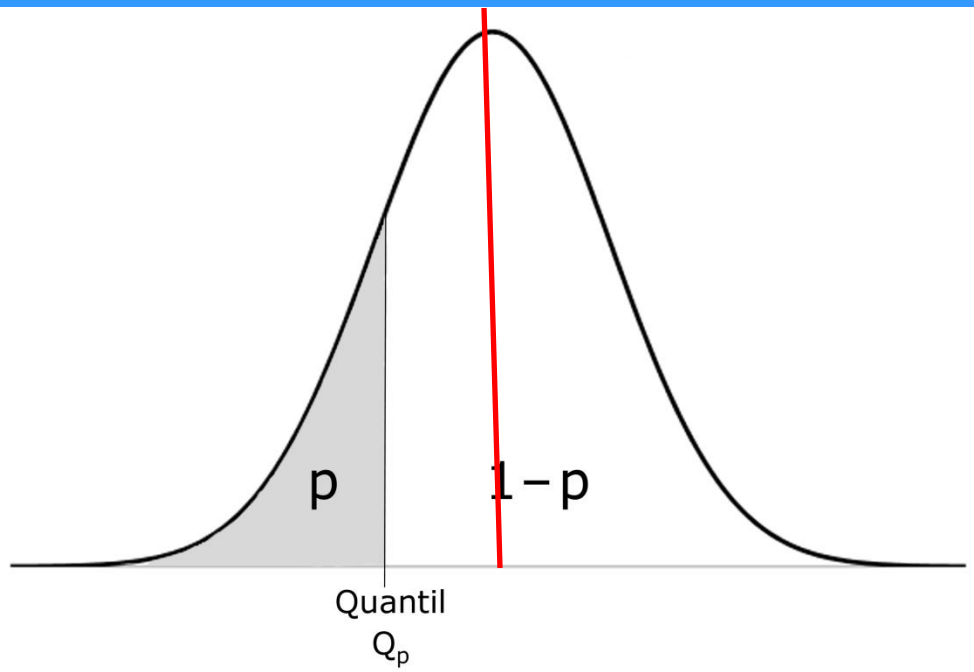
$$\text{variance} : \sigma^2$$

Gaussian distribution

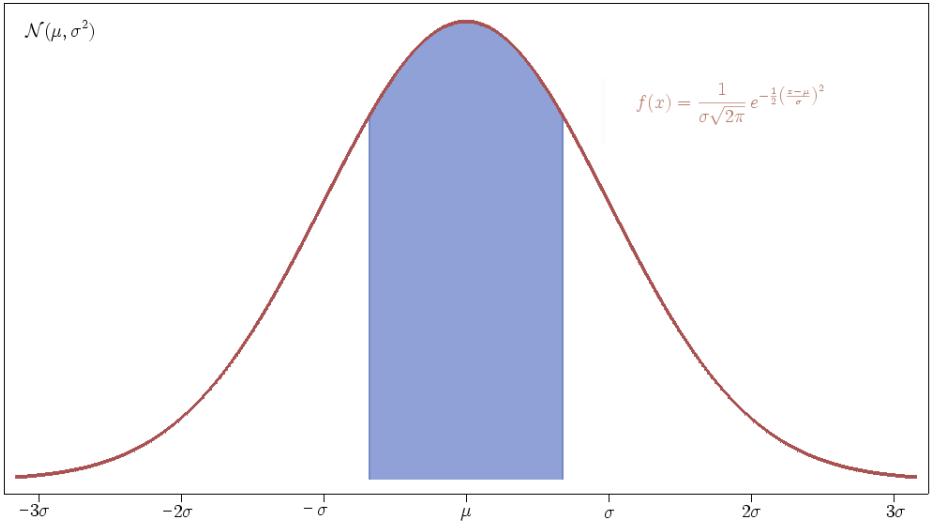


Quantiles and IQR in normal distributions

- p-quantile: p*100% lie within the gray area
- The 50% quantile (=median) equals the expectation value for symmetric PDFs (like the normal)
- It equals the mode of the PDF



- Interquartile range (IQR): difference between the 75% and 25% quantile



Application

- We measure the expression of the p53 gene in tumor cells.
- Suppose we know the data is $N(6, 1)$ distributed, in which range will our measurements lie with 95% probability?

- Answer: We know that 95% of the distribution is covered by the range $[\mu - 1.96\sigma; \mu + 1.96\sigma]$ (here: $\mu = 6, \sigma = 1$). Hence, any randomly drawn value from X will lie with ~95% probability in the range [4; 8].
- The range $[\mu - 1.96\sigma; \mu + 1.96\sigma]$ is the so-called *95% confidence interval* for normally distributed data.

Bayesian Reasoning with PDFs

- We measure the gene expression of p53 for 5 times.
  - Mean = 6; SD = 1

- Suppose we know that our data was drawn from a normal distribution with $\sigma = 1$ (i.e. $N(\mu, 1)$).
- Suppose we also know that $\mu \sim N(5, 0.1)$
  - Parameter μ follows a normal distribution for itself
- What can we infer about the distribution of μ from our data?

- Principle way: Bayes law

$$p(\mu \mid data) = \frac{p(data \mid \mu)\, p(\mu)}{\int p(data \mid \mu)\, p(\mu) d\mu}$$

Application: Bayesian Reasoning with PDFs

- Generally, if $\mu \sim N(\mu_0, \sigma_0)$ it can be shown that

$$\mu \mid data \sim N\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \right)$$

- Here:

$$\mu \mid data \sim N\left(\frac{\frac{5}{0.1} + \frac{5*6}{1}}{\frac{1}{0.1} + \frac{5}{1}}, \frac{1}{\frac{1}{0.1} + \frac{5}{1}} \right)$$

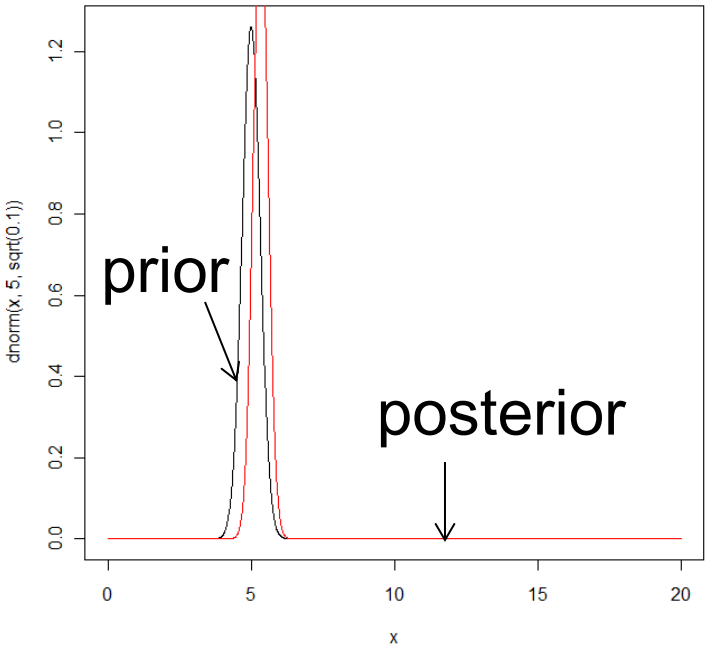
$$= N(5.33, 0.07)$$

- Hence: posterior expectation for μ: 5.33

Application: Bayesian Reasoning with PDFs

- Let‘s plot the prior and the posterior distributions:

```
> x=seq(0, 20, by=0.01)

> plot(x,dnorm(x, 5,
sqrt(0.1), type="l")

> lines(x,dnorm(x, 5.33,
sqrt(0.07), col="red")
```

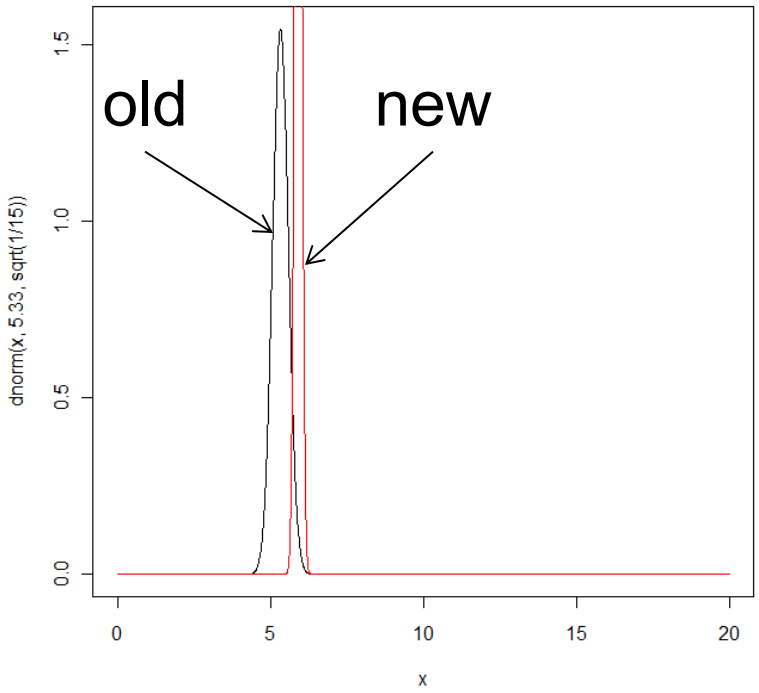


- Now suppose we have measured 100 times.

Application: Bayesian Reasoning with PDFs

$$\mu \mid data \sim N\left(\frac{\frac{5}{0.1}+\frac{100*6}{1}}{\frac{1}{0.1}+\frac{100}{1}}, \frac{1}{\frac{1}{0.1}+\frac{100}{1}}\right)$$

$$= N(5.91, 0.01)$$



- Expectation value of this distribution: 5.91
- 20 times more data provides a fold change of 5.91 / 5.33 = 1.11 of the posterior expectation.
- The variance (the uncertainty!) of the posterior shrinks due to the availability of more data

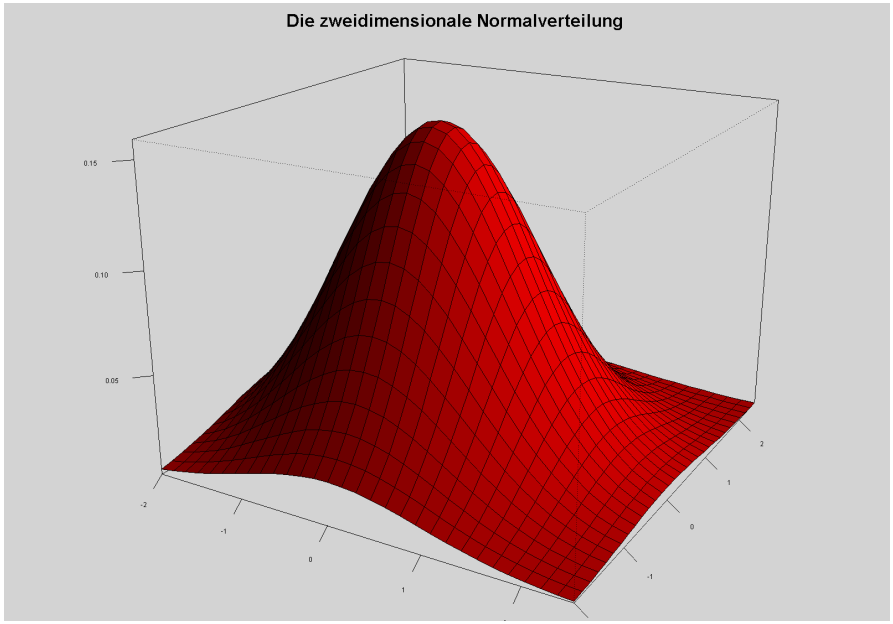
PDFs: Multivariate Normal Distribution

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi \mid \boldsymbol{\Sigma} \mid)^{d/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Mean vector

Coveriance matrix (d x d)

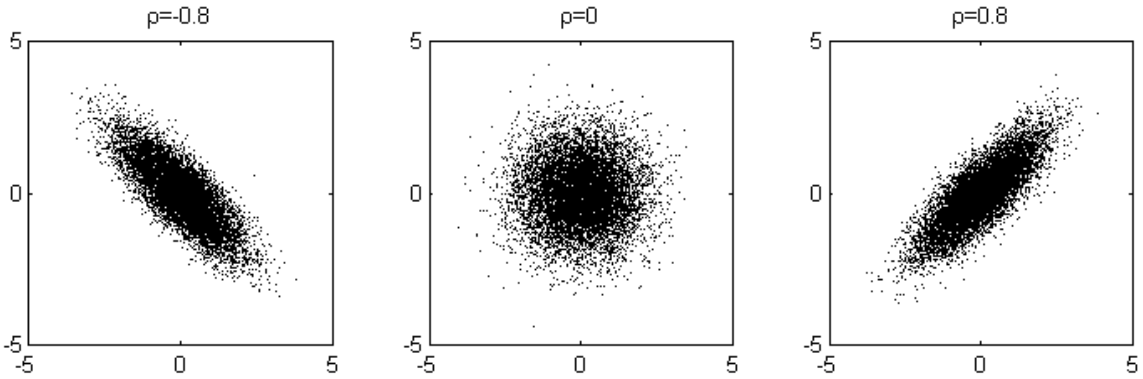
(d dimensions)



Covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{pmatrix}$$

Note: matrix is symmetric!



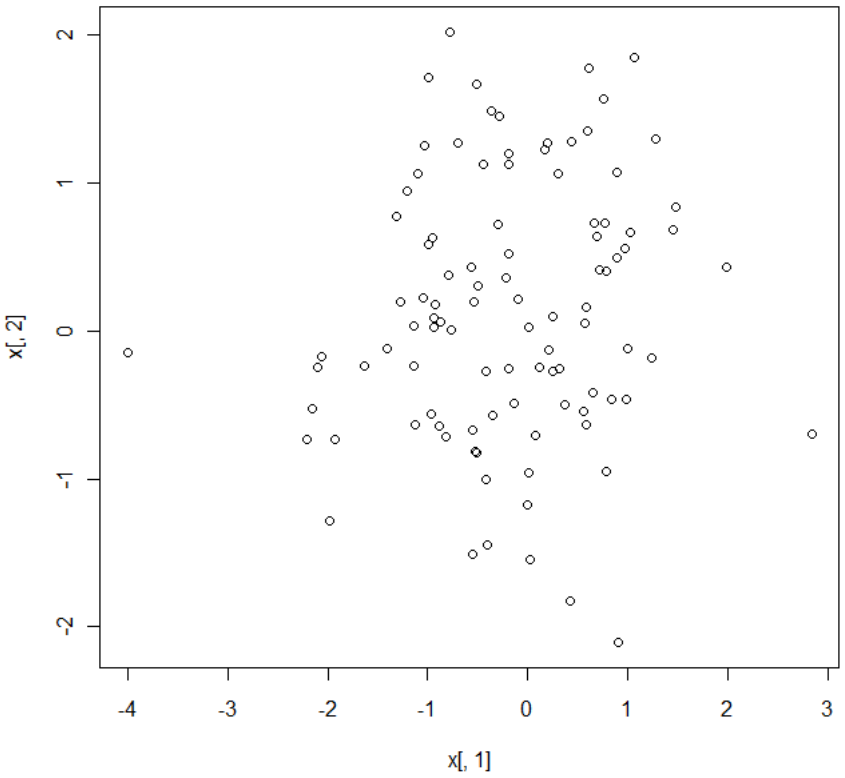
Example

- Draw 100 random values from a MVN distribution with

$$\mu = (0,0)^T$$

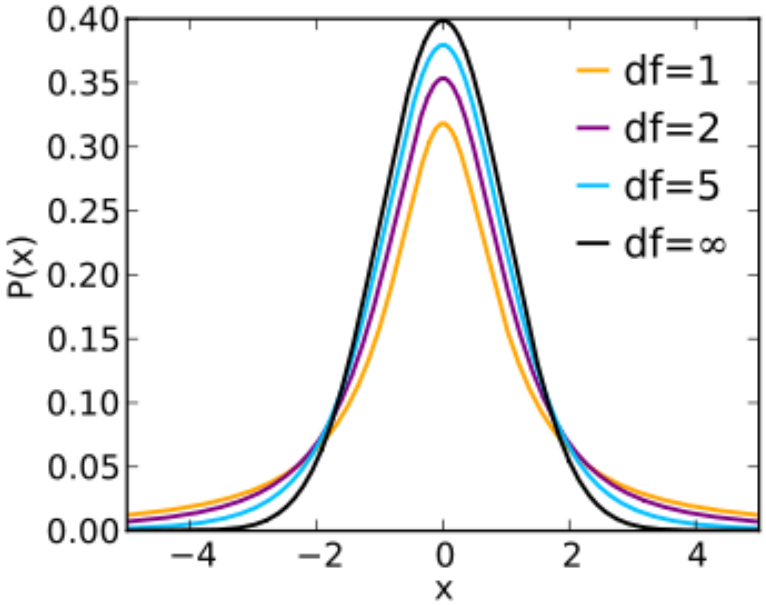
$$\Sigma = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$$

scatter plot



Student‘s t-Distribution

- **Student’s *t*-distribution** (or simply the ***t*-distribution**) is a continuous probability distribution that arises when estimating the mean of a normally distributed population in situations where the sample size is small.
- It plays a role in a number of statistical analysis (e.g. *t*-test)



$$\text{expectation value}: 0$$

$$\text{variance}: \frac{\nu}{\nu-2}, \nu > 2$$

- t-distribution looks similiar to normal distribution, but has heavier tails.

$$t_\nu(x) := t(x \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$$

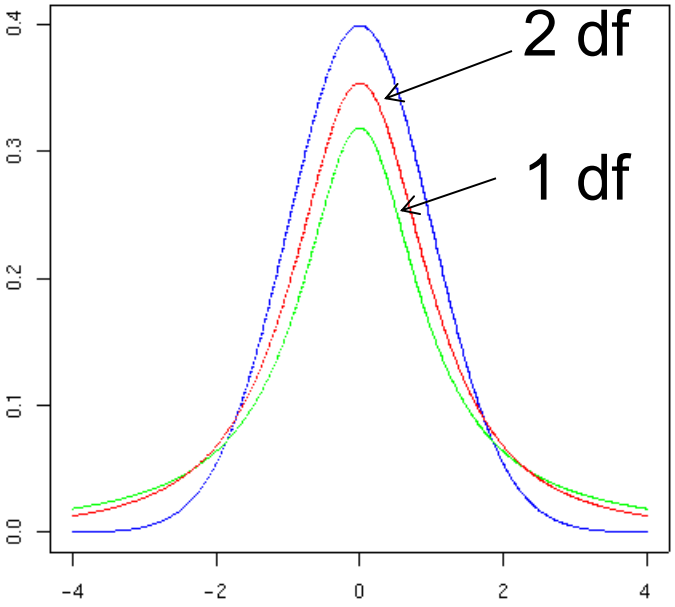
$$\Gamma(z) := \int_0^\infty t^{z-1}e^{-t}dt$$

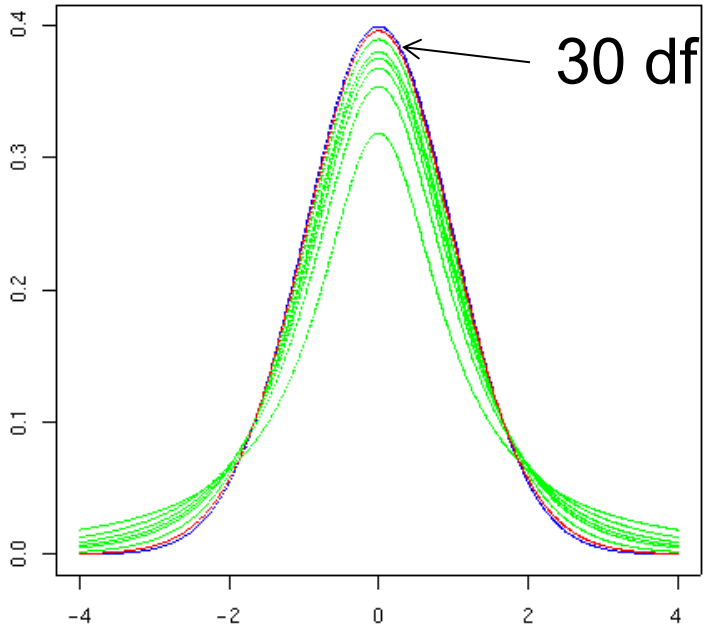
(so-called *Gamma* function)

- v is the so-called „**degree of freedom**“

Relationship to Normal Distribution

- With more df the t-distribution becomes closer and closer to a normal distribution (in blue)





- Expectation value of t-distribution: 0
- Variance:

$$\frac{\nu}{\nu-2} \text{ for } \nu > 2$$

Application

- Suppose we have a normal distribution N(μ, σ^2) of which we sample n data points with mean $\bar{x}$ and variance s^2
- The t-statistic

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

follows a t-distribution with n-1 degrees of freedom.

- The quantity $se(\bar{x}) := s / \sqrt{n}$ is called ***standard error*** of the mean.
- It is an estimate of the variance, which $\bar{x}$ would show under repeated drawings of n data points.

- The t-statistic describes the distribution of the (normalized) difference to μ, which $\bar{x}$ exhibits under (notional) repeated data samplings.

Cumulative Distribution Function (CDF)

- For every real number x, the CDF of a real-valued random variable X is given by

$$F(x) = P(X \leq x)$$

- For discrete random variables that implies (p = pmf):

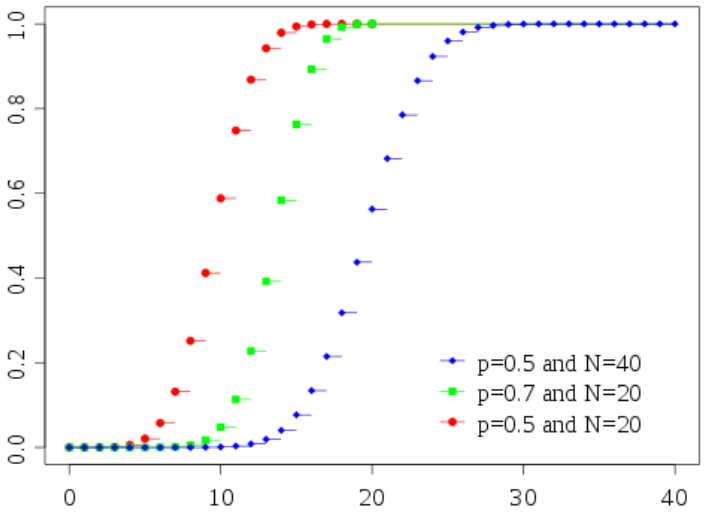
$$F(x) = \sum_{x_i \leq x} p(x_i)$$

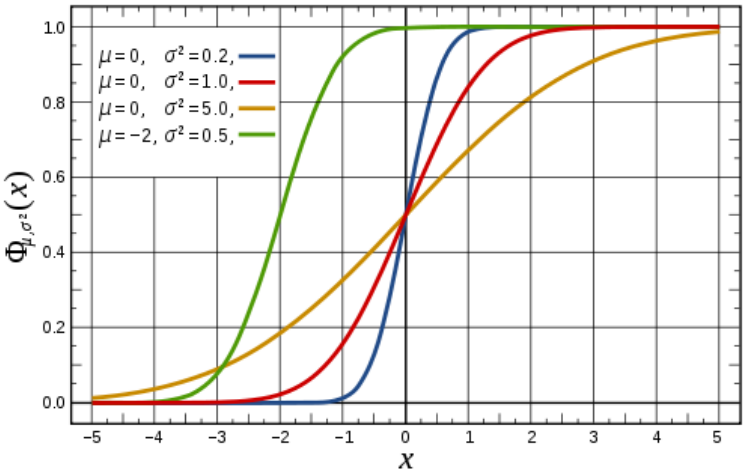
- Likewise, for continous random variables (p = PDF):

$$F(x) = \int_{-\infty}^{x} p(t)dt$$

  - The CDF of the standard normal distribution is:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$





Example

- What is the probability to find a male person with a height larger than 2.00m, if we know that μ=1.80m and σ=0.1m?

$$\Pr(X > 2 \mid \mu, \sigma) = 1 - \Pr(X \leq 2 \mid \mu, \sigma)$$

$$= 1 - \int_{-\infty}^{2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2} \right) dt$$

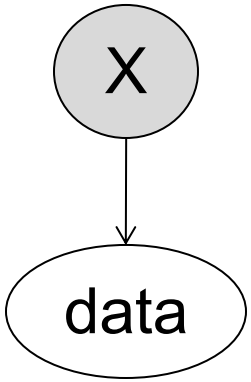
→ use R function pnorm to solve this:

```
> 1-pnorm(2,1.8,0.1)
0.02275013
```

- Answer: there is only a ~2.3% chance of seeing a male, who is larger than 2m.

Descriptive & Inductive Statistics, Predictions

- **Descriptive statistics:** describe collections of data in quantitative terms (see first slides).
- **Inductive / inferential statistics:** infer population characteristics (i.e. features of the underlying random variables) from data
  - Idea: data is obtained by sampling from a (possibly unknown) random variable
- **Predictions:** use inferred population characteristics to make predictions on so far unseen data (aka **machine learning**)



Statistical Models

- How to draw conclusions from data?
  - Need a **statistical model** of random data generation process
  - Statistical model = random variables, relationships these variables
- Result: estimates of population characteristics (we don't know the truth!)

- Statistical inference approaches
  - Bayesian ➔ Bayes theorem
  - Frequentist

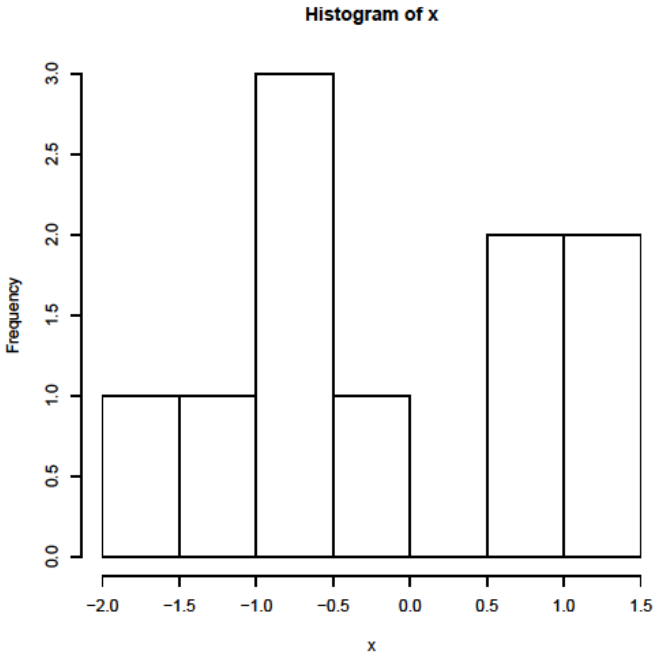
Example

- Supposed we are given a set of numbers

D = {-1.1485009, -0.6639319, 1.2620949, -0.3414443, -0.7440194, 0.7833574, 0.7375564, -0.5784752 , -1.7383867, 1.1788375}

- What could be a good model for these data?
  - One variable ➔ underlying statistical distribution?

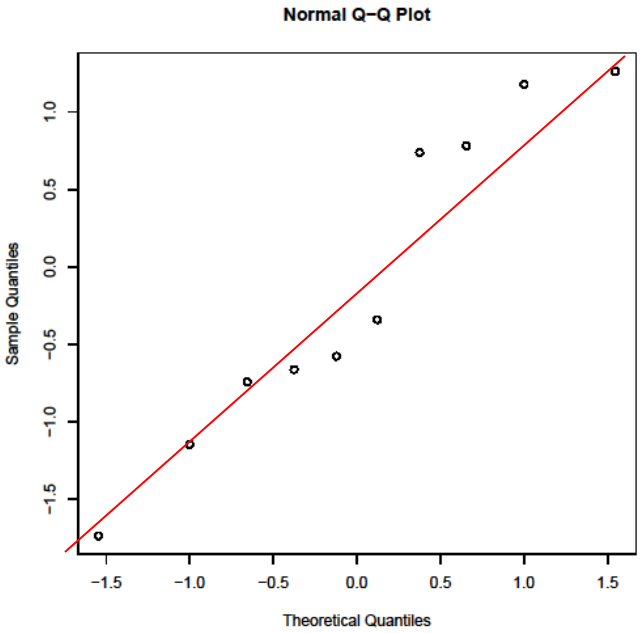
- We can plot a histogram of D (right)




```
> hist(D)
```

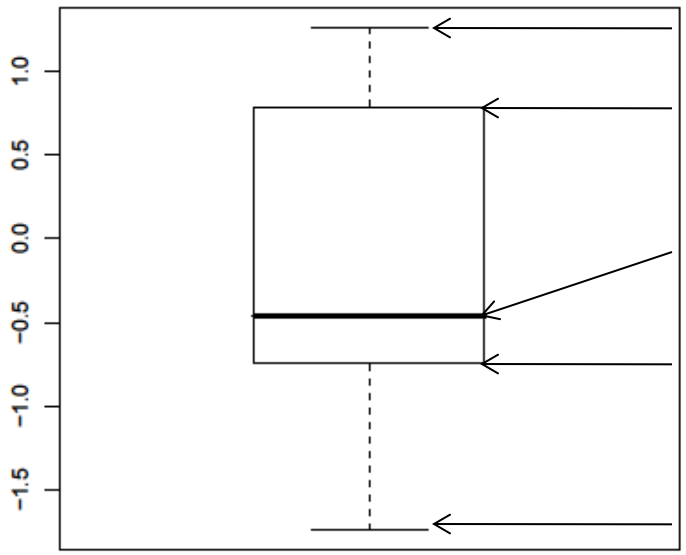
Example (cont‘d)

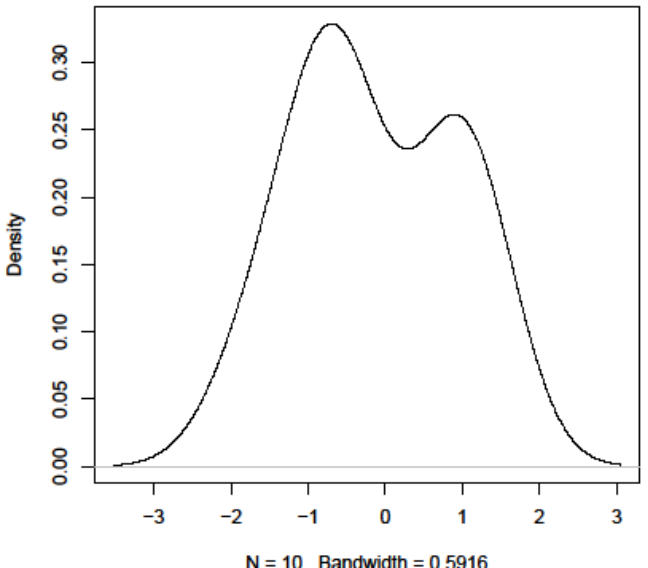
- We may further inspect a **QQ-plot** of the data
  - The quantiles of the sample are plotted against the quantiles of a theoretical distribution (e.g. standard normal)
  - Sample quantiles are calculated by replacing probabilities with relative frequencies
  - If values approximately lie on the line y=x, the distributions are similiar
  - If values approximately lie on a line y = x + a, the distributions are linearly related



Example (cont.'d)

- Also a boxplot and a **density plot** may help
- Both approaches allow to look at the sample distribution of values.





Example (cont'd)

- We may now hypothesize that D was drawn from a normal distribution with unknown expectation and variance.

- We can compute the sample mean and variance:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = -0.1252912$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1.041891$$

- Please note: D is a sample from a random variable X
- Sample mean and variance do not need to coincide with the E[X] and Var[X]!
  - indeed: data was drawn from N(0, 1) here!

Frequentist Inference: Maximum Likelihood (ML)

- How can we infer the parameters (μ, σ^2) of X given data D?
- Two principle ways of inference
  - Bayesian inference: rule of Bayes → posterior distribution over parameters
  - Frequentist inference: maximum likelihood → parameter point estimates

- Idea of ML: suppose E[X] = μ and Var[X] = σ^2. What would be the joint probability density for data D, if observations are statistically independent from each other?

$$p(D \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

Frequentist Inference: Maximum Likelihood (ML)

- View this as function of parameters μ, σ^2:
  - Find the parameter combination maximizing the ***likelihood function*** $L(\mu, \sigma^2) := p(D \mid \mu, \sigma^2)$

$$(\hat{\mu}, \hat{\sigma}^2) = \arg\max L(\mu, \sigma^2)$$

$$\text{with } L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

- For numerical reasons we usually consider the log-likelihood:

$$\log L(\mu, \sigma^2) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2)$$

Maximum Likelihood (cont'd)

- We look for the parameters yielding a local maximum of this function:

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma^2) = 0$$

$$\frac{\partial}{\partial \sigma} \log L(\mu, \sigma^2) = 0$$

$$\Rightarrow$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{n-1}{n} s^2$$

- Sometimes $s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is also called sample variance

- Conclusion: sample mean and s_n^2 are the ML estimates of the population mean and variance.

ML Estimates for Tumor Example

- ML estimates:

$$\hat{\mu}(tumor) = 6.070$$

$$\hat{\mu}(normal) = 1.257$$

$$\hat{\sigma}(tumor) = 1.18$$

$$\hat{\sigma}(normal) = 0.44$$

- Given these population estimates, could the observed difference (or an even more extreme one) $\hat{\mu}(tumor) - \hat{\mu}(normal)$ be due to chance?

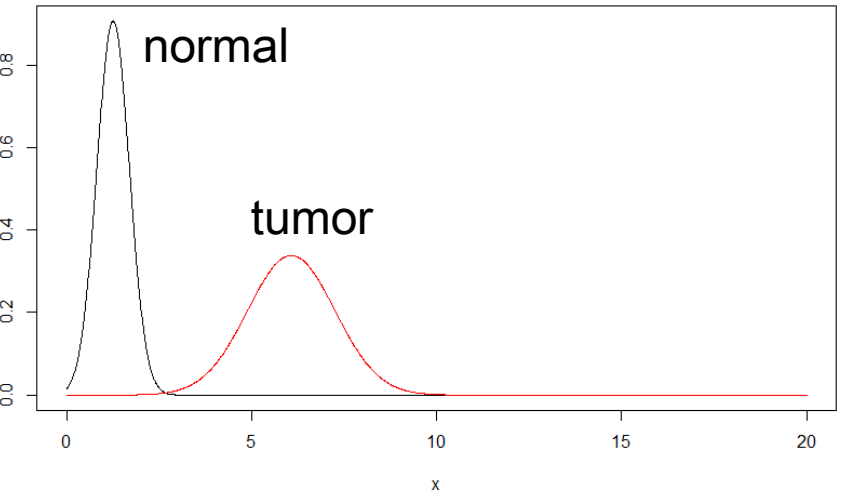
tumor	normal
5.423	1.234
6.239	0.283
8.288	1.488
4.999	1.048
5.399	0.599

mean (tumor) = 6.070

mean (normal) = 1.257

std(tumor) = 1.319

std(normal) = 0.486



Hypothesis Tests

- A **statistical hypothesis test** is a method of making decisions using data
- A result is called **statistically significant** if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the **significance level**.
- **Null hypothesis:** a hypothesis, which we like to falsify

- Basic approach: *Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?*

- Analogy: court trial
  - A defendant is considered not guilty as long as his guilt is not proven.
  - The prosecutor tries to prove the guilt of the defendant.
  - Only when there is enough charging evidence the defendant is convicted.

Hypothesis Testing (cont.'d)

- Two hypotheses
  - H_0: "the defendant is not guilty" (**null hypothesis**)
  - H_1: "the defendant is guilty" (**alternative hypothesis**)
- Hypothesis tests always try to reject the null hypothesis in order to gain confidence for the alternative hypothesis.
- Possible outcomes:

	Null Hypothesis (H_0) is true	**Alternative Hypothesis (H_1) is true**
Accept Null Hypothesis	Right decision	Type II error
Reject Null Hypothesis	Type I error	Right decision

T-Test

- A ***t*-test** is any statistical hypothesis test in which the test statistic follows a *t*-distribution, if the null hypothesis is supported.
- Types of t-test:
  - One-sample: tests whether the mean of a normally distributed population has a specified value (e.g. 0)
  - Unpaired two-sample t-test: tests whether the means of two normally distributed random variables are equal
    - Variant: Welch‘s t-test
  - Paired two-sample t-test: testes whether the difference between two normally distributed random variables is 0
    - Same patients before and after treatment

- T-test only works with normally distributed random variables

One-sample t-test

- Look back at our previous data: D = {-1.1485009, -0.6639319, 1.2620949, -0.3414443, -0.7440194, 0.7833574, 0.7375564, -0.5784752, -1.7383867, 1.1788375}
- We have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = -0.1252912$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1.041891$$

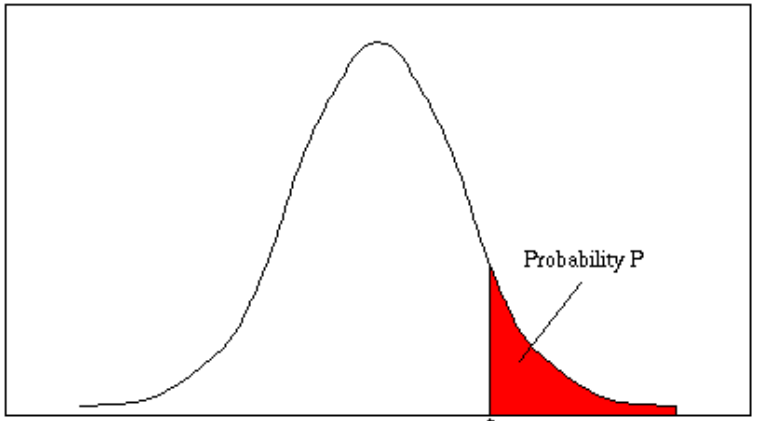
- H0: expectation value of random variable (μ) is <= 0 (generally: μ_0)
- H1: $\mu > \mu_0$
- We use the t-statistic

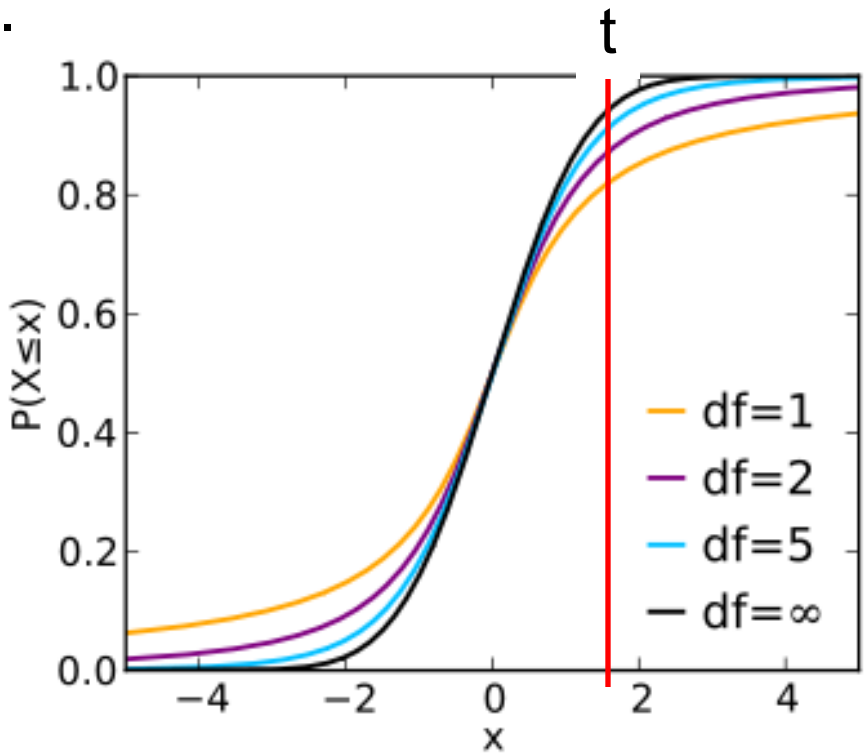
$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

which is t-distributed with df = n - 1

One-sample t-test (cont.'d)

- What is the probability to get a t-statistic higher than t?
  - Report the **area** under the t-distribution with n-1 df larger than t. How?
    - Take the CDF of the t-distribution with n – 1 df.
    - Look at the value at t: This is the Pr(X<=t)
    - But: we want Pr(X>t). How can we get it?
    - Pr(X>t) = 1 – Pr(X <= t)
- This value is the so-called p-value (p)
  - Here: p = 0.64





P-Values

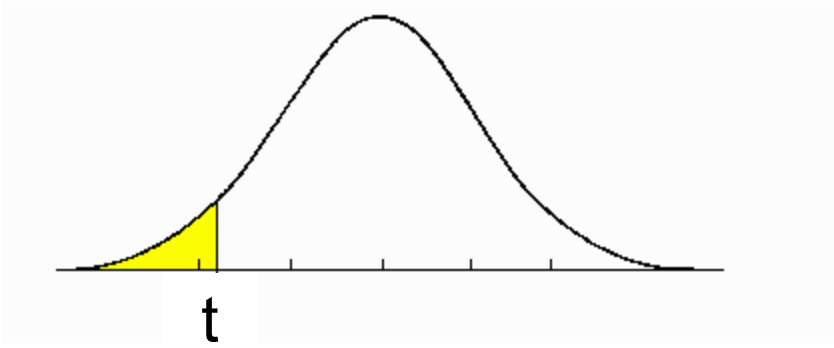
- The p-value tells the probability to see a result at least as extreme as the one observed given the null hypothesis is true.
- It is NOT the probability of the null hypothesis
- 1 – p is NOT the probability of the alternative hypothesis!!!

- One often rejects a null hypothesis, if the p-value is < 5%.
  - The result is then called significant.
- If the p-value < 1% it is called highly significant.

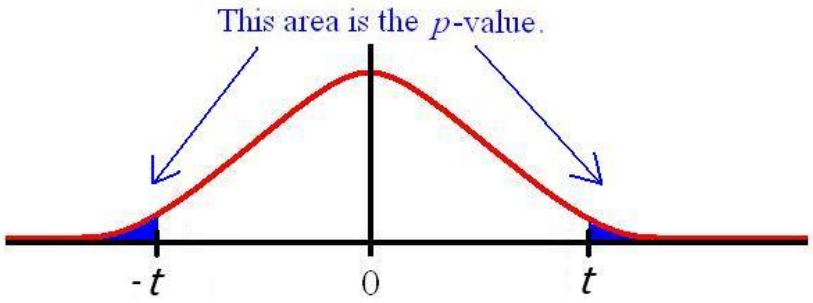
- The p-value is equal to the type I error. A statistical testing procedure therefore controls the type I error.
- The type II error (false acceptance of null hypothesis) is much more difficult to control.
- 1 – type II error is also called the *statistical power* of a test.

Other Hypothesis

- How do we test the opposite null hypothesis?
  - H0: expectation value of random variable (μ) is >= 0 (generally: μ_0)
  - H1: $\mu < \mu_0$
  - Same test procedure, compute Pr(X <= t) from CDF



- How do perform a symmetrical (two-sided) test?
  - H0: expectation value of random variable (μ) is = 0 (generally: μ_0)
  - H1: $\mu \neq \mu_0$
    - Compute Pr(X <= -t)
    - Compute Pr(X > t)
    - Add both together



Two-sample t-test

- Equal sample sizes, equal variances in both groups:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 + s_2^2)/n}}$$

$$df = 2n - 2$$

$$\bar{x}_1, \bar{x}_2 = \text{means of groups 1 and 2}$$

$$s_1^2, s_2^2 = \text{variances of groups}$$

- Unequal sample sizes, equal variances in both groups:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$df = n_1 + n_2 - 2$$

$$n_1, n_2 = \text{sample sizes}$$

Two-sample t-test (cont.'d)

- Unequal sample sizes, unequal variances in both groups (Welch-test):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 /(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2 /(n_2 - 1)}$$

- Paired t-test:
  - Compute differences between paired samples from both groups and take them as a new sample
  - Apply one-sample t-test to the new sample

Example

- P53 measurements in tumor and normal cells
- Case 1: Samples were taken from different individuals

tumor	normal
5.423	1.234
6.239	0.283
8.288	1.488
4.999	1.048
5.399	0.599

$$\bar{x}_{tumor} = 6.07, s_{tumor} = 1.32$$

$$\bar{x}_{normal} = 0.93, s_{normal} = 0.49$$

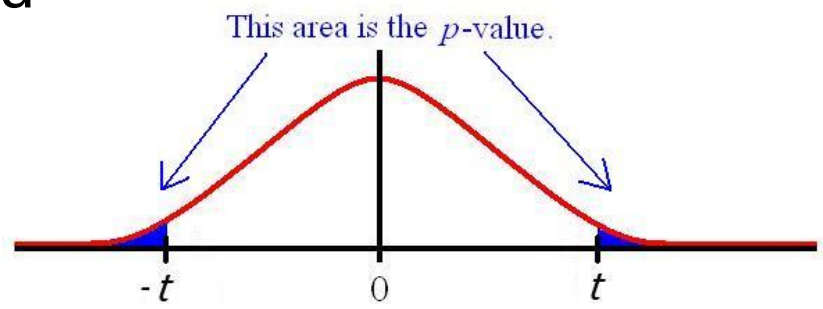
- We have the same sample size, but unequal variances in both groups ➔ Welch test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} = \frac{6.07 - 0.93}{\sqrt{\frac{1.32^2 + 0.49^2}{5}}} \approx 8.172$$

$$df = 5.067$$

Example (cont.'d)

- We decide for a two-sided test (H_0: $\mu_{tumor} = \mu_{normal}$)
- We have to compute Pr(X > t) and Pr(X <= -t) from the CDF of the t-distribution
  - Both are equal, because the t-distribution is symmetric
  - Both have to be added together
  - In R: `pvalue = pt(-8.172,5.067)*2`
  - p-value is 0.0004



Example (cont.'d)

- Case 2: Samples were taken from the **same** individual
  - We compute a paired t-test, which is identical to a one-sample t-test of differences

$$\bar{x} = 5.139, s = 1.21$$

$$t = \frac{\bar{x}}{s/\sqrt{5}} = 9.498$$

$$df = 5 - 1 = 4$$

  - Try to reject H0: $\mu = 0$
  - The test yields a p-value of ~0.0007

tumor	normal	difference
5.423	1.234	4.189
6.239	0.283	5.956
8.288	1.488	6.8
4.999	1.048	3.951
5.399	0.599	4.8

Other Tests

- T-test assumes normally distributed random variables
- What, if this is not the case?
  - *Mann-Whitney U test* / Mann–Whitney–Wilcoxon / Wilcoxon rank-sum test as a **non-parametric** alternative to a t-test
  - *Wilcoxon signed rank test* as alternative to a paired t-test
  - Both tests make NO assumptions about the distribution of the underlying random variables
  - ➔ Use rank statistics instead

  - Advantage: works, if normallity assumption does not hold true
  - Disadvantage: can loose statistical power, if it actually holds true

- *Kolmogorov-Smirnov test* can be used to test data against a reference distribution
- Specifically for testing against normal distribution: *Shapiro-Wilk test*

Permutation Tests

- P-value = probability to obtain test statistic at least as extreme as the observed one
- So far: known distribution of test statistic
- What, if this is not fulfilled?
- Idea:

$$p - value = \Pr(T(y) >= T(x))$$

  - x = original data
  - y = randomly permuted data
  - T(x) = statistic on original data
  - T(y) = statistic on permuted data

- We can do permutations on a computer!

Example

Species Richness	Lake Area
32	2.0
29	0.9
35	3.1
36	3.0
41	3.0

- Pearson correlation coefficient ~0.84
- Could this result have been expected by chance?

For i = 1...N (N should be LARGE – why?)

- Randomly shuffle around assigments of lake areas to species richness
- Compute correlation

$$p_{empirical} = \frac{\#(r_{random} \geq r_{real})}{N}$$

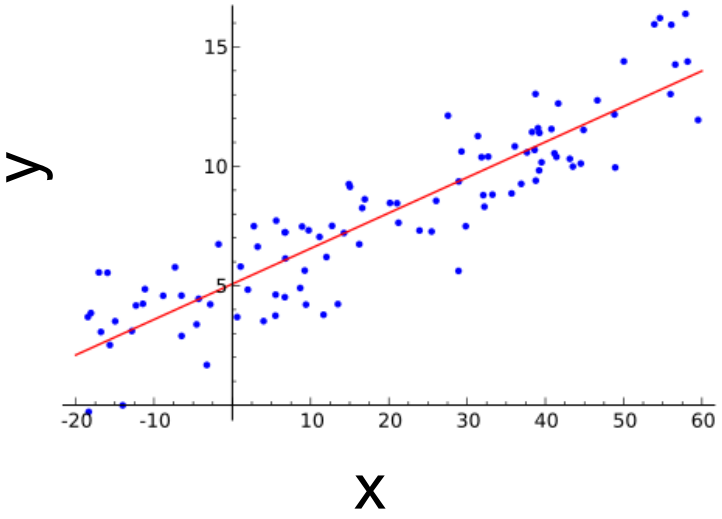
Back to Statistical Models

- **So far**: one random variable, associations (differences, correlations) between two random variables

- Now: more complicated relationships

Simple linear regression

- Consider a scatter plot (right)
- There appears to be a linear relationship (red line) between x and y
- At the same time there is some noise
  - We assume this noise to be normally distributed
- Denote sample points by $\{x_i, y_i\}$, i=1,...,n.
- We have the following equation for y_i:



$$y_i = \beta_o + \beta x_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \eta)$$

Least squares fitting

- An obvious approach for fitting parameters (β_o, β) is to minimize the squared difference (also called residual) between model fits and data (so-called least squares fit):

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0, \beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i)^2$$

- Take the partial derivatives w.r.t. (β_o, β) and set them to 0:

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i)^2 = -\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta x_i) x_i = 0$$

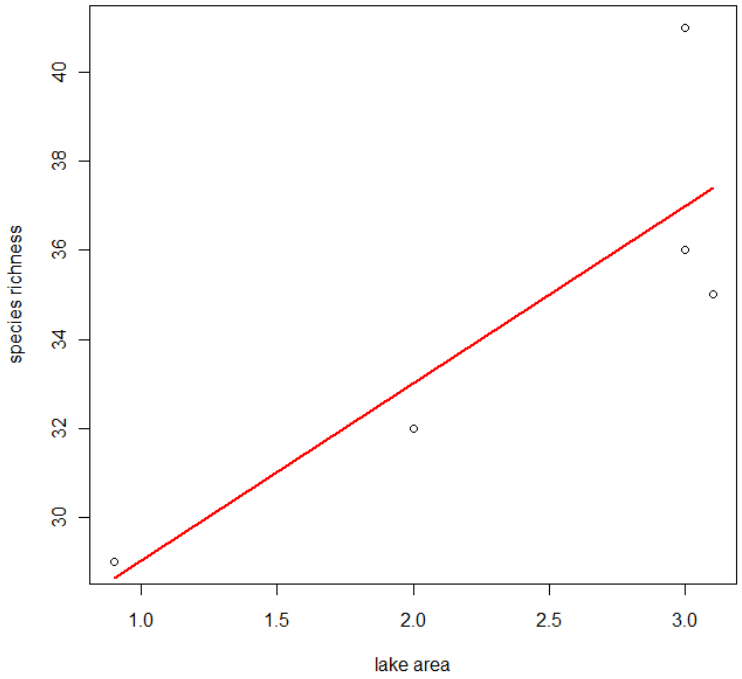
$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i)^2 = -\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta x_i) = 0$$

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$$

Example

Species Richness	Lake Area
32	2.0
29	0.9
35	3.1
36	3.0
41	3.0



$$\hat{\beta} = \frac{\text{cov}(\text{species richness}, \text{lake area})}{\text{var}(\text{lake area})} = \frac{3.6}{0.905} \approx 3.978$$

$$\hat{\beta}_0 = mean(\text{species richness}) - \hat{\beta} * mean(\text{lake area}) = 25.053$$

Comparison with Maximum Likelihood

- We assumed the noise to be normally distributed.
- Hence: data points should be normally distributed around the regression line.

- For the i-th data point we have the density $N(y_i - \beta_0 - \beta x_i, \varepsilon_i)$
- For all n data points the joint density is

$$p(\mathrm{x}, \mathbf{y} \mid \beta, \beta_0, \boldsymbol{\varepsilon}) = \prod_{i=1}^{n} N(y_i - \beta_0 - \beta x_i, \varepsilon_i)$$

- The *negative* log-likelihood is thus

$$-\log p(\mathbf{x}, \mathbf{y} \mid \beta, \beta_0, \boldsymbol{\varepsilon}) = \sum_{i=1}^{n} \left(\frac{(y_i - \beta_0 - \beta x_i)^2}{2\varepsilon_i^2} + \sqrt{2\pi}\varepsilon_i \right)$$

- The expression is minimized by minimizing the sum of squared model residuals ➔ identical to least squares fit

Standard Errors of Regression Coefficients

- Every estimation is subject to uncertainty
  - Resampling of new data may lead to different regression coefficients
  - Comparative situation: (re-)calculation of the empirical mean for different samples
  - We should report the estimated variation of regression coefficients under notional resampling of data ➔ standard errors

$$se(\hat{\beta}) = \hat{\eta}\sqrt{\frac{1}{(n-1)\,\mathrm{var}(x)}}$$

$$se(\hat{\beta}_0) = \hat{\eta}\sqrt{\frac{\sum_i x_i^2}{n(n-1)\,\mathrm{var}(x)}}$$

$$\hat{\eta}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Testing significance

- Does the predictor variable have a ***significant*** influence on the fitted value?
  - We test the hypothesis $H_0 : \beta = 0$
  - We form the standardized coefficient or *Z-score*:

$$z = \frac{\hat{\beta}}{se(\hat{\beta})}$$

  - Under the null hypothesis that $\beta = 0$ we have

$$z \sim t_{n-2}$$

  - Significance can be tested via a one-sample t-test

Example

$x = \text{lake area},\ y = \text{species richness}$

regression coefficients:

$\hat{\beta} \approx 3.978$

$\hat{\beta}_0 \approx 25.053$

estimated noise/residual variance:

$$\hat{\eta}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}x_i)^2 \approx 7.97$$

$$se(\hat{\beta}) = \hat{\eta}\sqrt{\frac{1}{(n-1)\,\mathrm{var}(x)}} \approx 1.484$$

$$z = \frac{\hat{\beta}}{se(\hat{\beta})} \approx 2.86$$

$$p - value = 2 * pt(2.86, n-2) \approx 0.075$$

Species Richness	Lake Area
32	2.0
29	0.9
35	3.1
36	3.0
41	3.0

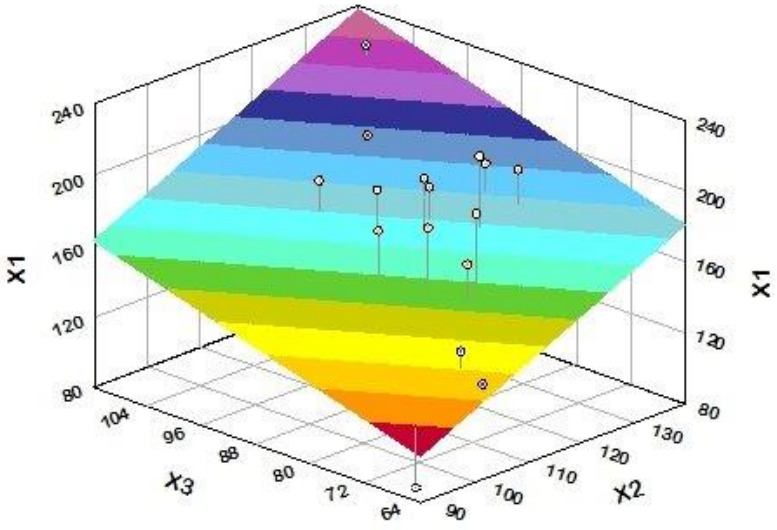
- $\hat{\eta}^2$ is called residual variance
- It gives an indication, how much of the y-variance is unexplained by the model

Linear regression with several regressors / predictor variables

- **Goal:** Estimate linear relationship between *regressors / predictor variables* $\{\mathbf{x}_1,...,\mathbf{x}_n\}$ and the dependent response variable y_i.
- Regressors in general can be vectors of dimension p.
- Linear relationship is disturbed by some normally distributed noise ε_i
- Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \eta)$$



Linear Regression Analysis with several regressors / predictors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

$$= \beta_0 1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

$$= \left(1, x_{i1}, x_{i2}, \ldots, x_{ip}\right)^T \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \varepsilon_i$$

$$= \tilde{\mathbf{x}}_i^T \tilde{\boldsymbol{\beta}} + \varepsilon_i$$

- We can stack all n equations together into a matrix-vector form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{x}}_1^T \\ \tilde{\mathbf{x}}_2^T \\ \vdots \\ \tilde{\mathbf{x}}_n^T \end{pmatrix} \tilde{\boldsymbol{\beta}} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$$

- Matrix X is called **design matrix** of the linear model.
- How to estimate $\tilde{\boldsymbol{\beta}}$ via maximum likelihood?

Regression Analysis (cont.'d)

- ML estimate: minimize the sum of squared model residuals!

$$\sum_{i=1}^{n} \underbrace{\left(y_i - \tilde{\mathbf{x}}_i^T \tilde{\boldsymbol{\beta}}\right)}^{2} \quad \text{residual}$$

$$= \| \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} \|_2^2$$

$$= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

- Solution: Take derivative w.r.t. parameters and set it to 0 (maximum likelihood approach)

$$\frac{d}{d\tilde{\boldsymbol{\beta}}} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

$$= \frac{d}{d\tilde{\boldsymbol{\beta}}} \left(\mathbf{y}^T\mathbf{y} - 2\tilde{\boldsymbol{\beta}}^T\mathbf{X}\mathbf{y} + \tilde{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}\right)$$

$$= -2\mathbf{X}\mathbf{y} + \mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\beta}}$$

$$= 0$$

$$\Rightarrow \quad \tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$

Important Assumptions of Multiple Linear Regression

- Model residuals are normally distributed
- p < n (fewer regressors than data points)
- No collinearities: No variable can be written as a linear combination of other variables
- Example:

```
 [1,]  4.04414119  8.0882824
 [2,] -2.21462876 -4.4292575
 [3,]  0.08596091  0.1719218
 [4,]  3.77235181  7.5447036
 [5,]  2.01672458  4.0334492
 [6,]  0.05481749  0.1096350
 [7,]  1.99911007  3.9982201
 [8,] -1.79666442 -3.5933288
 [9,] -0.38206636 -0.7641327
[10,]  0.60313553  1.2062711
```

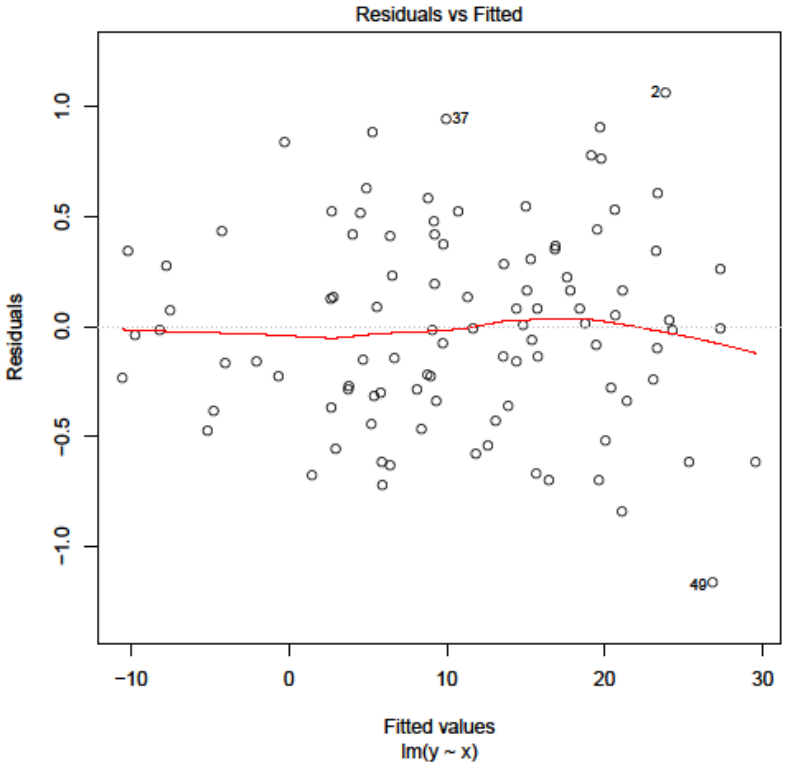
- Second column is just twice the first one ➔ assumption violated!

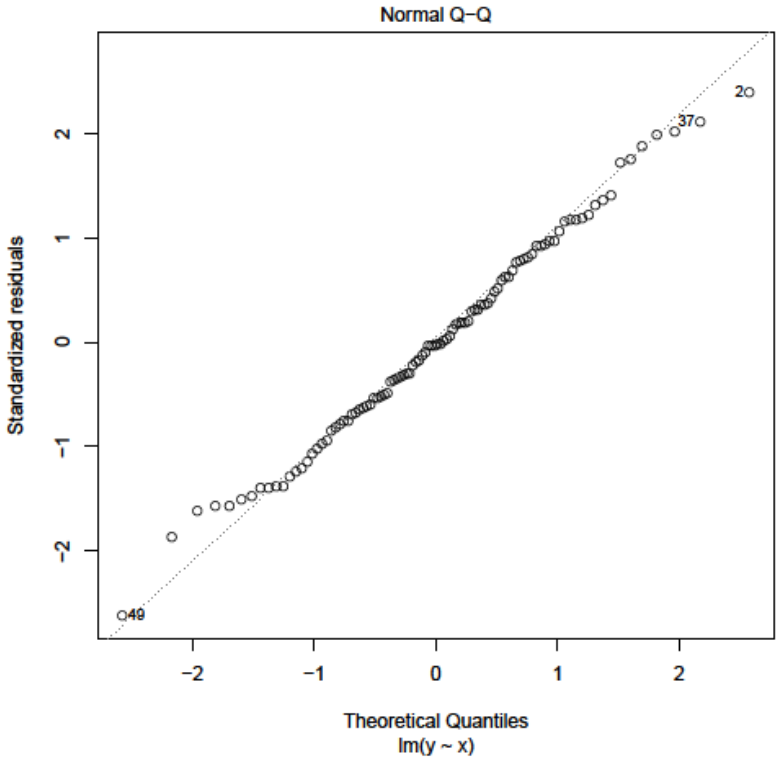
Example

- Regression analysis in R:

```
> X = cbind(2*rnorm(100, mean=1), 5*rnorm(100, mean=-1))
   # generate some normally distributed data

> Y = X[,1] - 2*X[,2] + rnorm(100, sd=0.5) # generate a
   functional dependency of X and Y + some noise
> fit = lm(Y~X)
> coef(fit) # guess, how our coefficients look like

> plot(fit)
```

Quality Control: Plots





- Residuals are symmetric around 0 and do not show a trend
- Residuals are normally distributed

Quality Control: Model Summary

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3093 -0.2996  0.0469  0.3504  1.1858

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) -0.16442    0.10426   -1.577    0.118
X1           0.99371    0.02837   35.031   <2e-16 ***
X2          -2.02442    0.01103 -183.455   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5118 on 97 degrees of freedom
Multiple R-Squared: 0.9971,     Adjusted R-squared: 0.9971
F-statistic: 1.686e+04 on 2 and 97 DF,  p-value: < 2.2e-16
```

residual distribution

regression coefficients

Overall model fit

Model Summary: Residual Distribution and Coefficients

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.3093 -0.2996  0.0469  0.3504  1.1858
```

- Shows information on the distribution of residuals ➔ they are symmetric around 0

```
Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -0.16442    0.10426   -1.577    0.118
X1           0.99371    0.02837   35.031   <2e-16 ***
X2          -2.02442    0.01103 -183.455   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
   ' ' 1
```

- 1st column: estimated coefficients, 2nd column: their std. errors.
- What do the other columns mean?

Regression Analysis: Quality Control, step by step

- 3rd and 4th columns: t-value and Pr(>|t|)
  - We test the hypothesis that a particular coefficient $\tilde{\beta}_i = 0$
  - We form the standardized coefficient or *Z-score*:

$$z_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

  - Under the null hypothesis that $\tilde{\beta}_i = 0$ we have

$$z_i \sim t_{n-p-1}$$

  - This is the reported „t-value“
  - Significance can be tested via a t-test

Quality Control: Overall Model Fit

```
Residual standard error: 0.5118 on 97 degrees of freedom
Multiple R-Squared: 0.9971, Adjusted R-squared: 0.9971
F-statistic: 1.686e+04 on 2 and 97 DF,  p-value: < 2.2e-
    16
```

- Multiple R-squared (*coefficient of determination*): The fraction of *y*-variance explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

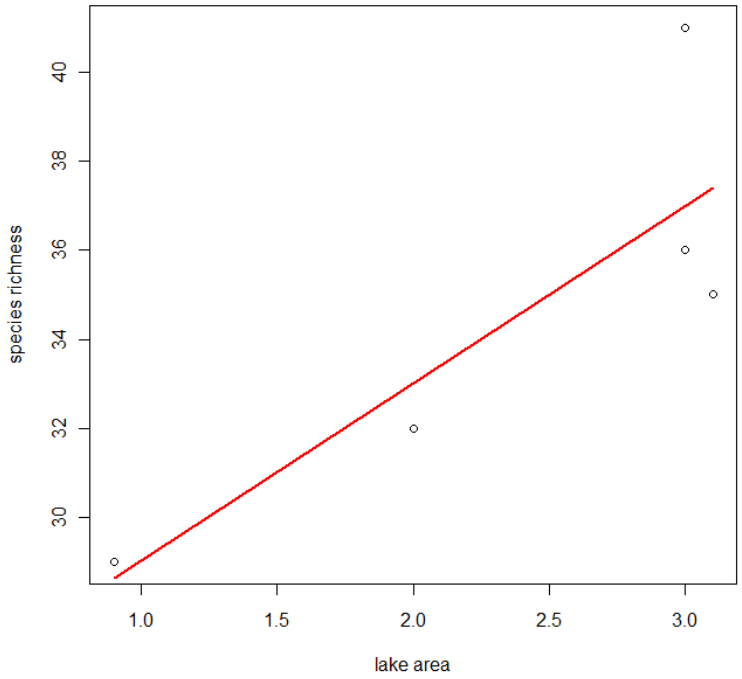
simple linear regression :

$$R^2 = r_{xy}^2 = \frac{\mathrm{cov}^2(x,y)}{\mathrm{var}(x)\,\mathrm{var}(y)}$$

- Equivalent to the squared correlation between **y** and $\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)_{i=1}^{n}$

Example

Species Richness	Lake Area
32	2.0
29	0.9
35	3.1
36	3.0
41	3.0



$$R^2 = \frac{\text{cov}^2(\text{species richness, lake area})}{\text{var}(\text{species richness})\,\text{var}(\text{lake area})} = \frac{12.96}{20.3 * 0.905} = 0.7054$$

- The lake area explains ~70% of the observed variance of the species richness
- 30% are unexplained and may be due to unobserved variables

- Problem with R^2 for multiple linear regression: Increases the more variables are added, regardless of whether these additional variables improve the model

- Adjusted R-squared: adjusts R^2 for the number of regressors in a model. Unlike R^2, the adjusted R^2 increases with increasing number of regressors only if the new term improves the model more than would be expected by chance:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Quality Control: Overall Model Fit

```
Residual standard error: 0.5118 on 97 degrees of freedom
Multiple R-Squared: 0.9971, Adjusted R-squared: 0.9971
F-statistic: 1.686e+04 on 2 and 97 DF,  p-value: < 2.2e-16
```

- What does the p-value mean?
  - ➔ Result of an ANOVA F-Test

ANOVA (Analysis of Variance) F-Statistic / F-Test

```
> anova(fit)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          2 8833.9  4416.9   16861 < 2.2e-16 ***
Residuals 97   25.4     0.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Does the regression model differ significantly from an empty / null model without any predictor variables?
  - Null hypothesis: all coefficients are 0; alt. hyp.: at least one is not 0

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_1 : \exists j : \beta_j \neq 0$$

ANOVA F-Test

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

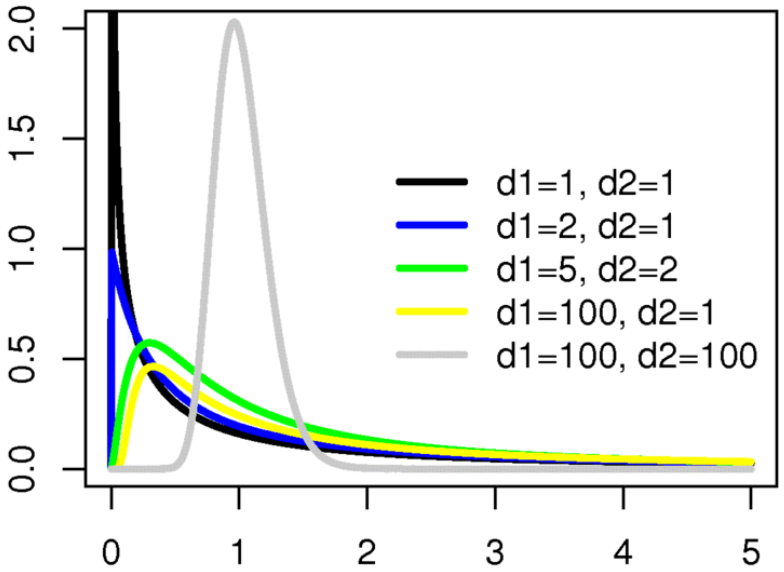
$$\sum_{i=1}^{n}(y_i - \tilde{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}})^2 = \underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{pure error}} + \underbrace{\sum_{i=1}^{n}(\bar{y} - \tilde{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}})^2}_{\text{lack of fit}}$$

- Test statistic:

$$F = \frac{\text{lack of fit / df}}{\text{pure error / df}} = \frac{\sum_{i=1}^{n}(\bar{y} - \tilde{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}})^2 / p}{\sum_{i=1}^{n}(y_i - \bar{y})^2 /(n-p-1)}$$

follows a $F_{p,n-p-1}$ distribution.

- Compute p-value as usual.



ANOVA F-Test for Comparing Groups

- ANOVA can also be used to study the differences between two or more independent groups. This is the most common use of ANOVA.
  - In case of two groups it is equivalent to a t-test
  - Idea: x now indicates the membership to a particular group
- Example: ANOVA model with 3 groups (A, B, C)

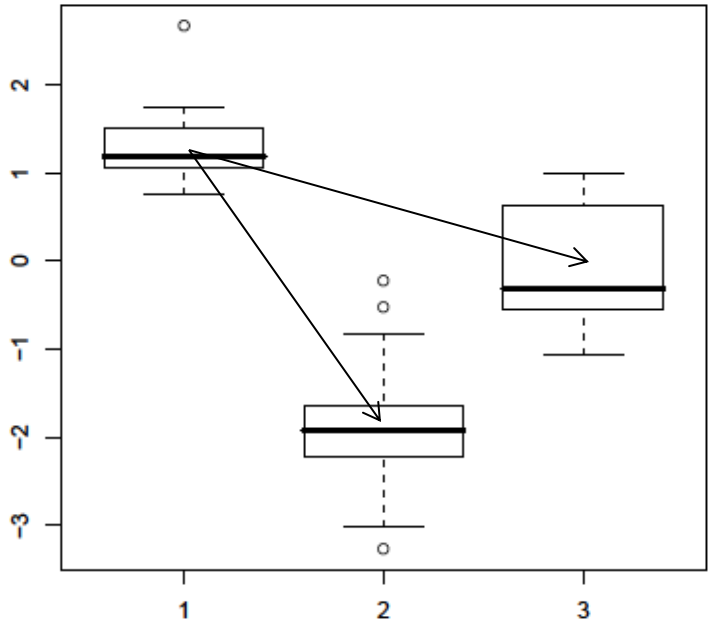
$$\mathbf{y} = \begin{pmatrix} \text{intercept} & B-\text{int.} & C-\text{int.} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Important assumption of ANOVA: Variances in groups should be the same (so-called homoscedasticity)

Example

```
> X = c(rnorm(10,mean=1), rnorm(15,mean=-2), rnorm(7))
> Y=factor(c(rep(1,10),rep(2,15),rep(3,7)))
> boxplot(X~Y)
```

Is the difference between groups on average larger than expected by chance?




```
> fit=lm(X~Y)
> fit

Call:
lm(formula = X ~ Y)

Coefficients:
(Intercept)           Y2           Y3
      1.356       -3.222       -1.388
```

First group average

Difference of second to first group average

Difference of third to first group average

Example (cont.'d)

```
> anova(fit)
Analysis of Variance Table

Response: X
          Df Sum Sq Mean Sq F value    Pr(>F)
Y          2 63.913  31.957  54.726 1.432e-10 ***
Residuals 29 16.934   0.584
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
   0.1 ' ' 1
```

Group means differ significantly

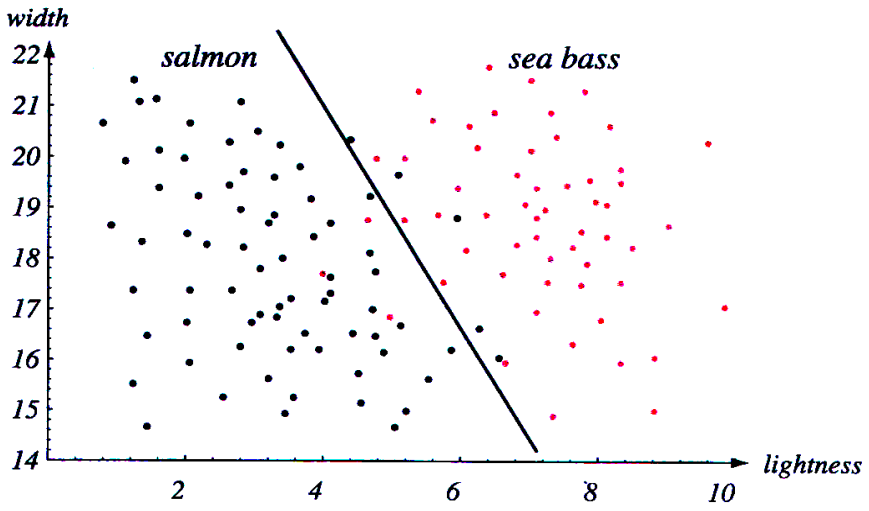
Reason: We can reject the null hypotheses that all regression coefficients (i.e. differences to the first group mean) are 0.

Logistic Regression

- Previously: linear regression (real valued response variable *y*)
- Now: classification (sorting objects into *K* categories / classes), i.e.

$$y \in \{0,1,\ldots K-1\}$$

- Let **x** be the vector of predictors for one case sample drawn from *X*
- We model the *log-odds ratio* for **x** belonging to each class as a linear function:



Source: Duda, Hart, Stork, Pattern Classification, Wiley Interscience, 2000

For K = 2:

$$\ln \frac{\Pr(y=0 \mid X=\mathbf{x})}{\Pr(y=1 \mid X=\mathbf{x})} = \beta_0 + \sum_{i=1}^{d} \beta_i x_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}$$

General case:

$$\ln \frac{\Pr(y=0 \mid X=\mathbf{x})}{\Pr(y=K \mid X=\mathbf{x})} = \beta_{00} + \boldsymbol{\beta}_1^T \mathbf{x}$$

$$\ln \frac{\Pr(y=1 \mid X=\mathbf{x})}{\Pr(y=K \mid X=\mathbf{x})} = \beta_{10} + \boldsymbol{\beta}_2^T \mathbf{x}$$

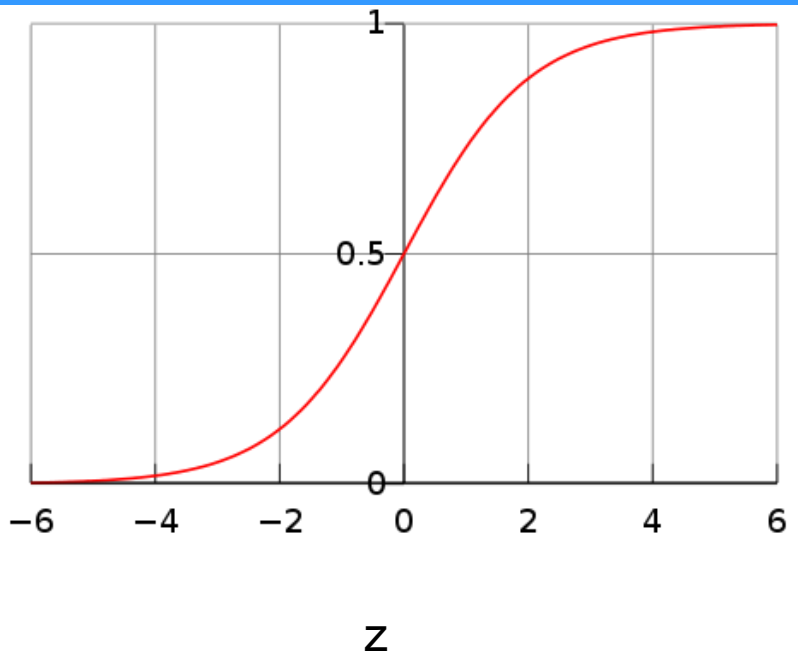
$$\vdots$$

$$\ln \frac{\Pr(y=K-1 \mid X=\mathbf{x})}{\Pr(y=K \mid X=\mathbf{x})} = \beta_{(K-1)0} + \boldsymbol{\beta}_{K-1}^T \mathbf{x}$$

Logistic Function

- Let us from now on consider K = 2 for the sake of simplicity.
- Classes are encoded by 0 or 1.

$$\Pr(y = 1 \mid X = \mathbf{x}) = \frac{1}{1 + \exp(\underbrace{-(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})}_{=:z})}$$



- This is called *logistic function*

Logistic Regression

- Suppose N data points / data vectors to be given
- Obervation: responses y_i can be supposed to be drawn from a conditional binomial distribution p(Y|X)

$$\Pr(Y = y_i \mid X = \mathbf{x}_i) = \begin{cases} \dfrac{1}{1+\exp(-z_i)} & y_i = 1 \\ 1 - \dfrac{1}{1+\exp(-z_i)} & y_i = 0 \end{cases}$$

- Probability for N data points (= likelihood function):

$$\ell(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{N} \underbrace{\left(\frac{1}{1+e^{-z_i}}\right)^{y_i} * \left(1 - \frac{1}{1+e^{-z_i}}\right)^{1-y_i}}_{\text{Binomial PMF}}$$

Parameter Estimation

- Like in linear regression we set

$$\boldsymbol{\beta} \leftarrow (\beta_0, \boldsymbol{\beta}^T)^T$$

$$\mathbf{x}_i \leftarrow (1, \mathbf{x}_i^T)^T$$

- We take the logarithm of likelihood function:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[y_i \ln \frac{1}{1+\exp(-z_i)} + (1-y_i) \ln\left(1 - \frac{1}{1+\exp(-z_i)}\right) \right]$$

$$= \sum_{i=1}^{N} \left[y_i \boldsymbol{\beta}^T \mathbf{x}_i - \ln(1+\exp(\boldsymbol{\beta}^T \mathbf{x}_i)) \right]$$

- We take the derivative and set it to 0:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \mathbf{x}_i \left(y_i - \frac{1}{1+\exp(-\boldsymbol{\beta}^T \mathbf{x}_i)} \right) = 0$$

- Cannot be solved algebraically ➔ numerical solvers.

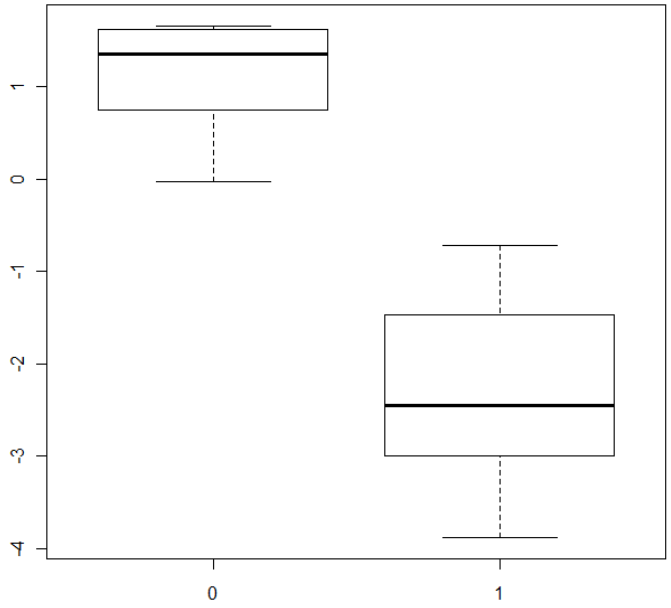
Model Fitting and Important Assumptions of Logistic Regression

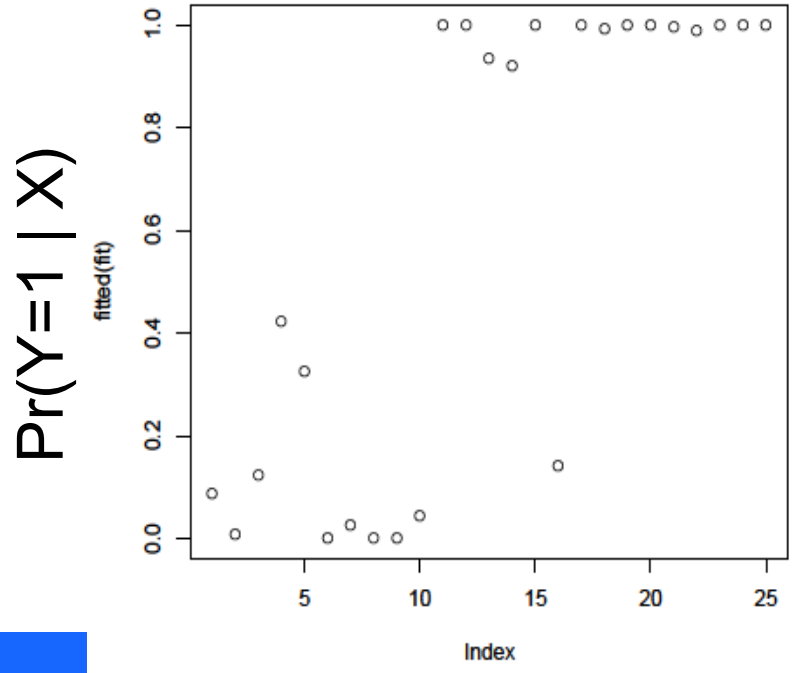
- Fitting of logistic regression models is done in an iterative process (e.g. Newton-Raphson method)
  1. Start with tentative solution
  2. Follow gradient to improve solution
  3. Continue until no sufficient further improvement can be obtained (*convergence*)

- Non-convergence may occur due to violuation of model assumptions:
  - Linear class separation possible
  - Not an error-free class separation
  - $p < n$ (fewer regressors than data points)
  - No colinearities among regressor variables

Logistic Regression in R

```
> X = c(rnorm(10,mean=1),
   rnorm(15,mean=-2))
> Y=factor(c(rep(0,10),rep(1,15)))
> fit=glm(Y~X,family=binomial)
> plot(fitted(fit))
> summary(fit)
```





Logistic Regression in R (cont.'d)

```
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.047418  -0.124378   0.006357   0.041119   1.984991

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5438     1.0009  -0.543   0.5870
X            -4.0887     2.2270  -1.836   0.0664 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33.6506  on 24  degrees of freedom
Residual deviance:  6.7687  on 23  degrees of freedom
AIC: 10.769

Number of Fisher Scoring iterations: 8
```

Residual quantiles

Coefficients and their significances

Model fit

Logistic Regression in R (cont.'d)

```
Null deviance: 33.6506  on 24  degrees of freedom
Residual deviance:  6.7687  on 23  degrees of freedom
AIC: 10.769
```

- „Null model" = model just containing the intercept
- Deviance = - 2 * model log-likelihood
- Hence:
  - 1st row = deviance of null model
  - 2nd row = deviance of built model

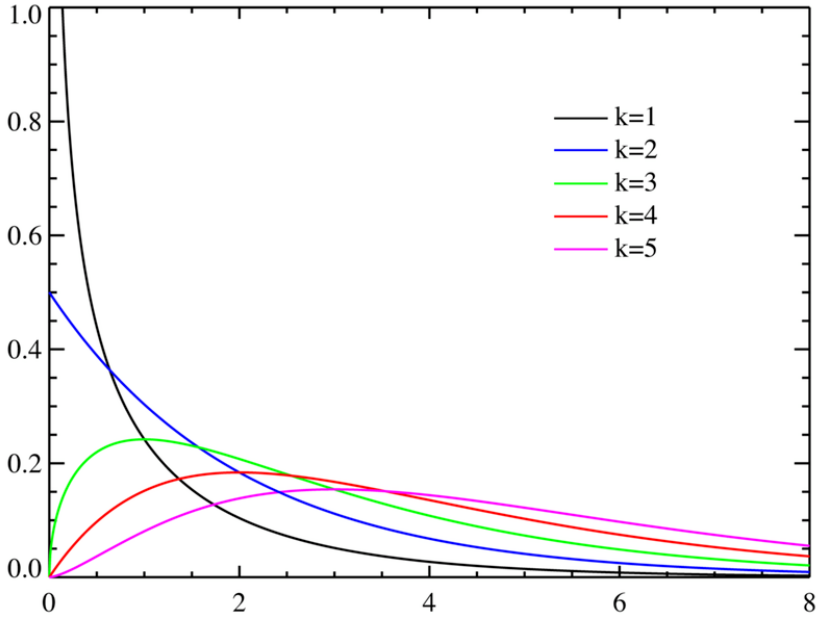
- Significance can be tested via a *likelihood ratio test*:
  - H0: likelihood (null model) >= likelihood (built model)
  - H1: built model has a higher likelihood
- Test statistic:

$$D = -2 \ln \frac{\text{likelihood for null model}}{\text{likelihod for built model}}$$

Asymptotically follows χ^2 distribution

Chi-Square Distribution

- The **chi-square distribution** (also **chi-squared** or ***χ*²-distribution**) with ***k* degrees of freedom** is the distribution of a sum of the squares of *k* independent standard normal random variables.
- It is one of the most widely used probability distributions in inferential statistics, e.g. in hypothesis testing or in construction of confidence intervals.



$$\chi_k^2(x) := \chi^2(x \mid k) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$

is the so-called *Gamma* function

Logistic Regression in R (cont.'d)

```
> anova(fit, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    24     33.651
X     1   26.809        23      6.841 2.246e-07
```

Likelihood of built model differs significantly from null model

Model Selection

- Summary of logistic regression contains an „AIC“ value in R.
- What does it mean?

- The „Akaike Information Criterion“ (AIC) is a measure of the relative goodness of fit of a statistical model.
  - AIC provides a means for comparison among (two or more) models (***model selection***)
  - It is NOT a statistical test, like the likelihood-ratio test
  - It is does NOT tell anything about the quality of the fit on an absolute scale.
- Definition:

$$AIC = 2*\#\text{parameters} - 2*\log\text{-likelihood}$$

- Select model with minimal AIC

Model Selection (cont.'d)

- Alternative to the AIC is the „Bayesian Information Criteration" (BIC)
- It penalizes the number of parameters higher than the AIC:

$$BIC = \log(N) * \#\text{parameters} - 2 * \log - \text{likelihood}$$

- All these approaches implement a general principal called **Occam's razor**: Given a set of competing models that equally well explain our data, select the one that makes the fewest assumptions / has the fewest parameters / is least complex.

Principal Component Analysis (PCA)

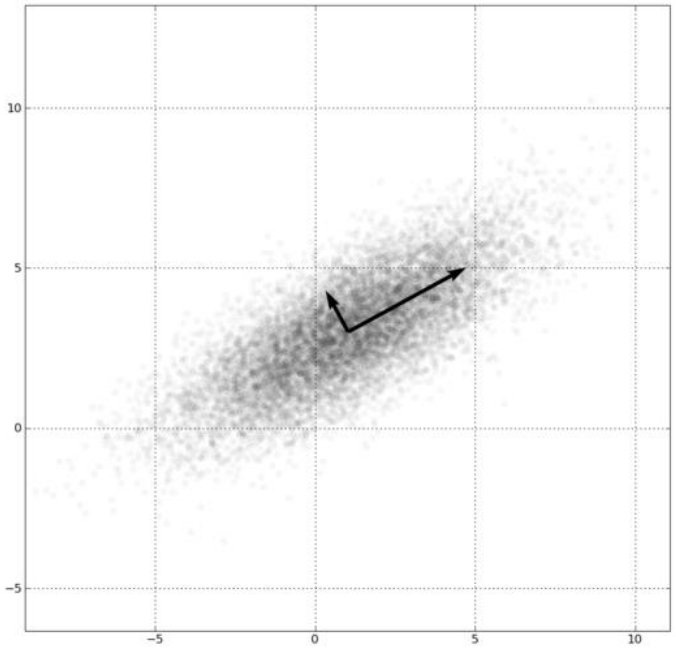
- **Goals:**
  - Dimensionality reduction of multi-dimensional data
  - Visualization of high-dimensional data (2D, 3D)
  - De-correlation of variables in case of jointly normally distributed data

- Example: Is there a way to visualize this data in 2D?

```
> head(USArrests)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

Approach

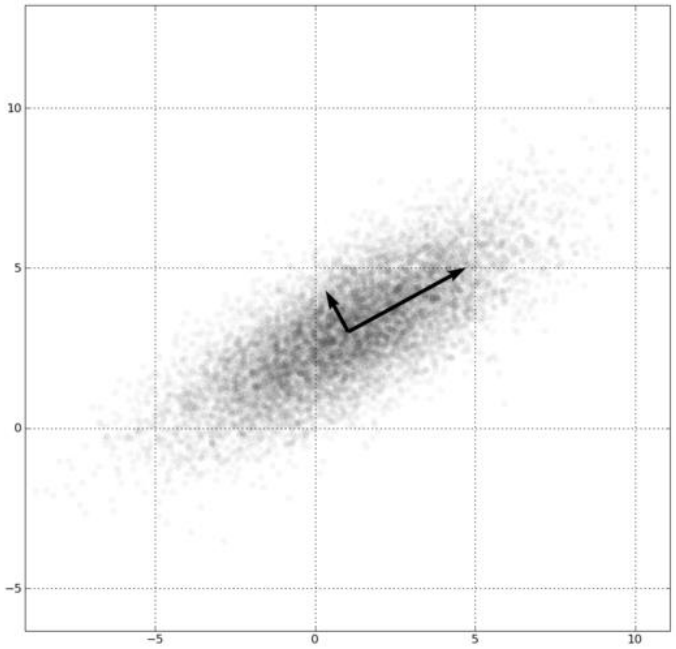
- Assumption: data is organized in a n x d matrix **X**
  - Rows = observations
  - Columns = variables (mean centered)



- **Idea:**
  1. Look for direction of maximal variance in our data (= 1st principal component)
  2. Look for direction of second highest variance in our data (= 2nd principal component)
  3. ...
  4. Project data on first principal components

How does PCA work?

- How do we find these principal components?



- One way: PCs = eigenvectors corresponding to largest eigenvalues of the data covariance matrix Σ

$$\Sigma v = \lambda v \quad v \in \Re^n, \lambda \in \Re$$

- Vector **v** is called eigenvector for eigenvalue λ
- There are d eigenvectors and eigenvalues alltogether

Example

```
> head(USArrests)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
> pr = prcomp(USArrests, scale = TRUE)

> pr$rotation
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

Meaning of eigenvalues

- Mathematical insight: Larger eigenvalue ➔ more variance explained

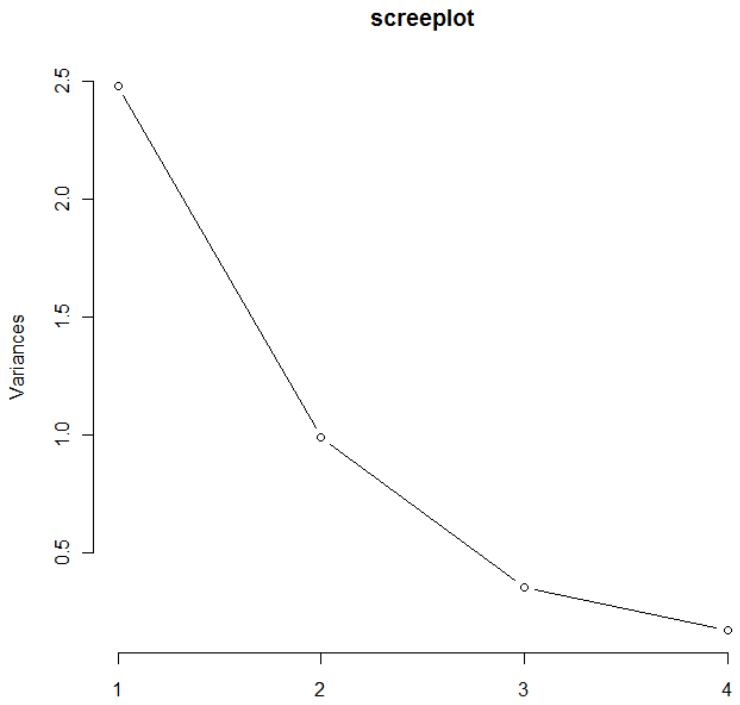
- How much of the total variance is explained by m components?

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

explained variance

total variance

- Heuristics for dimensionality reduction:
  - choose suitable cutoff (i.e. 95%) ➔ 2 – 3 principal components here
  - Look at steepest decrease in scree plot.



Component / eigenvector number

```
> screeplot(pr, type="lines", main="screeplot")
```

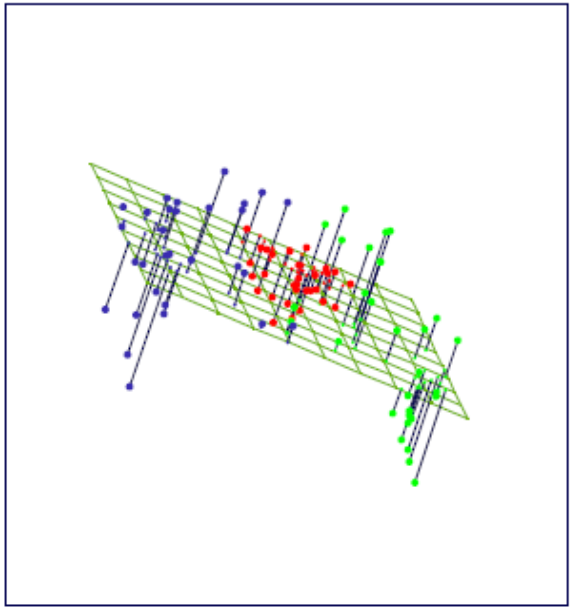
Example

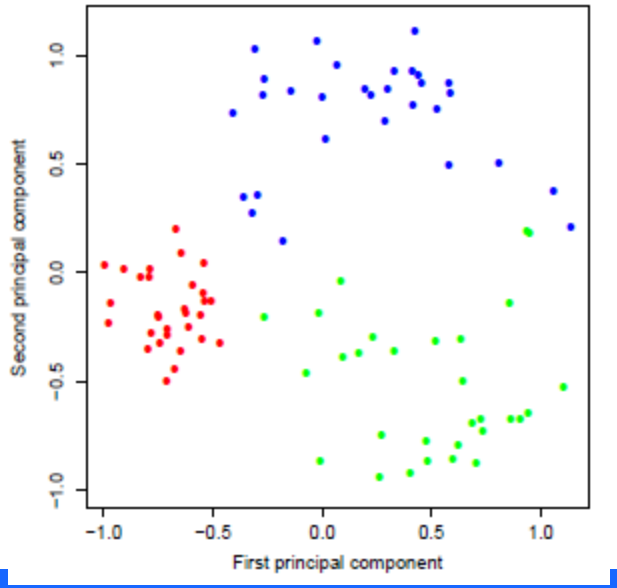
```
> pr$rotation
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432

> summary(pr)

Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.57 0.995 0.5971 0.4164
Proportion of Variance 0.62 0.247 0.0891 0.0434
Cumulative Proportion  0.62 0.868 0.9566 1.0000
```

Projection of Data into Lower Dimensional Subspace

- Let $\mathbf{V}_m$ be the matrix, whose columns contain the m first principal components
- We can project our data $\mathbf{X}$ onto these principal components via:

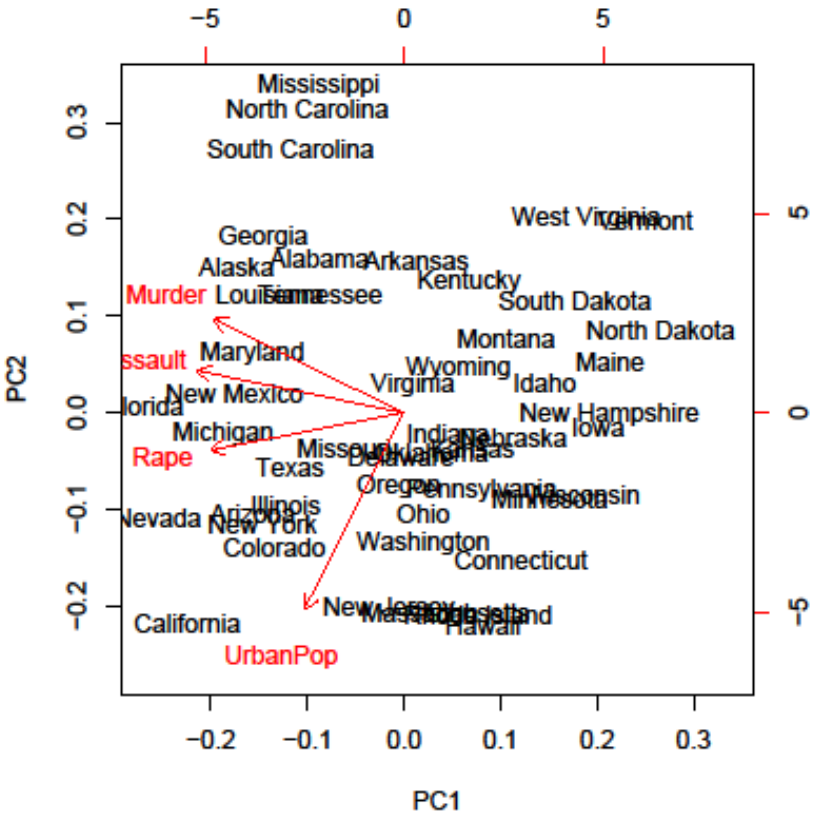
$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{V}_m$$





2D PCA-Plot

```
> biplot(pr)
> pr$rotation # prints matrix of
    eigenvectors = rotation
    matrix
> head(pr$x[,1:2])
                    PC1         PC2
Alabama     -0.9756604  1.1220012
Alaska      -1.9305379  1.0624269
Arizona     -1.7454429 -0.7384595
Arkansas     0.1399989  1.1085423
California -2.4986128 -1.5274267
Colorado    -1.4993407 -0.9776297
```



red arrows = original axis

Features of PCA

- Advantages:
  - Dimensionality reduction: fewer variables, suppression of irrelevant information
  - Low dimensional (visual) data representation
  - Principal components are linearly uncorrelated

- Drawbacks:
  - Choice of number of PCs
  - Low dimensional projection not guaranteed to be distance preserving
  - PCs are usually not interpretable. They are just linear combinations of existing variables.

- Further considerations:
  - PCA is a **linear** projection technique: non-linear structures not detectable
  - The core idea behind PCA is that the directions of maximal variance also contain maximal information.

What you should know and be able to apply

- Probability:
  - What is probability?
  - How can we compute with probabilities (→ Conditional probability, joint probability, Bayes‘ law)?
- Random variables:
  - What is a random variable and for what reasons is it needed?
  - Which types of random variables do exist (→ discrete / continous)?
  - What is a PDF, a CDF and a PMF? What are the differences between those? Examples of PDFs and PMFs?
  - How can we compute with random variables? (→ conditional density, joint density, Bayes law)
  - Being able to apply rules on examples!
- Parameter estimation:
  - Why do we need parameter estimation?
  - Which mechanisms for parameter estimation are there (→ ML, Bayes’ law) and how do they work in principle?
  - Being able to apply them on a simple example

What you should know and be able to apply

- Hypothesis tests:
  - What is the logic behind a hypothesis test and how does it work principally?
  - What is a p-value?
  - Being able to tell, which test should be used in a particular situation
- Multivariate Linear Modeling Methods:
  - What is the purpose of linear regression and how does it work in principle (→ ML estimate minimizes sum of squared residuals)? Under which circumstances can it be applied (→ normal distribution of residuals, more observations than variables, no colinearities)?
  - What is an ANOVA and for which purposes is it applied? How does it work in principle?
  - What is the purpose of logistic regression (→ classification) and how does it work in principle (→ modeling of log odds ratios as a linear function of predictors)? Under which circumstances can it be applied (→ more observations than variables, linear class separation, no colinearities)
  - What is the purpose of PCA and how does it work in principle? How can the number of relevant PCs be determined?
  - Being able to apply correct approach in an example and interpret results