Grundpraktikum Bioinfo - Week 1

Biological Data Analysis

Carl Herrmann
IPMB - Universität Heidelberg

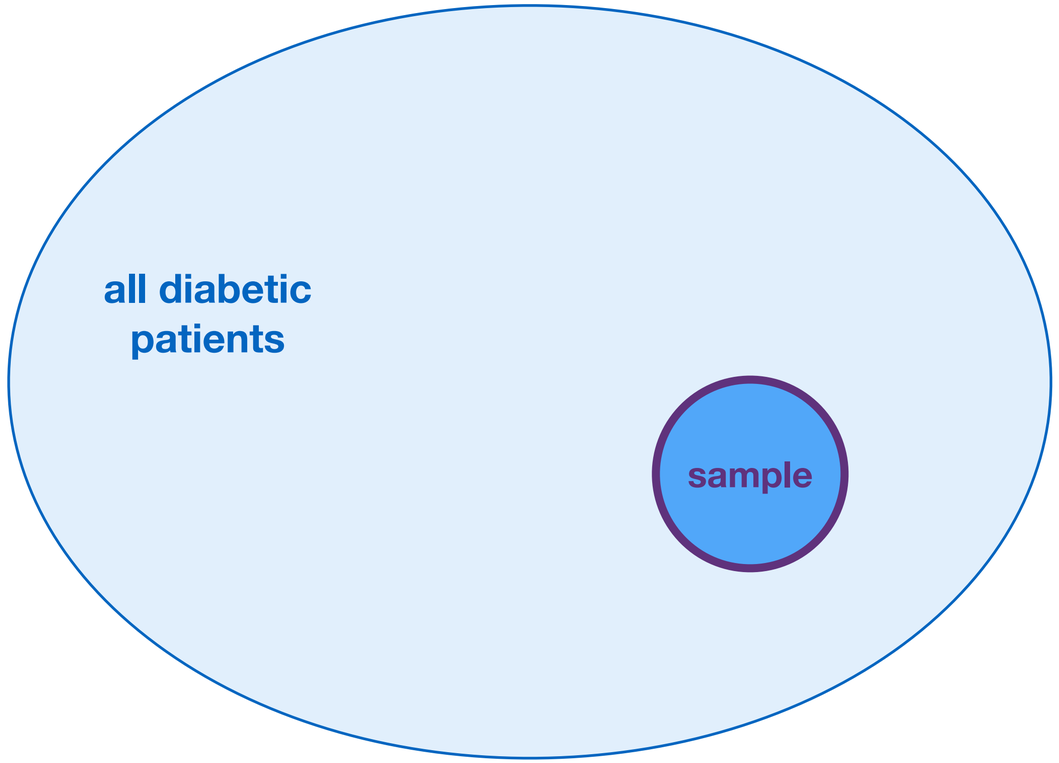
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

5. Distributions

Principles of statistical inference

- can we learn general principles from a small sample?
- Example: opinion polls / clinical study / ...



Is the sample representative of the general population?

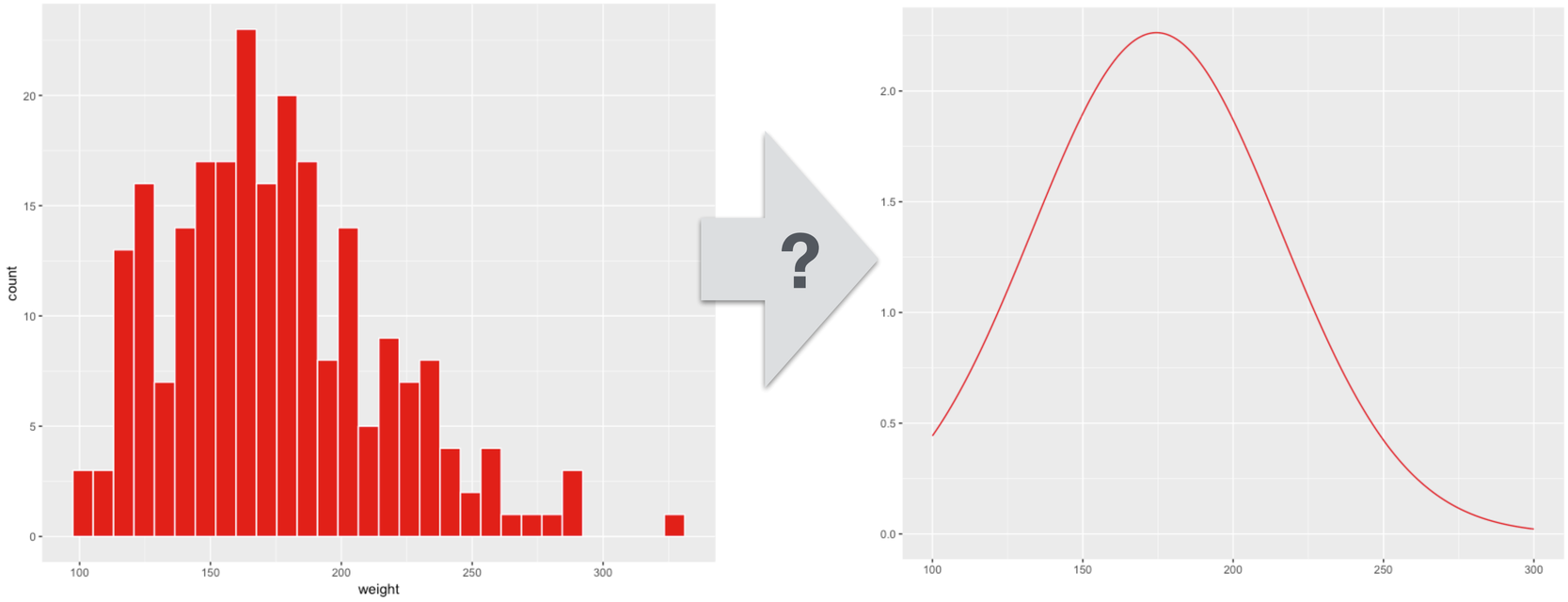
<u>Assumption</u>:

sample is drawn from the same distribution as the general population!

Diabetic patients

sample

population?



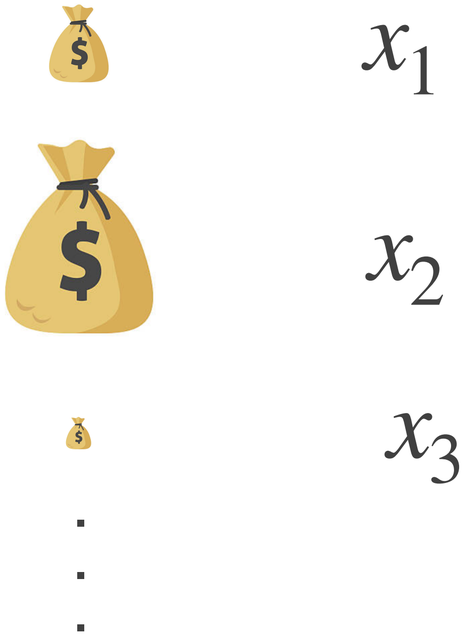
inference: How can we estimate the parameters (mean, spread,…) of the population based on the sample?

Random variables

- a **random variable** *X* describes a **stochastic process**

Random variable X

Realizations x_i



x_1

x_2

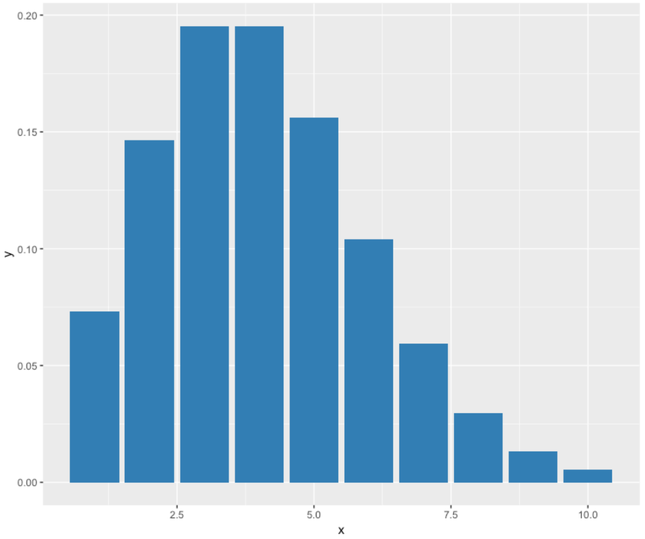
x_3

Random variables

- Random variable X can be described using
  - **expectation *E(X)***: “theoretical" mean of the entire population
  - **Variance *Var(X)***: “theoretical" variance of the entire population
  - **Density distribution** (continuous RV) *f(X)*
  - **Probability distribution** (discrete RV) *p(X)*

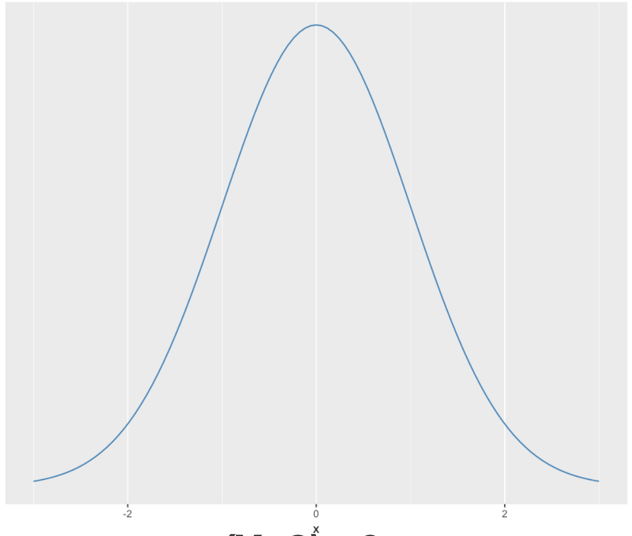
Discrete probability distribution

$$\sum_{i \in A} p(X_i) = 1$$



p(X=2) : probability that X takes value 2

Continuous density distribution



$$\int_{x \in A} f(x)dx = 1$$

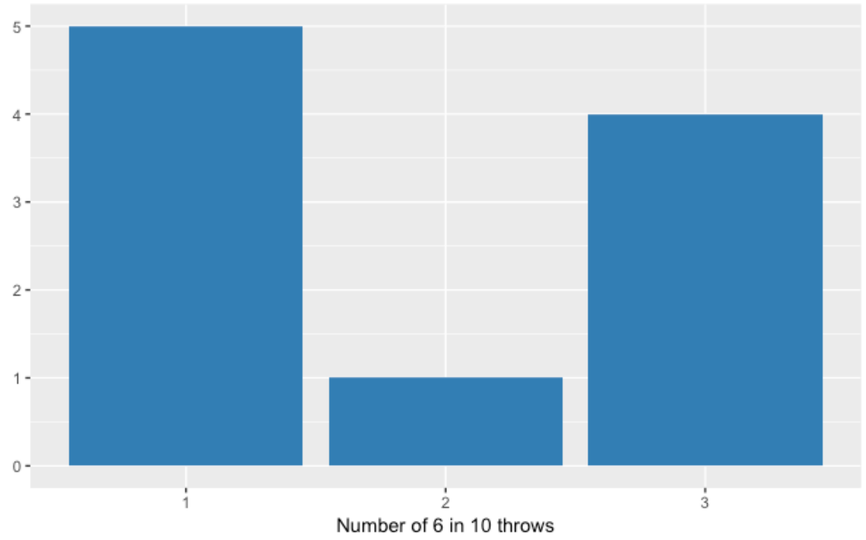
p(X=2) =0
p(2 < X < 3) >0

Example of a discrete RV

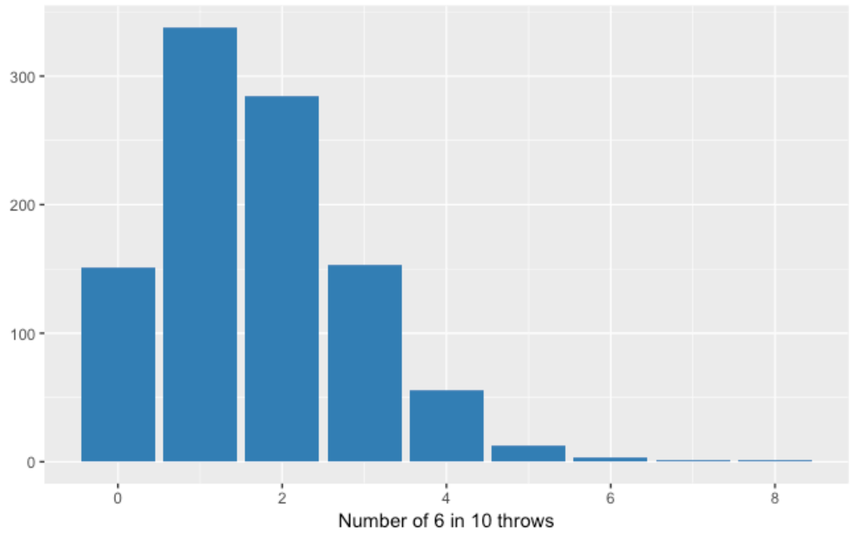
- Random variable X : number of 6 obtained in 10 dice throws

$$X \in \{0,\ldots,10\}$$

10 games

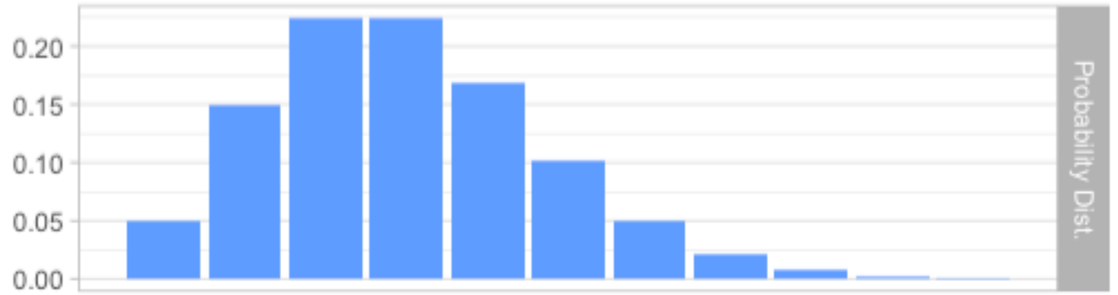


1000 games



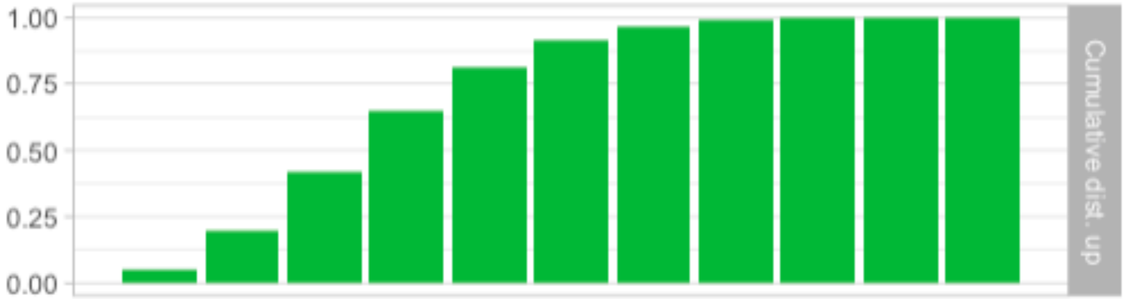
Cumulative distribution function (CDF)

- Example: **discrete RV**
- Probability distribution *P(k)*



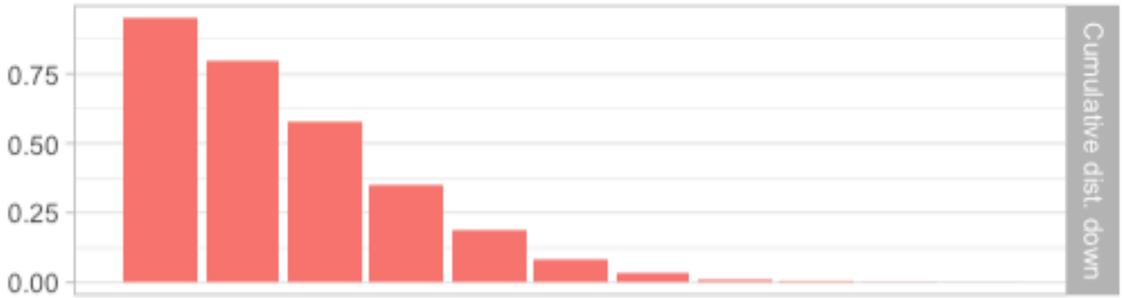
  - Cumulative distribution

$$F_X(k) = P(x \leq k)$$



  - Complementary cumulative distribution

$$\bar{F}_X(k) = p(x > k) = 1 - F_X(k)$$

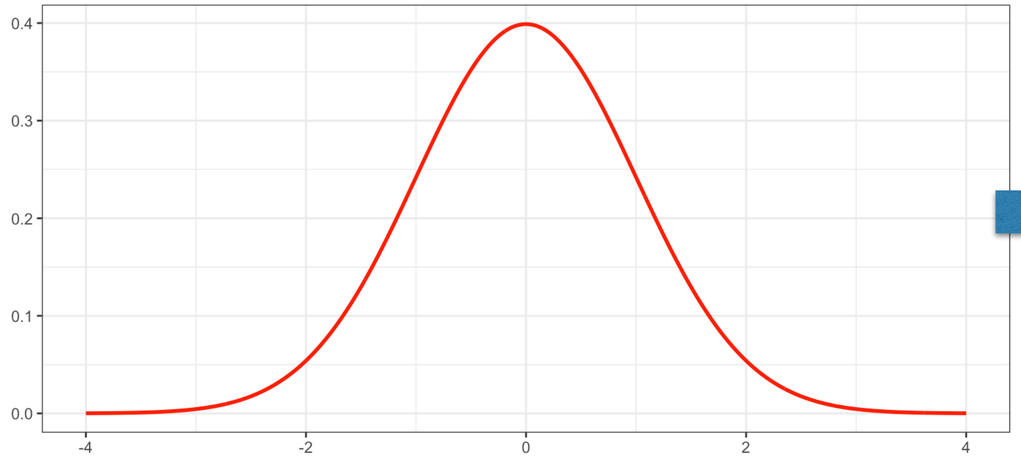


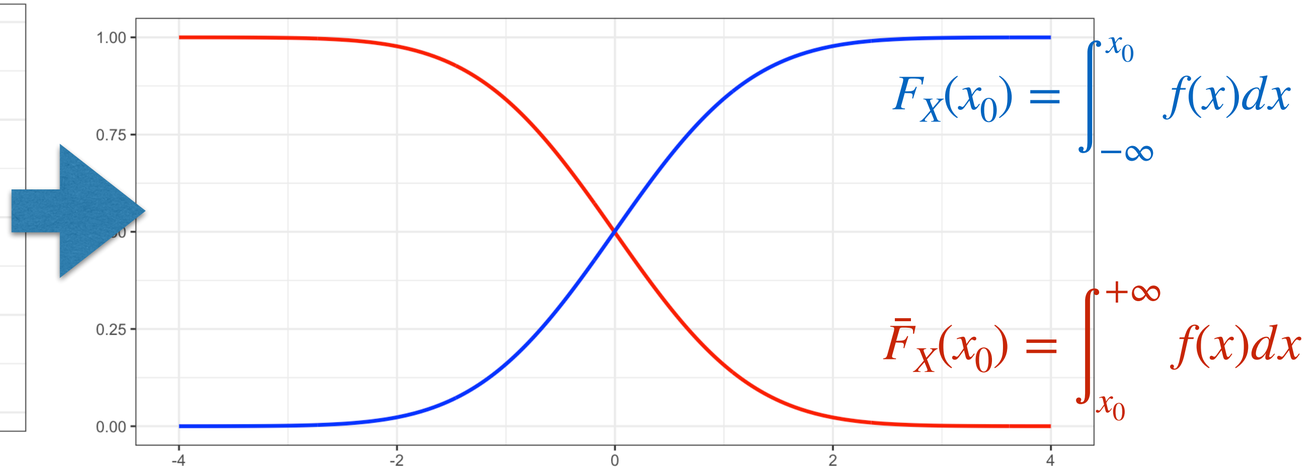
Cumulative distribution (CDF)

- **Continuous** random variable X
- density distribution *f(X)*
  - cumulative distribution
  - complementary cumulative distribution

$$F_X(x_0) = P(x < x_0) = \int_{-\infty}^{x_0} f(x)\,dx$$

$$\bar{F}_X(x_0) = P(x > x_0) = \int_{x_0}^{+\infty} f(x)\,dx$$





Cumulative distribution of the uniform distribution?

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Discrete distributions

Binomial distribution

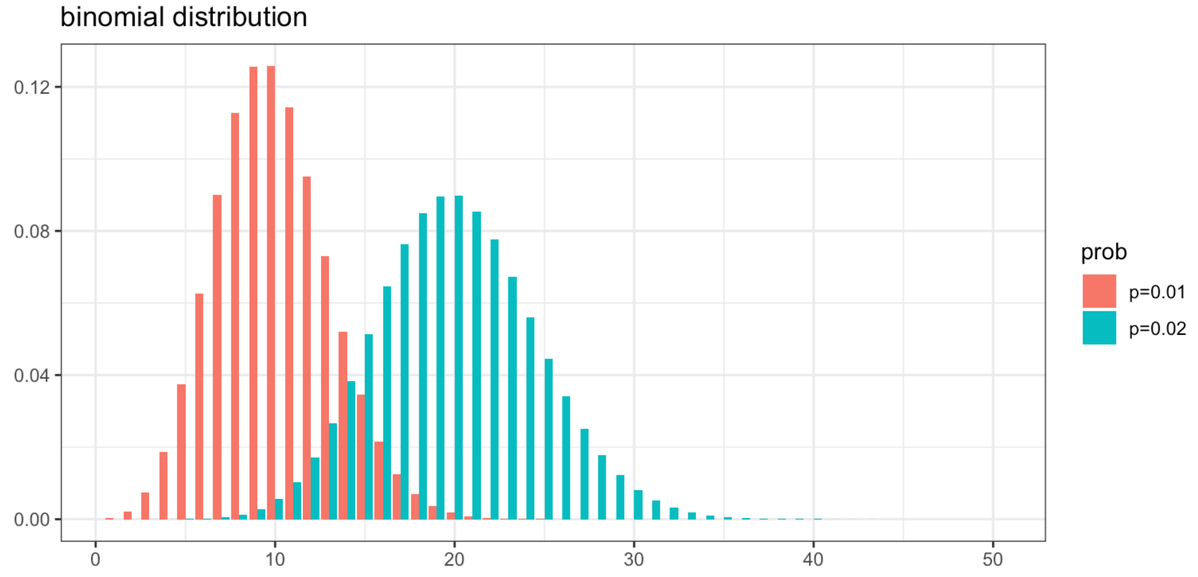
- Number of successes in ***L*** independent trials, each with probability of success ***p***
- Example of DNA polymerase
  - error rate at each position ***p*** (constant, independent of the previous position)
  - sequence of length ***L*** → number ***k*** of errors (= “successes”) ?

$$p(k) = \binom{L}{k} p^k (1-p)^{L-k}$$

$$E(X) = Lp$$

$$Var(X) = E(X)(1-p)$$

(how to compute the cumulative distribution?)



(what is L here?)

Poisson distribution

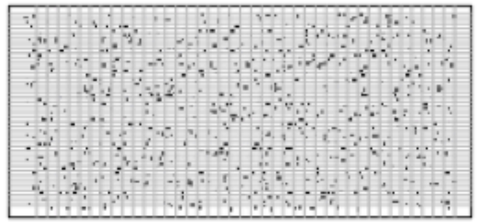
- Measures the number of events in a time period for a given **rate** of events
- Example: number of drops per tile during a rainshower
- Rate needs to be constant and independent of previous period!

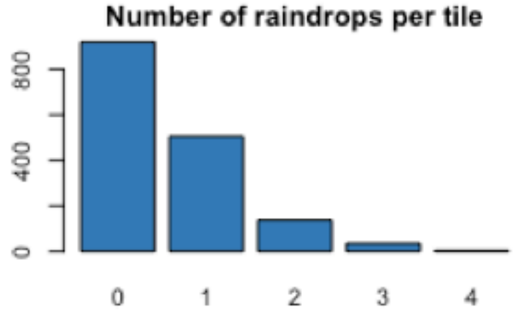
$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$E(X) = \lambda$$

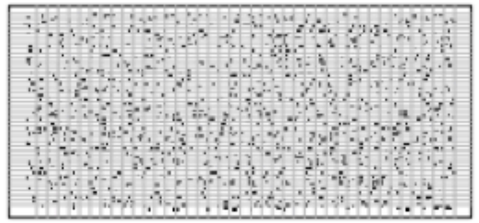
$$Var(X) = \lambda$$

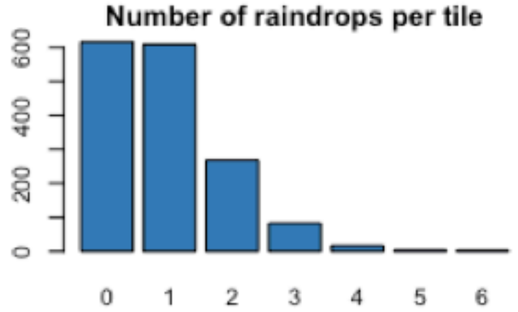
$\lambda = 0.56$



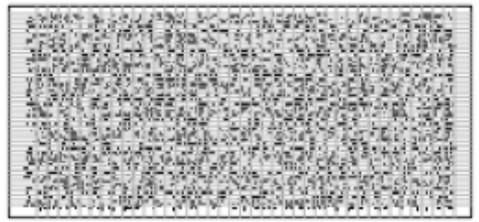


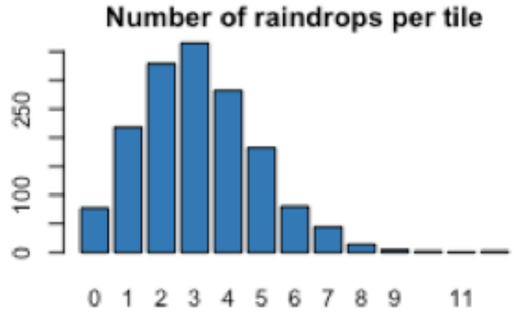
$\lambda = 0.94$





$\lambda = 3.125$





λ = drop per tile per unit time

Negative binomial

- Independent trials, constant probability p
- Different questions can be asked
  - ***Number of successes in L trials?***
    → **Binomial distribution** with 2 parameters *(L, p)*
  - ***Waiting time until r successes reached?***
    → **Negative binomial distribution** with 2 parameters *(r, p)*
  - Example: Taq-polymerase has accuracy (1-p) with p ~ 1/900 (one error every 900 bases)
    → length distribution of sequences with r = 3 errors?

Negative binomial

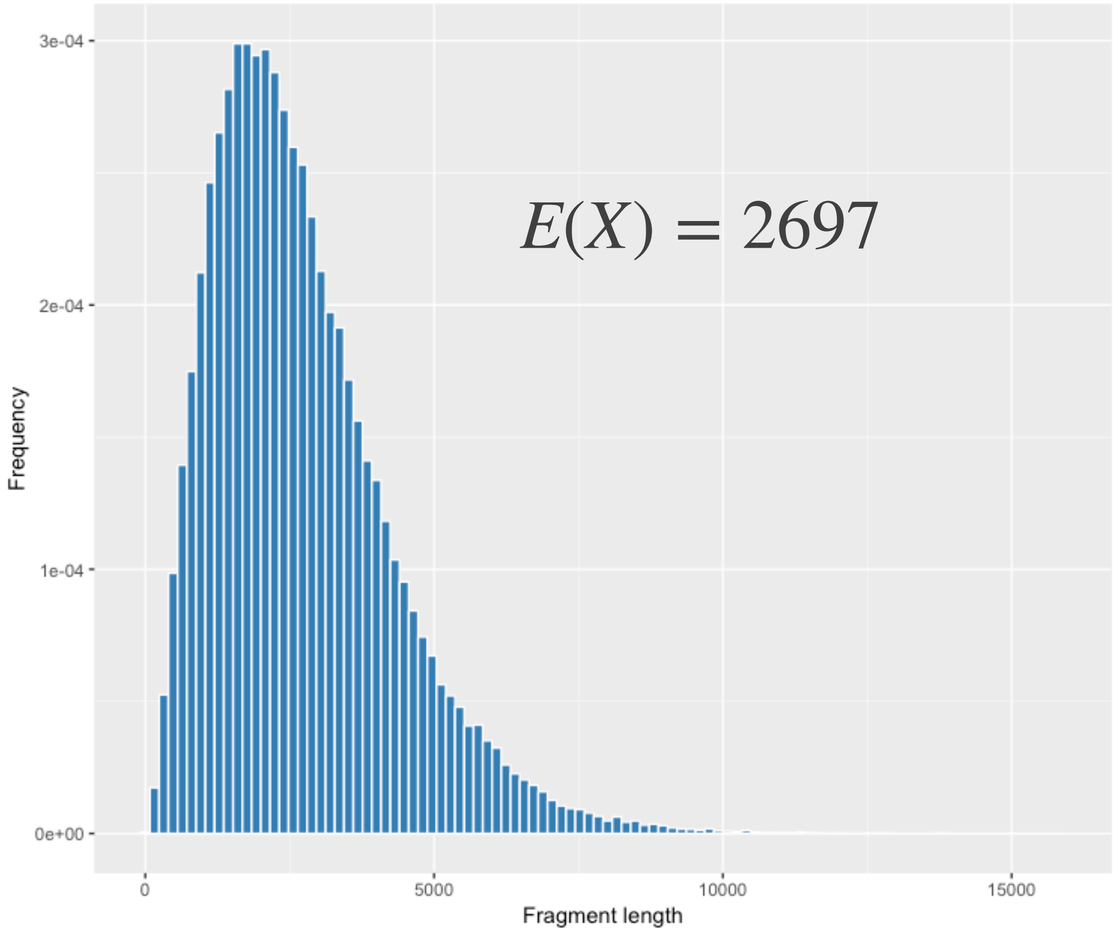
$$p(k) = \binom{k+r-1}{r-1} p^r (1-p)^k$$

$$E(X) = \frac{(1-p)r}{p}$$

$$Var(X) = \frac{E(X)}{p} = \frac{(1-p)r}{p^2}$$

k = number of error-free bases

p = 1/900, r = 3 errors

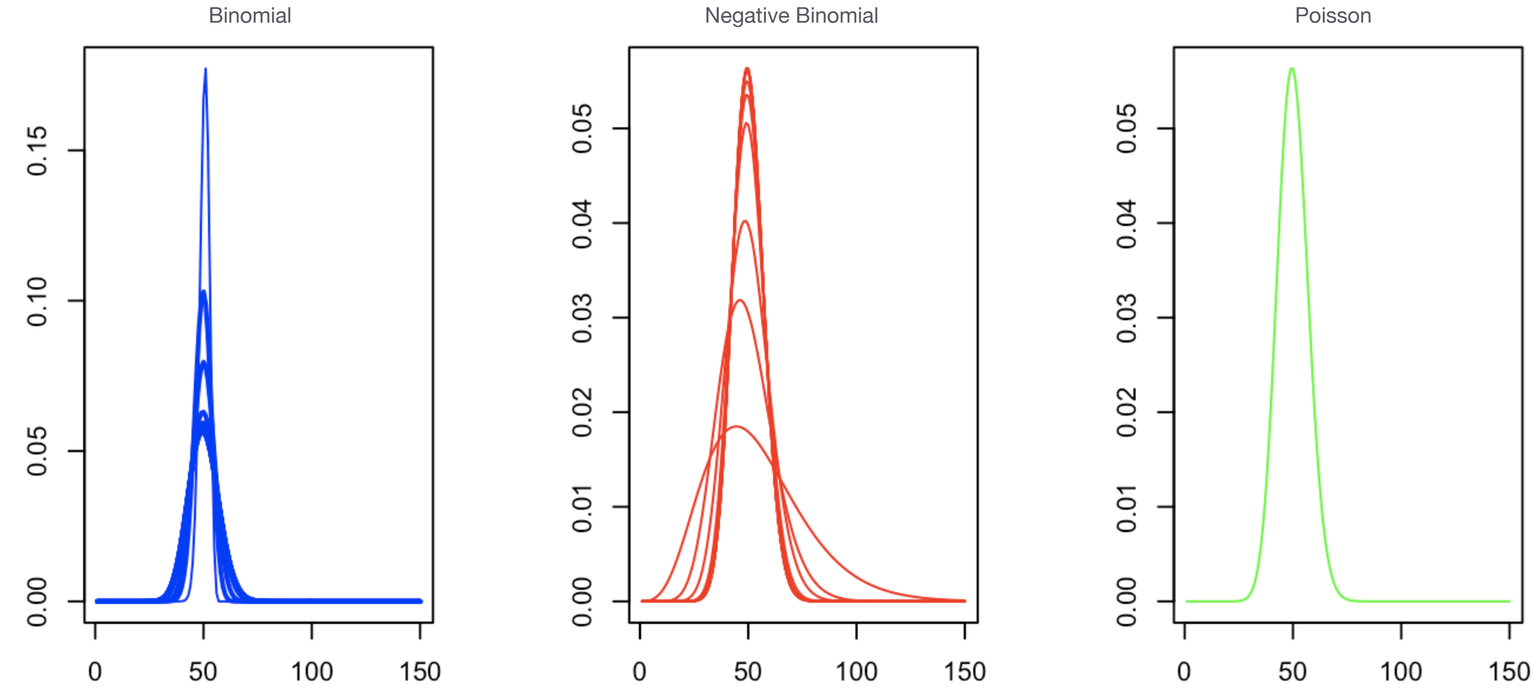


Negative binomial

- The variance of the NB distribution can be made **arbitrarily wide**, by making p smaller

$$Var(X) = \frac{E(X)}{p} = \frac{(1-p)r}{p^2}$$

- This can be usefull in modelling processes showing an **over-dispersion**



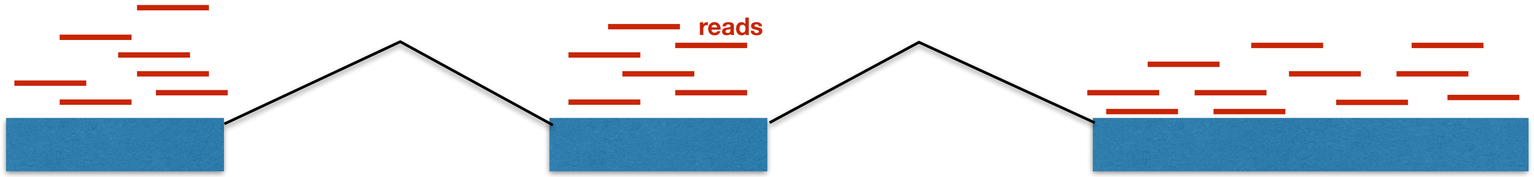
$$Var(X_B) = (1-p)E(X_B)$$

$$Var(X_{NB}) = \frac{E(X_{NB})}{p}$$

$$Var(X_P) = E(X_P)$$

Over-dispersion

- RNA-seq: mRNA molecules are fragmented and sequenced as short “reads” (e.g. 75-150 bp)
- reads are mapped onto the genome
- for each transcript, the number of reads is recorded



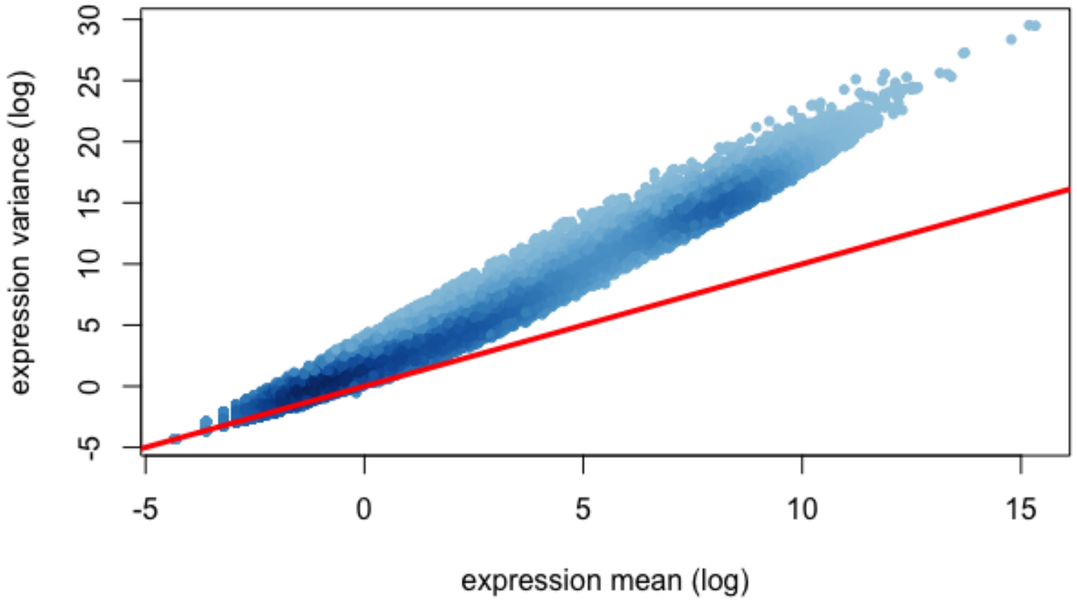
samples / replicates

transcripts

Description	GTEX-1117F-0226-SM-5GZZ7				GTEX-111CU-1826-SM-5GZYN				GTEX-111FC-0226-SM-5N9B8				GTEX-111VG-2326-SM-5N9I		
DDX11L1 3	4	1	1	0	2	1	3	5	1	0	1	3	1	1	3
WASH7P 616	395	826	364	301	419	340	451	424	331	304	430	508	414	911	505
MIR1302-11	2	1	1	0	1	0	2	3	1	1	1	0	0	1	0
FAM138A 1	0	1	1	0	0	2	3	2	0	1	2	0	2	1	2
OR4G4P 0	0	0	0	0	0	2	1	0	0	0	0	0	1	1	0
OR4G11P 0	2	2	0	0	1	0	0	0	1	0	0	1	1	1	2
OR4F5 0	0	0	0	2	0	0	0	1	1	0	0	2	1	1	0
RP11-34P13.7	8	3	12	12	2	4	10	9	3	14	9	6	4	1	12
CICP27 11	29	9	18	5	5	7	4	11	18	11	12	9	9	18	14
AL627309.1	264	300	114	1364	227	276	518	52	91	105	558	354	91	195	587
RP11-34P13.15	0	2	1	1	1	1	2	0	0	1	0	0	1	3	3
RP11-34P13.14	5	0	0	16	4	9	9	0	0	2	16	10	1	8	12
RP11-34P13.13	57	31	24	104	31	69	104	18	65	18	116	34	23	82	146
RNU6-1100P	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
RP11-34P13.9	1	0	0	1	1	0	0	0	0	1	1	0	0	0	0
AP006222.2	52	113	168	7	100	46	25	29	44	256	35	152	79	159	511
RP4-669L17.10	18	15	47	42	54	42	23	33	20	26	20	38	15	17	128
RP4-669L17.8	1	1	2	2	6	0	3	0	3	0	0	0	1	1	2
CICP7 5	6	5	2	2	1	2	4	4	4	2	7	11	8	3	13
RP4-669L17.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OR4F29 0	1	0	0	0	3	0	1	0	0	0	0	0	2	1	2
WBP1LP7 0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	2
AL732372.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP4-669L17.2	0	0	0	1	1	0	0	2	1	1	0	1	1	0	0
RP5-857K21.15	3	0	4	2	2	1	2	2	2	2	0	1	2	2	2
RP4-669L17.1	3	3	1	0	2	2	1	0	5	2	2	6	1	2	5

RNA-seq data

- each dot is a transcript; several replicates (~patients) available
- x-axis : mean number of reads per transcript (over all replicates)
- y-axis : variance of number of reads per transcript (over all replicates)
- **variance is larger than mean**: cannot be described by Poisson or binomial process!



if number of reads per transcript would be a Poisson distribution, dots should lie on the y = x line

Continuous distributions

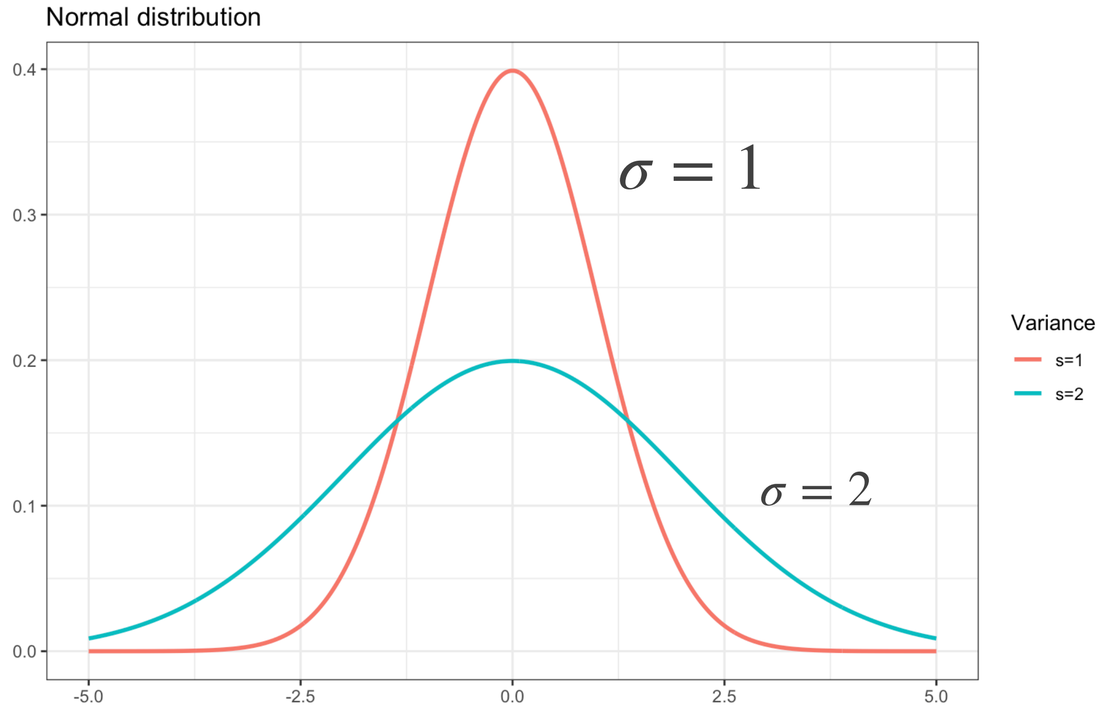
Normal distribution

- Normal distribution plays a central role in statistics and data analysis due to *Central Limit Theorem*
- 2 parameters: expectation $\mu$and standard deviation σ

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

$$E(X) = \mu$$

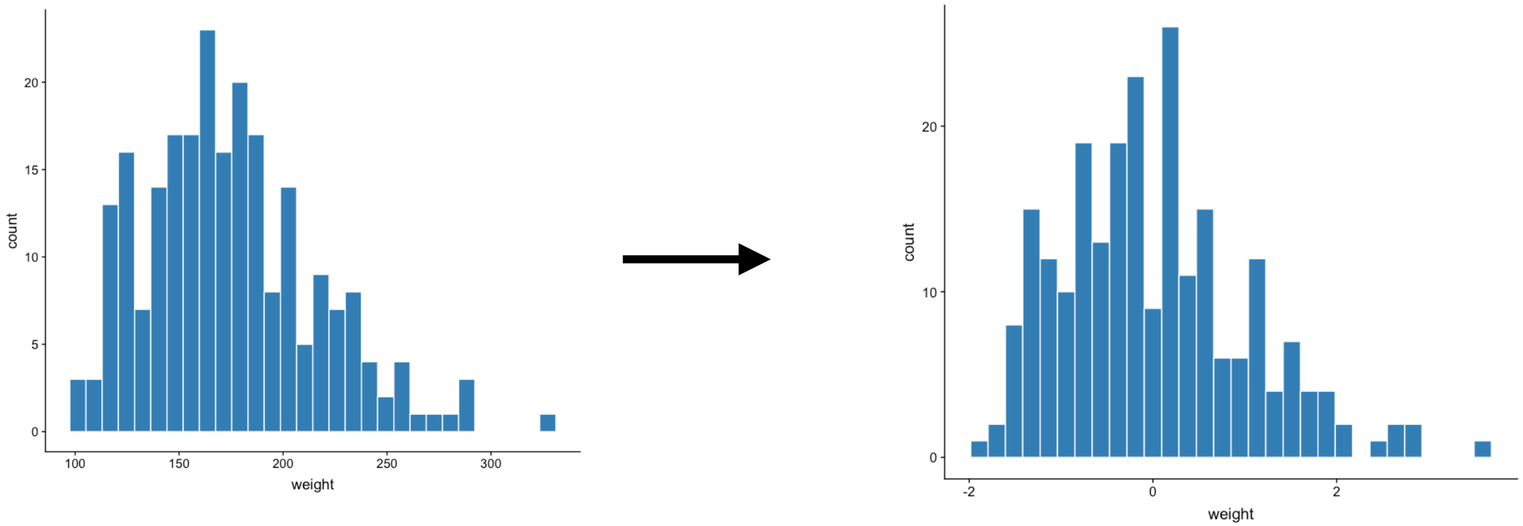
$$Var(X) = \sigma^2$$



Normal distribution

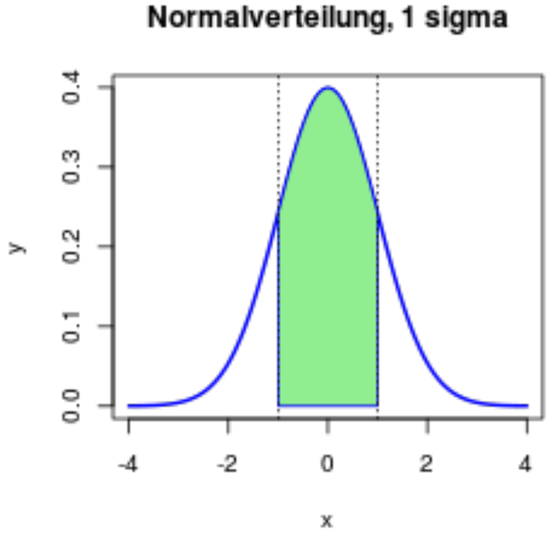
- The normal distribution with $\mu = 0$ and $\sigma = 1$ is called the **Standard Normal distribution** (SND) $\mathcal{N}(0,1)$
- Every normal distribution $\mathcal{N}(\mu, \sigma)$ can be transformed into the SND through a Z-transformation
  By definition, the new RV Z has expectation 0 and variance 1
- The Z-transformation can be applied to any distribution!

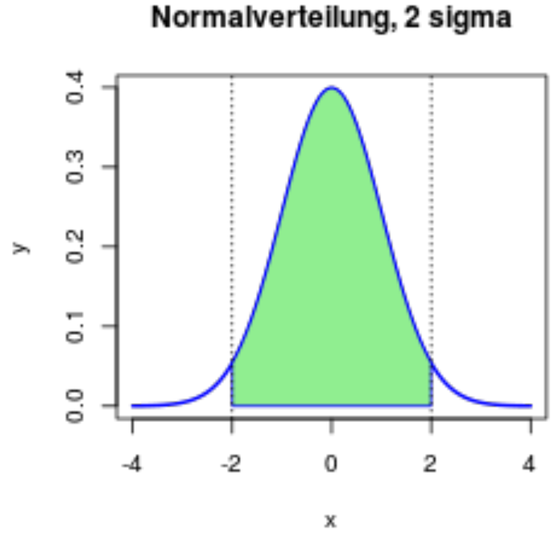
$$X \longrightarrow Z = \frac{X - \mu}{\sigma}$$



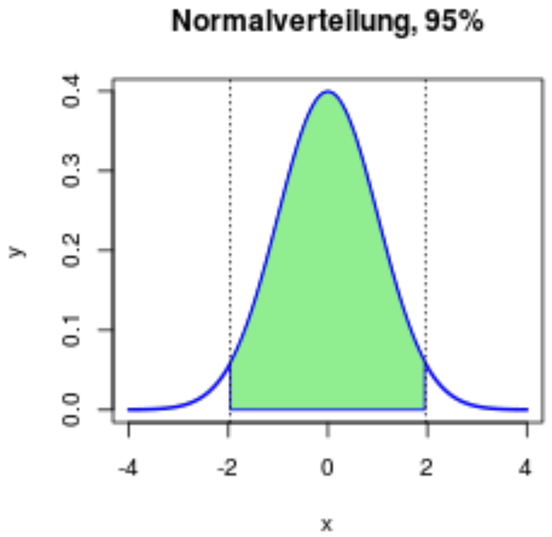
Properties of SND

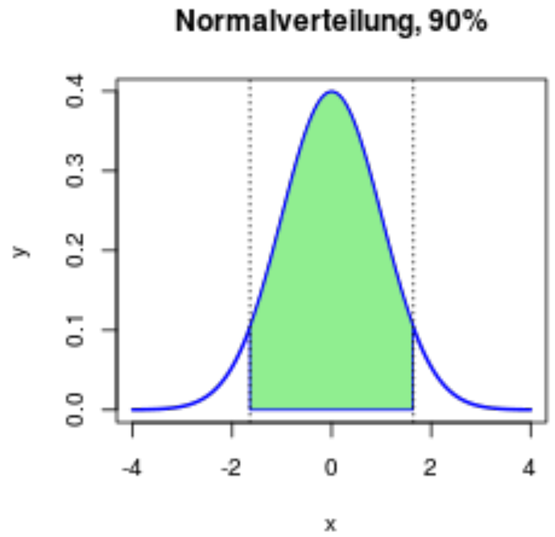
- the area represents…
  - [-1 , +1] = 68%
  - [-2 , +2] = 95.4%
  - [-1.96 , +1.96] = 95%
  - [-1.64 , +1.64] = 90%
- Critical values
  - t_{95} = 1.96
  - t_{90} = 1.64





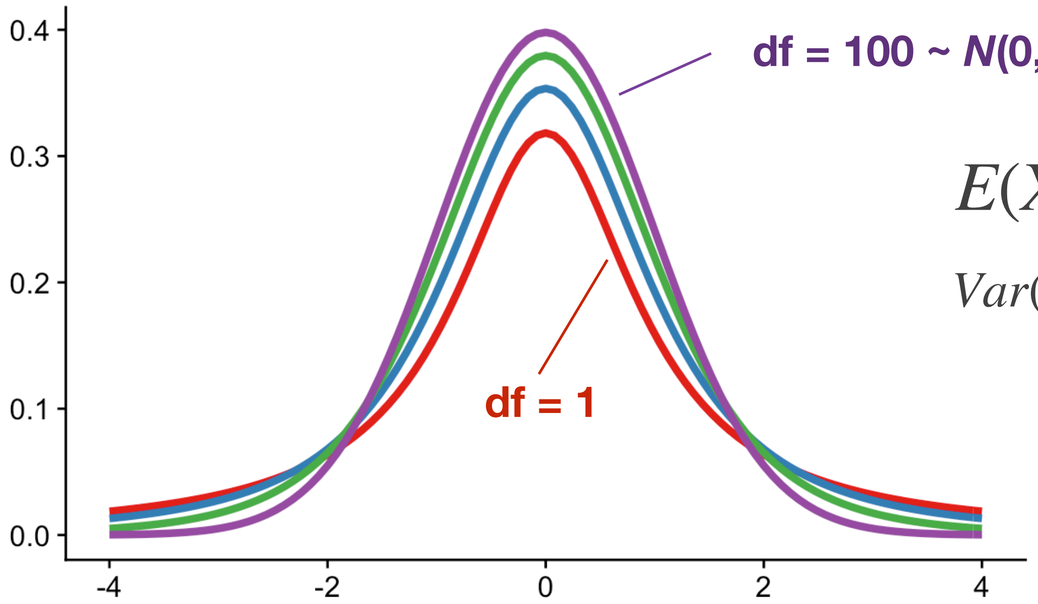
$$\int_{-\infty}^{+\infty} f(X)dX = 1$$





t-distribution

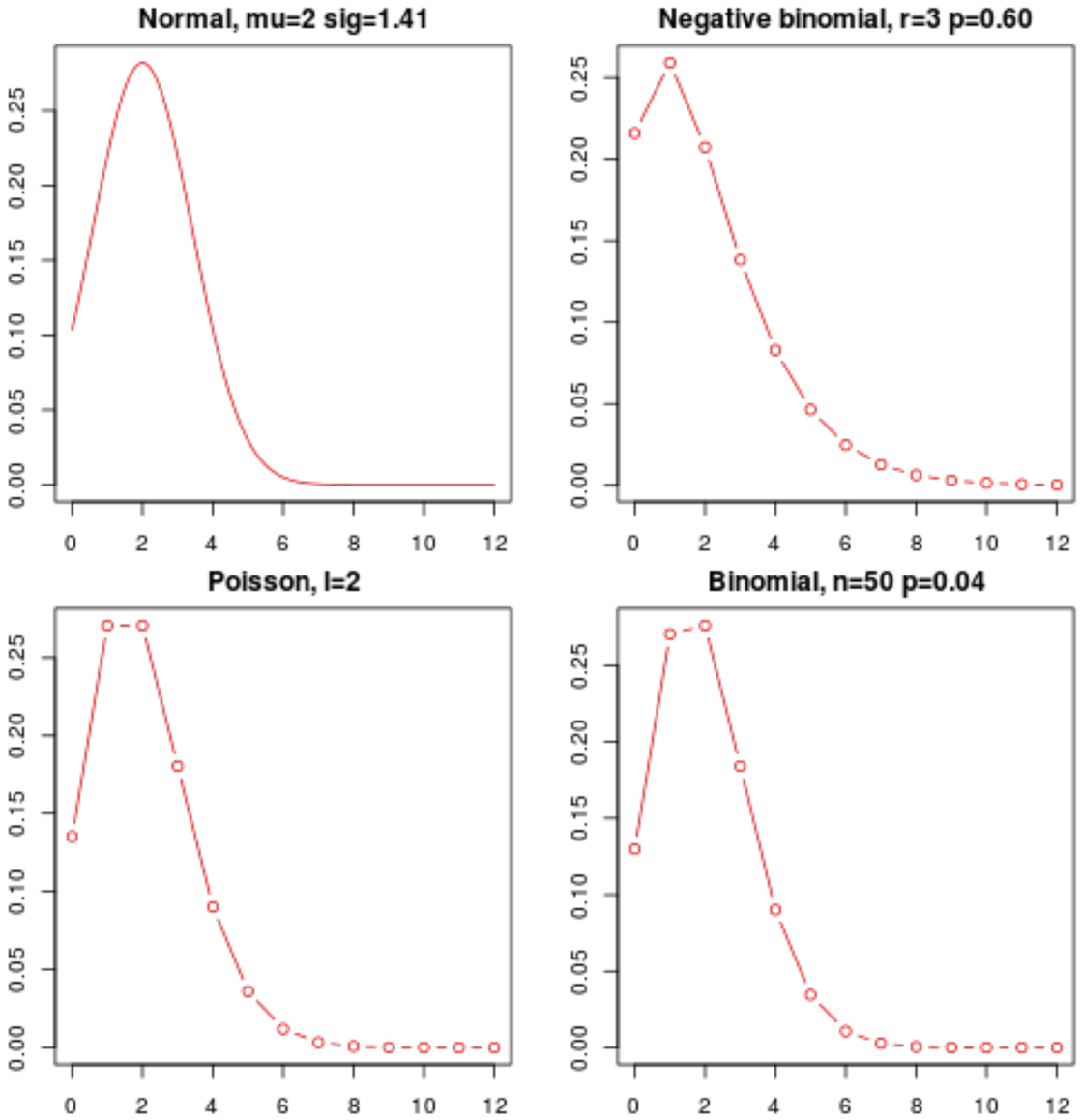
- Student's t-distribution is a continuous distribution describes the **distribution of mean values over small samples**
- Parameter: *degree of freedom (df)*
- Tends to the SND for large number of degrees of freedom



$$E(X) = 0$$

$$Var(X) = \frac{df}{df - 2} \quad \text{for} \quad df > 2$$

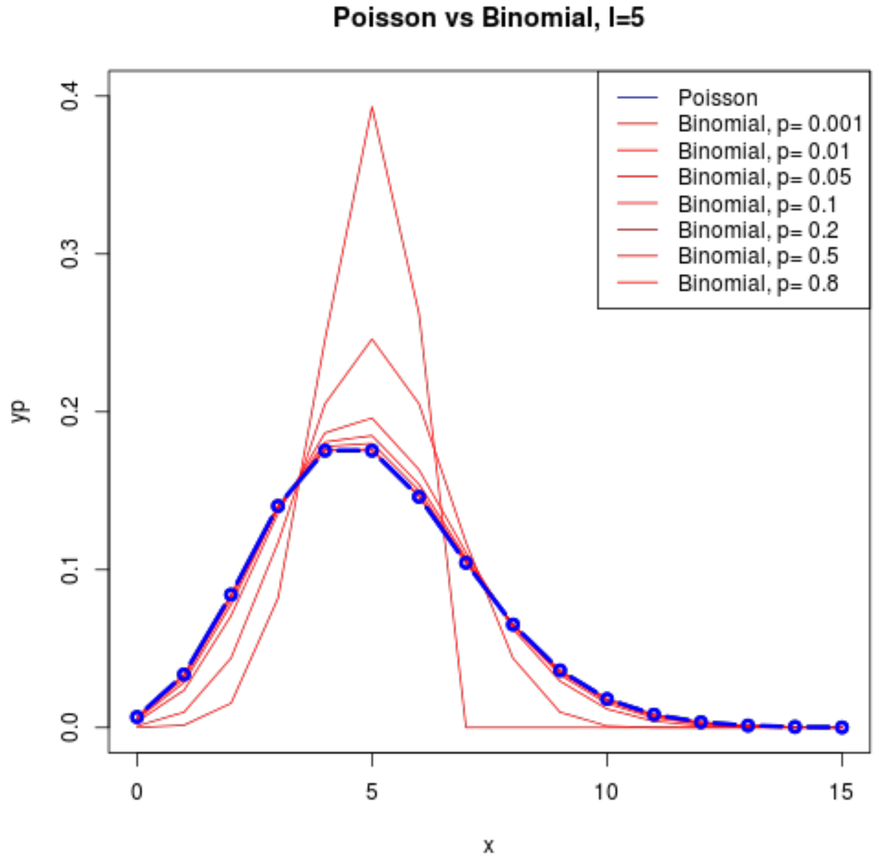
Related distributions



Poisson vs. binomial

$$Binom(n, p) \longrightarrow Pois(\lambda = np)$$

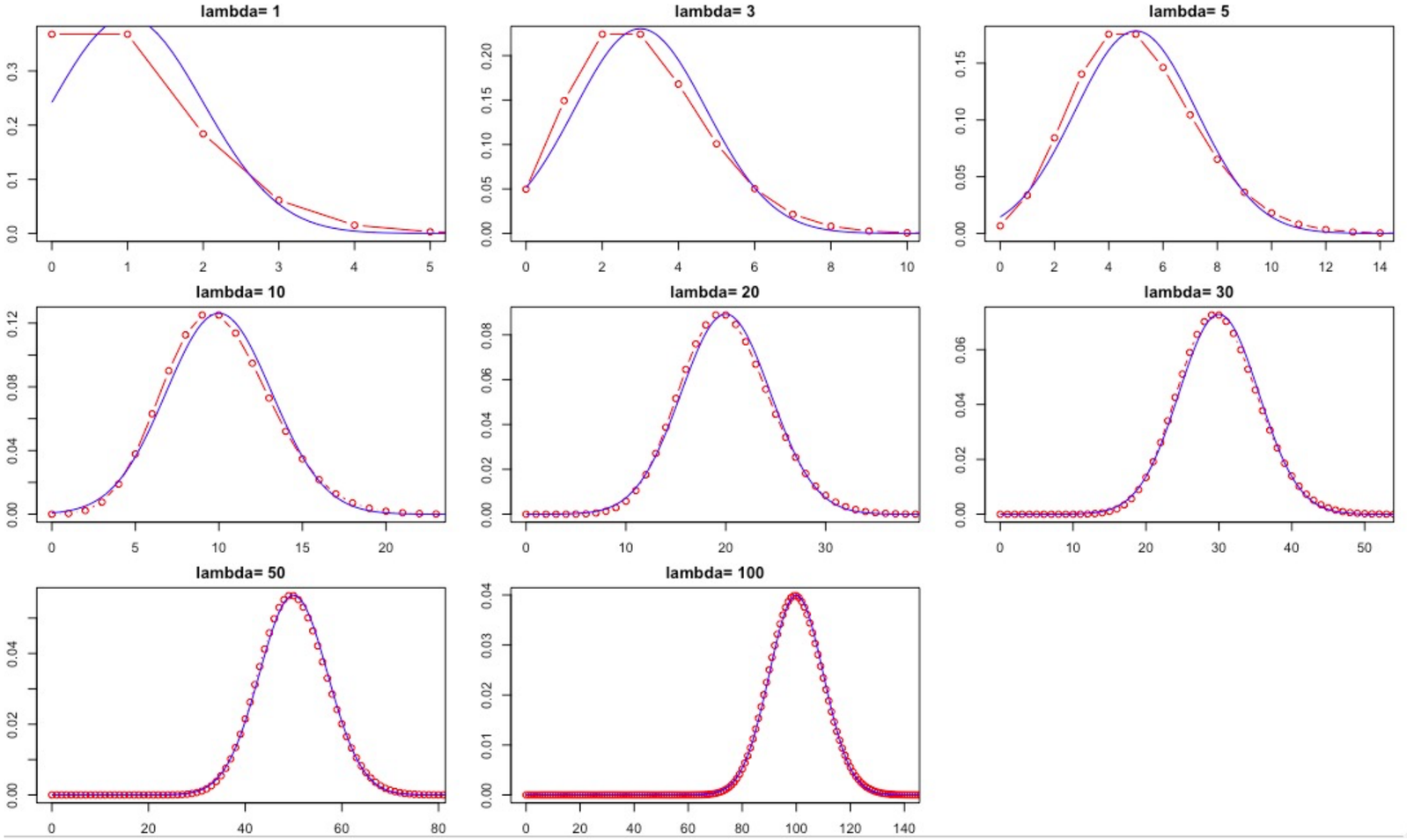
$$n \to \infty; p \to 0; np = \lambda$$



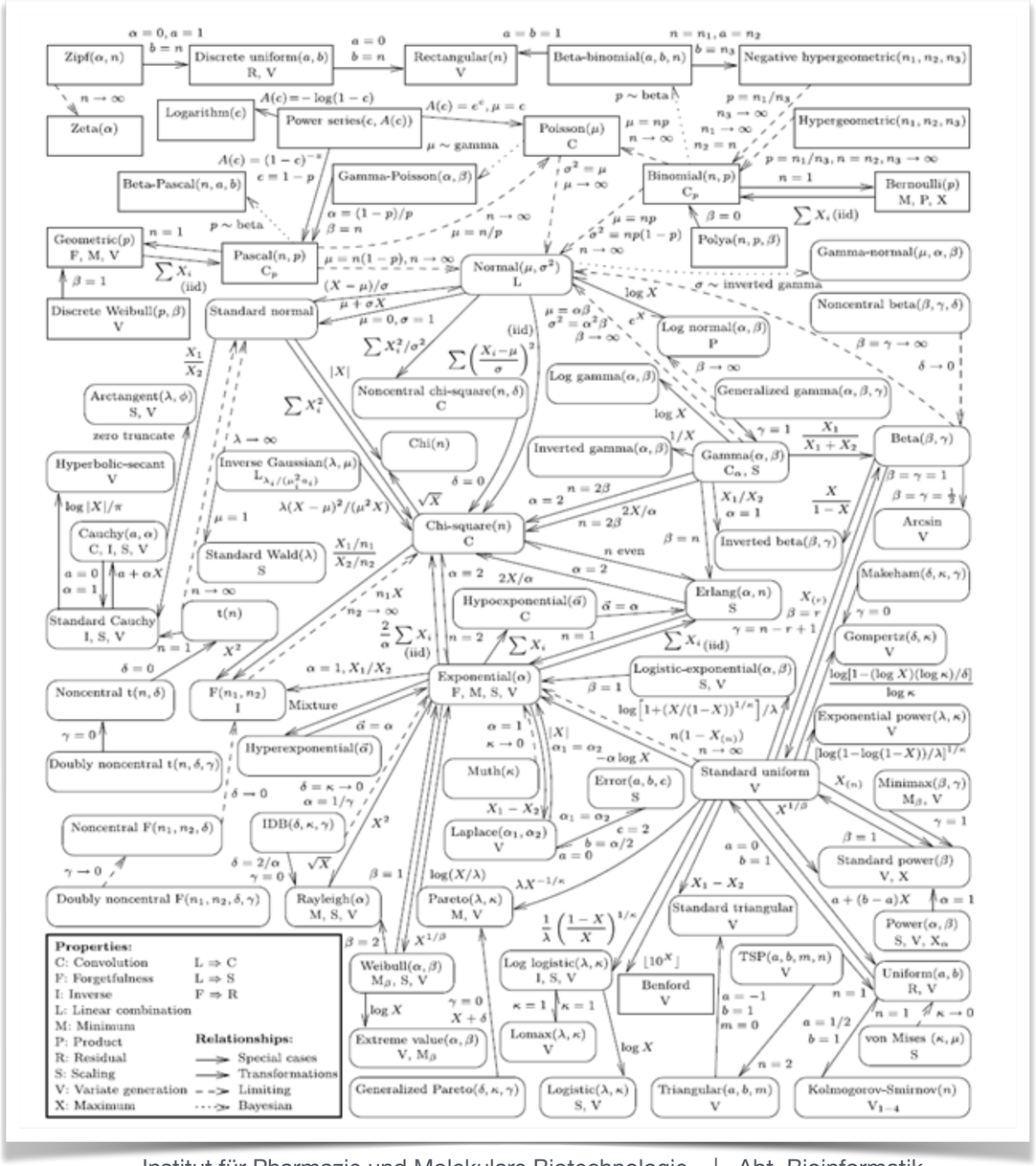
Poisson vs. normal

$$Pois(\lambda) \longrightarrow \mathcal{N}(\lambda, \sqrt{\lambda})$$

$$\lambda \gg 1$$



Distribution = italian wedding party!

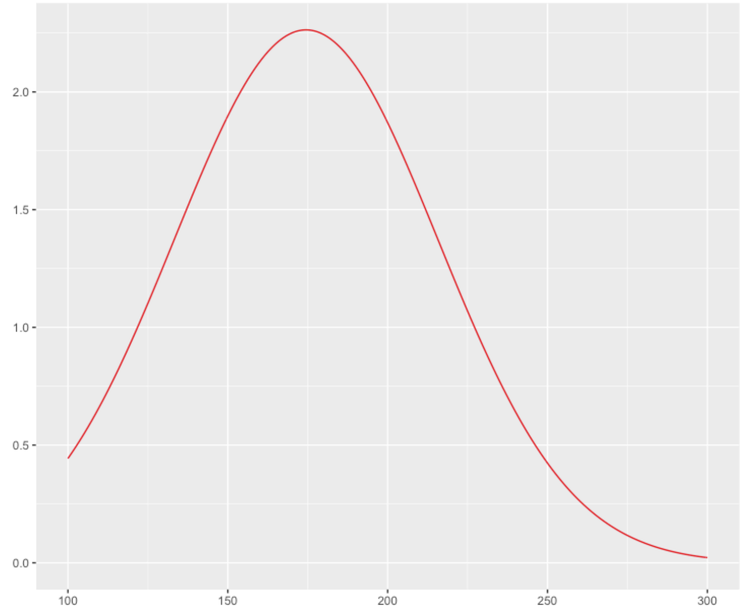


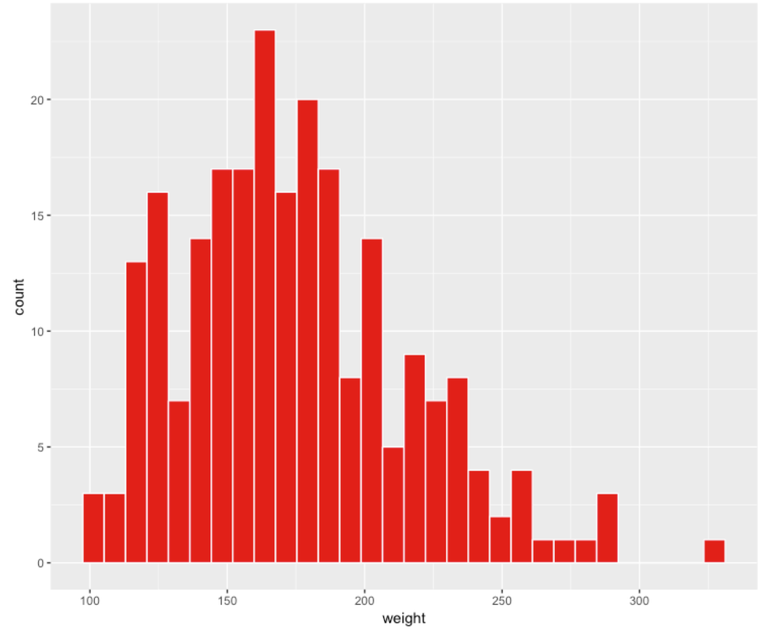
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Statistical inference

Statistical inference





How can we estimate the parameters of the population (expectation, variance, ..) …

… using the sample distribution?

Random variables and samples

- Parameters associated to the **random variable** are indicated in **greek letters**
- Parameters associated to **sample** are in **roman letter**

Random variable	Sample
random variable X	realizations $x_1, x_2, \ldots, x_n$
probability distribution p(k) density distribution f(X)	histogram
Expectation $E(X) = \mu$	Mean / median $\bar{x} = m$
Variance $Var(X) = \sigma^2$	Sample variance $Var(x_i) = s^2$

Can we use m and s to estimate μ and σ ?

Infering the expectation

- Example: **number of children per family**
  → we conduct a survey and interview every day *n* families
- Random variable : ***N_i* = {number of children in the i-th family}**
- The N_i are random variables, independent of each other, and identically distributed (“i.i.d. = independent, identically distributed")

$$M_n = \frac{1}{n} \sum_{i=1}^{n} N_i$$

- M_n is a **new random variable**
  → estimate its probability distribution p(k), expectation and variance

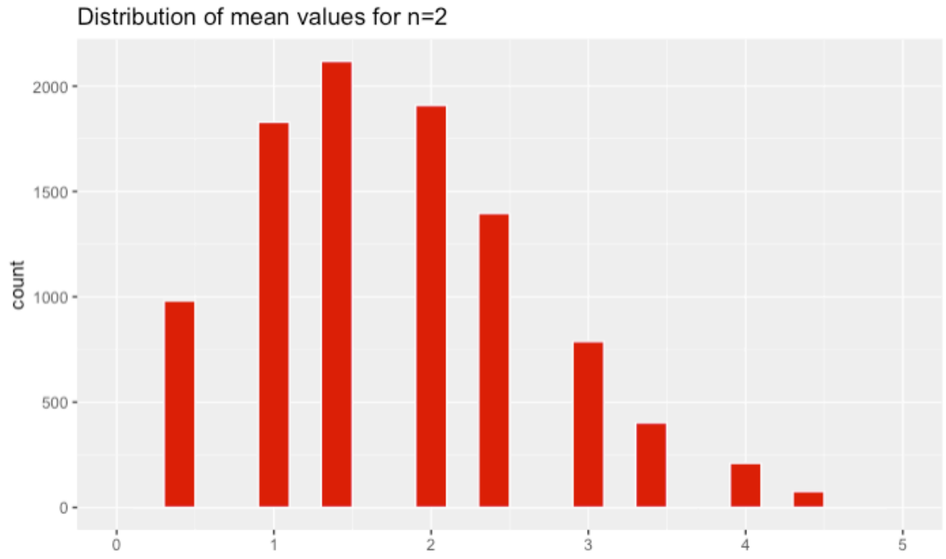
Example n=2

	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7	Day_8	Day_9	Day_10
Family_1	1	0	4	4	3	2	4	2	0	0
Family_2	3	2	1	1	2	2	4	3	1	1

...

	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7	Day_8	Day_9	Day_10
m_2	2	1	2.5	2.5	2.5	2	4	2.5	0.5	0.5

...



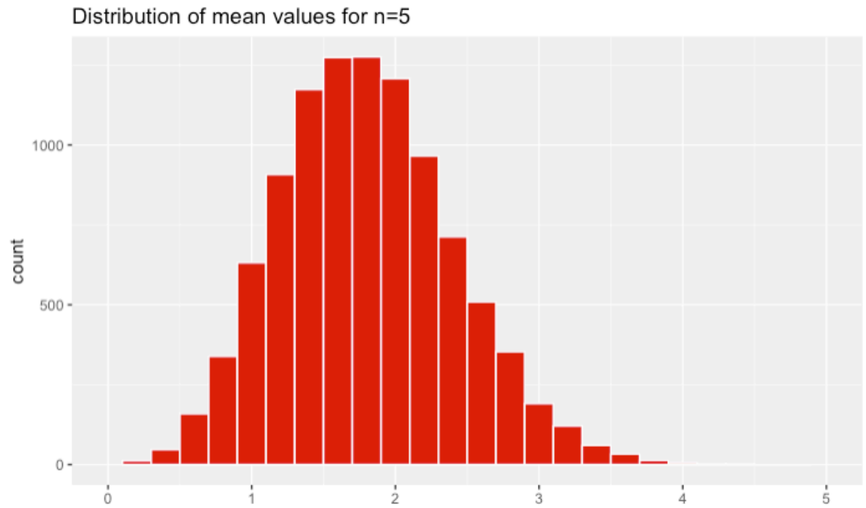
Example n=5

	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7	Day_8	Day_9	Day_10
Family_1	1	0	4	4	3	2	4	2	0	0
Family_2	3	2	1	1	2	2	4	3	1	1
Family_3	1	4	2	0	2	2	2	1	1	2
Family_4	3	2	2	1	6	1	3	1	1	1
Family_5	4	1	0	4	2	0	0	1	0	3

...

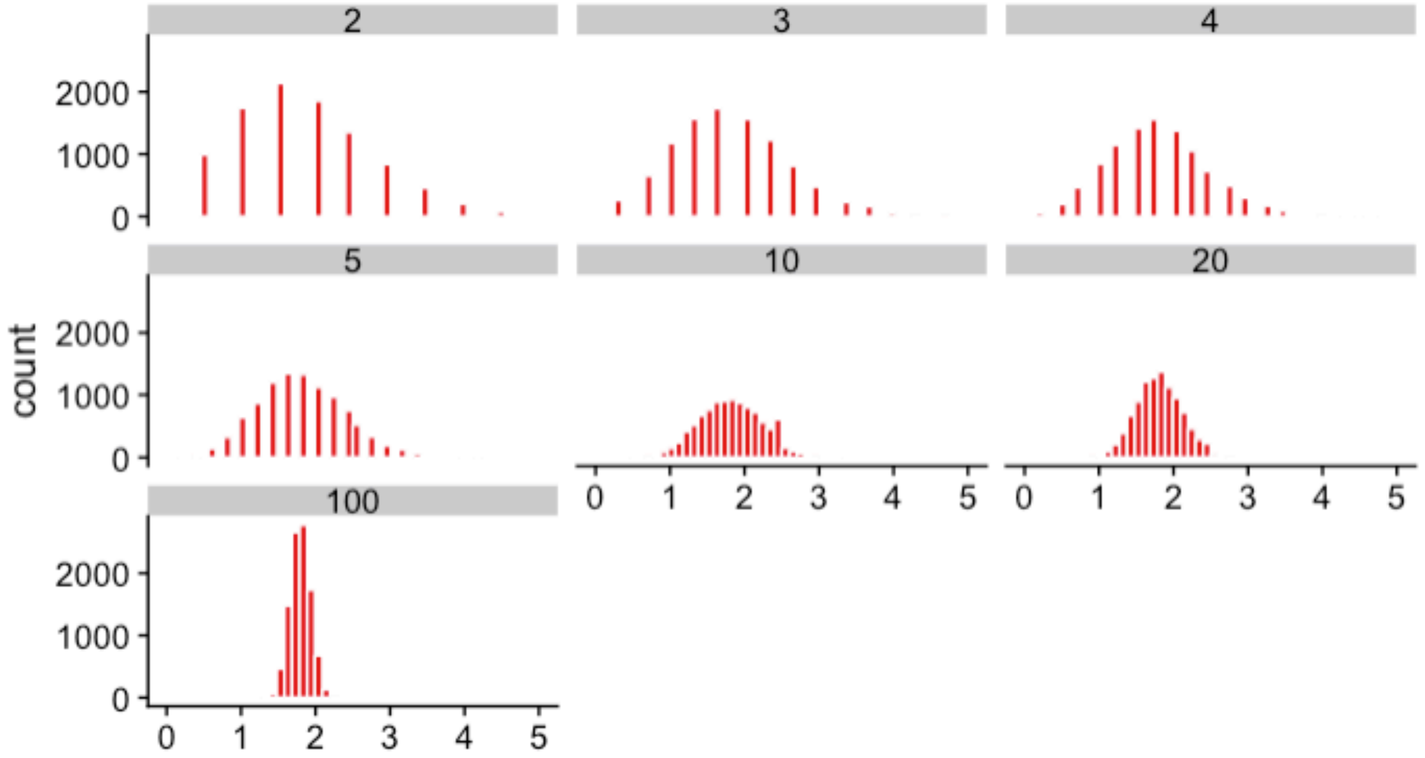
	Day_1	Day_2	Day_3	Day_4	Day_5	Day_6	Day_7	Day_8	Day_9	Day_10
m_5	2.4	1.8	1.8	2	3	1.4	2.6	1.6	0.6	1.4

...

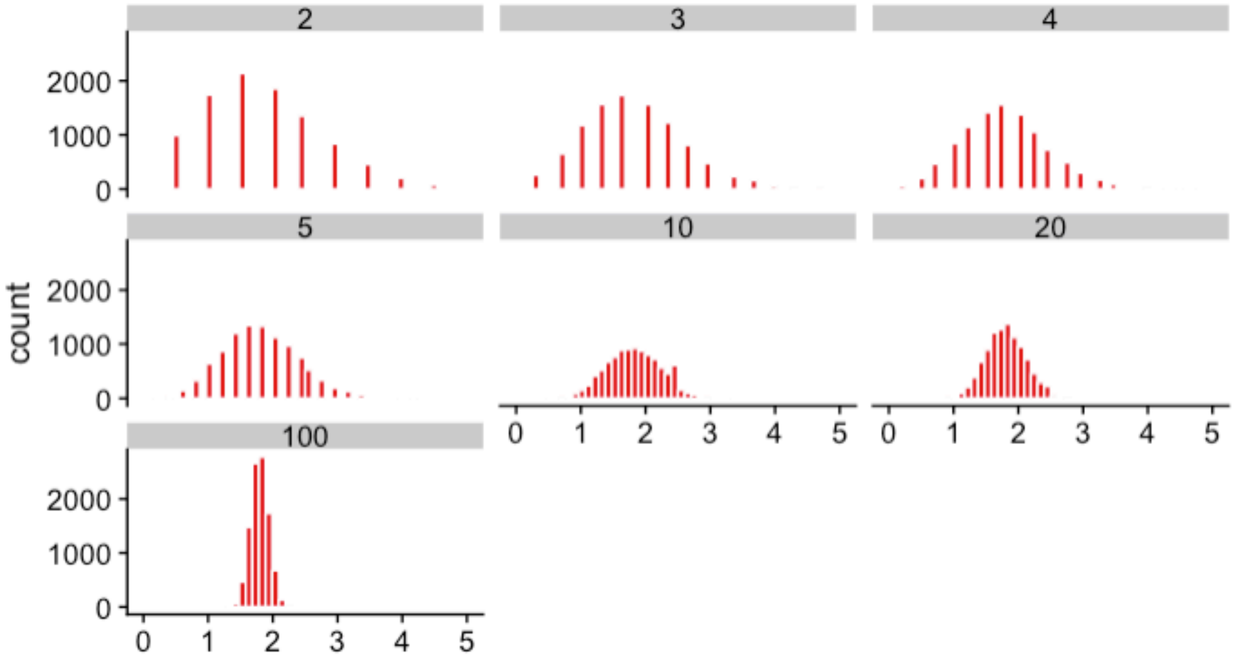


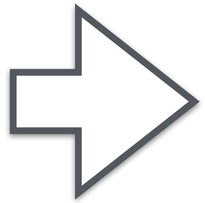
Distribution of means

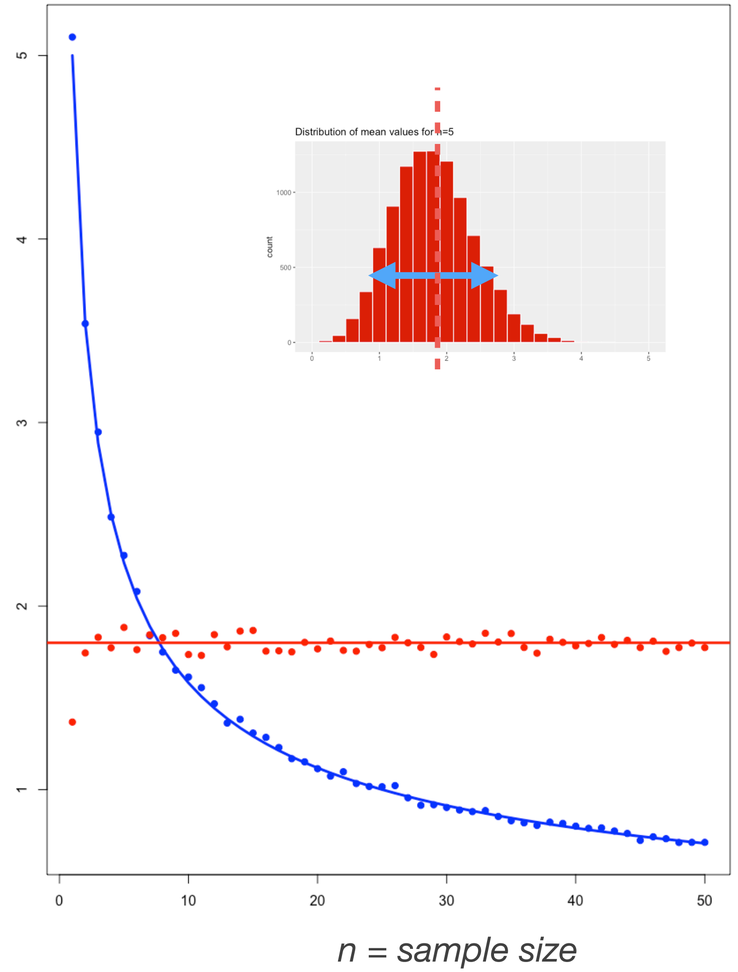
- For increasing sample size n:
  - distribution of mean values **tends to a normal distribution**
  - distribution of mean values **becomes narrower**
- **less fluctuations of the mean between samples**



Distribution of means







$$M_n \longrightarrow \mathcal{N}(\mu, \sigma/\sqrt{n})$$

mean of the sample is a good ("unbiased") estimator of the expectation

Standard error vs. Standard deviation

- **Standard deviation** describes the *spread in a distribution*

$$\text{Standard deviation} : \sigma$$

- **Standard error** measures the *spread in the estimation of the mean*

$$\text{Standard error} : \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

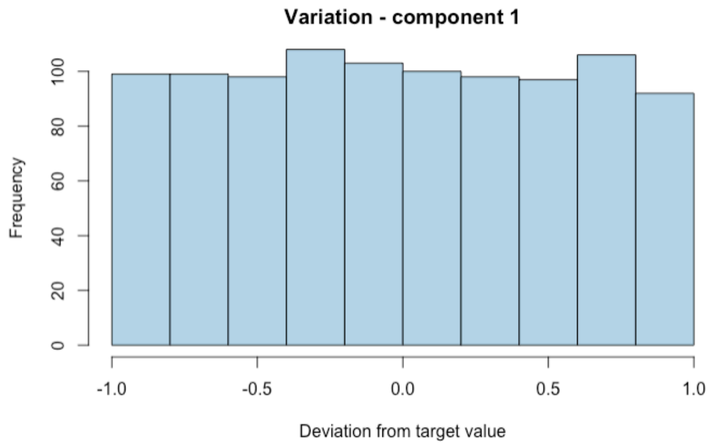
- $\{X_1, X_2, \ldots, X_n\}$ *n* random variables such that
  - independent
  - identically distributed
  - $E[X_i] = \mu$
  - $Var(X_i) = \sigma^2$

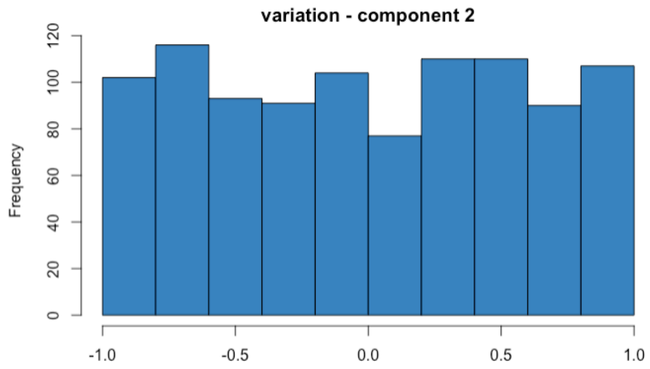
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \xrightarrow[n \to \infty]{} \quad \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \xrightarrow[n \to \infty]{} \quad \mathcal{N}(0,1)$$

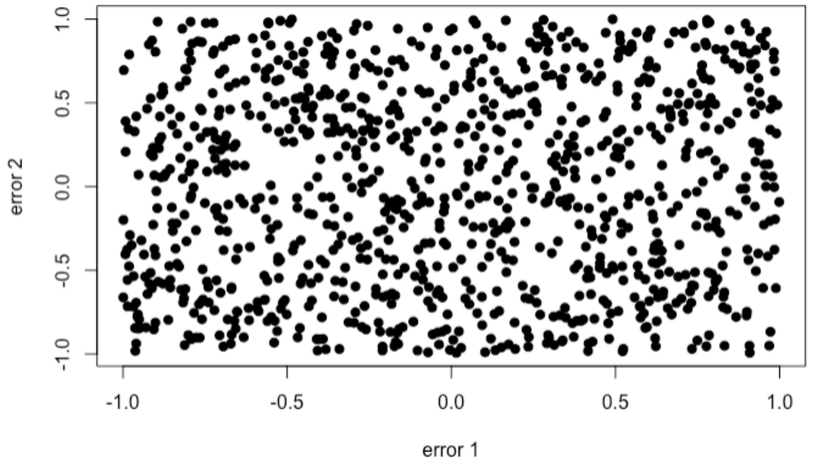
Central Limit Theorem

- **Example**: effect of multiple sources of errors
- Production process - multiple sources of error X_i on the final weight of the product
- Assume these errors are i.i.d (independent, identically distributed - for example uniformly distributed)





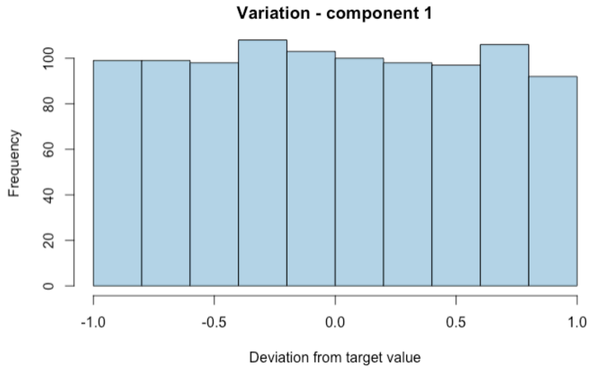
identically distributed (uniform)

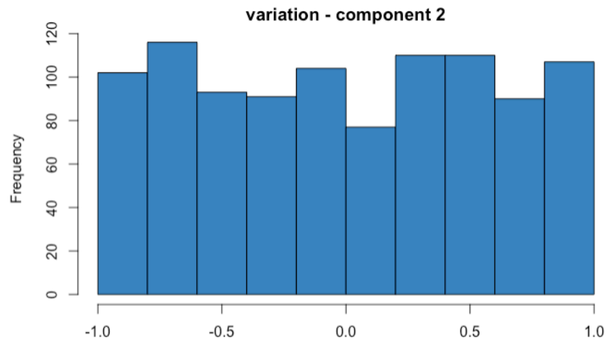


independent (no correlation)

Central Limit Theorem

- **Example**: effect of multiple sources of errors
- Total error on final product = mean of errors (for example on weights)



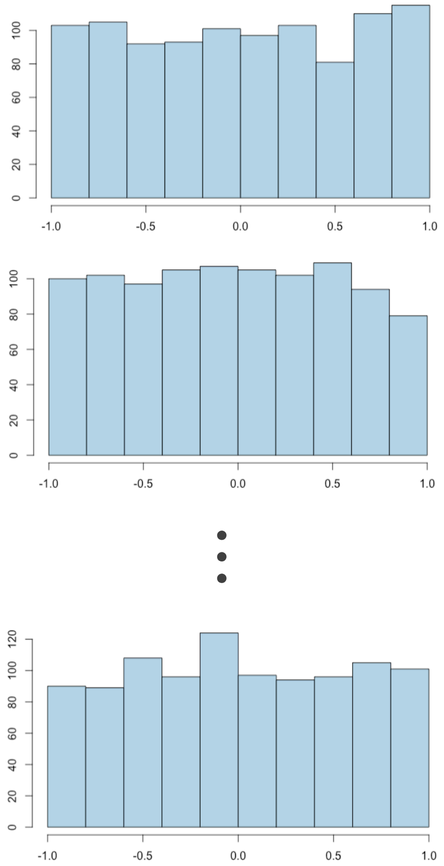


mean

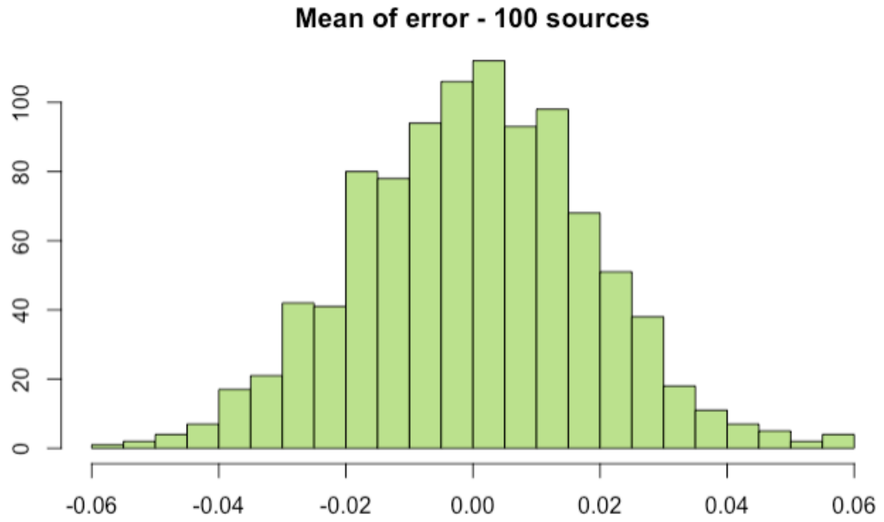
This is no longer a uniform distribution!

Central Limit Theorem

- **Example**: effect of multiple sources of errors
- Total error on final product = sum of errors (for example on weights)



mean



Converges to a normal distribution!

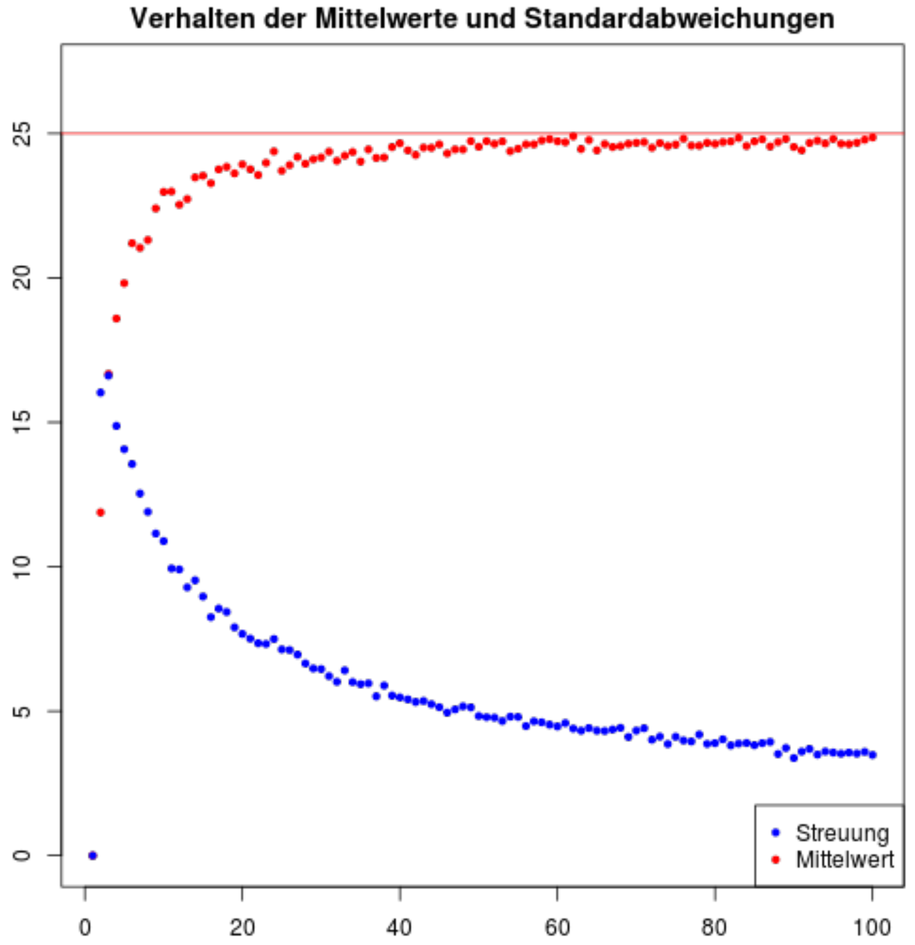
Estimation of variance

- Estimated sample variance

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

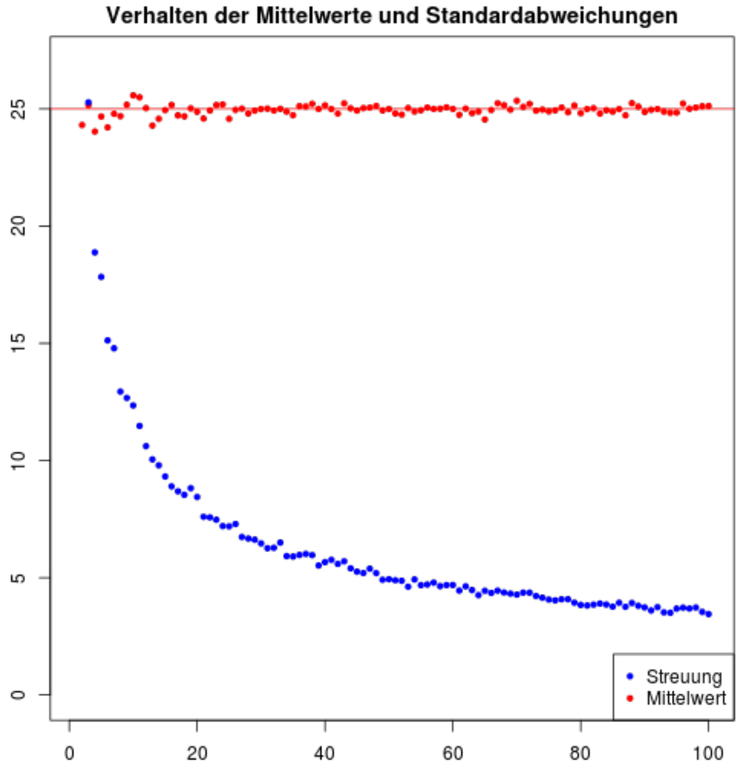
- converges **only asymptotically** to the true variance of the random variable

sample variance is a <u>biased</u> estimator of the population variance!



Estimating variance

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



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

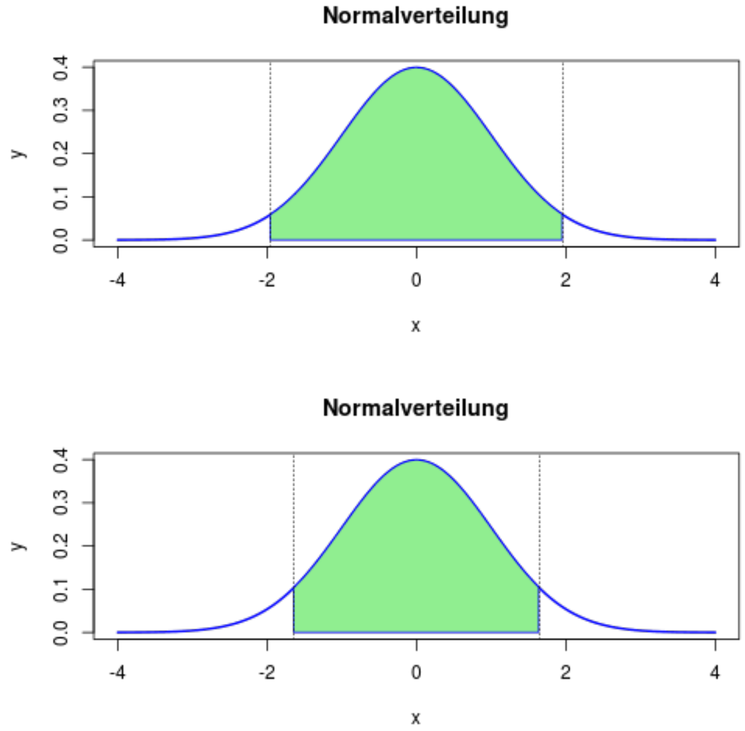
is a unbiased estimator of the population variance!

Confidence interval

$$M_n \xrightarrow{n \gg 1} \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Central Limit Theorem

$$\mathrm{M}_n \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}); \frac{M_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$



- in 95% of the cases, $\left|\frac{m_n - \mu}{\sigma/\sqrt{n}}\right|$ is smaller than t_{95}=1.96
- 95% confidence interval:

$$\mu \in [m_n - t_{95}\frac{\sigma}{\sqrt{n}} \; ; \; m_n + t_{95}\frac{\sigma}{\sqrt{n}}]$$

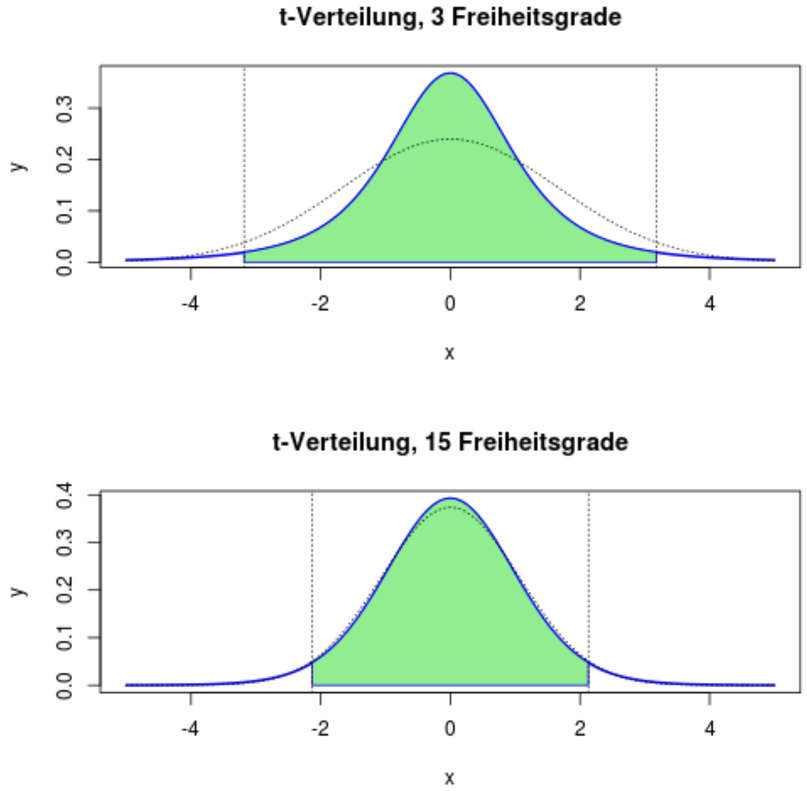
- **Problem**: we do not know the value of the population standard deviation σ
- Replace by its estimator

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

Confidence interval

for small samples:

$$\frac{M_n - \mu}{\sigma/\sqrt{n}} \sim t_{n-1}$$



- in 95% of the cases, $\left|\frac{m_n - \mu}{\sigma/\sqrt{n}}\right|$ is smaller than $t_{95,n-1}$
- *95% confidence interval:*

$$\mu \in [m_n - t_{95,n-1}\frac{\sigma}{\sqrt{n}} \;;\; m_n + t_{95,n-1}\frac{\sigma}{\sqrt{n}}]$$

- **Problem**: we do not know the value of the population standard deviation σ
- Replace by its estimator

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

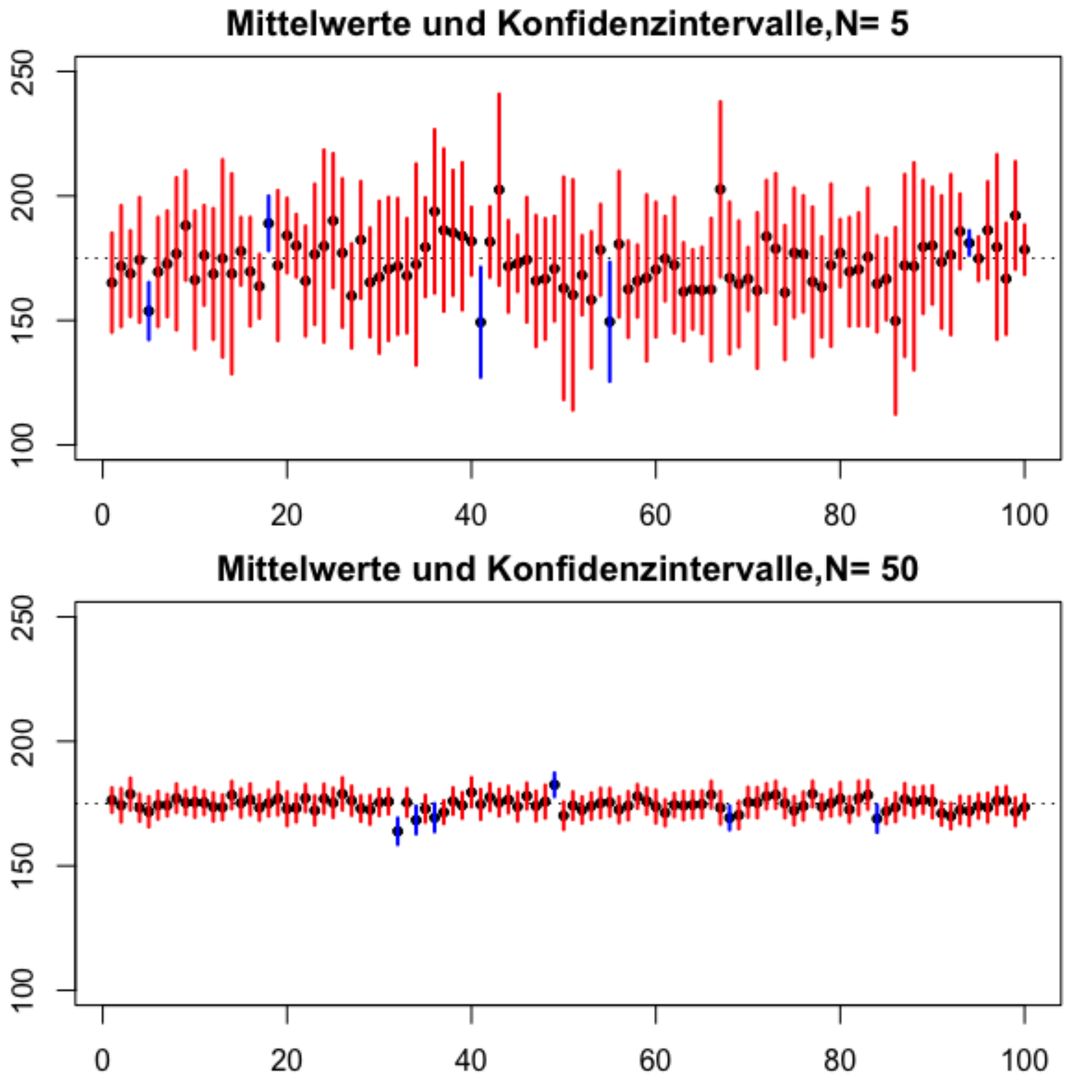
Critical values for t-distribution

df	t_95	t_90
2	4.302653	2.919986
3	3.182446	2.353363
4	2.776445	2.131847
5	2.570582	2.015048
8	2.306004	1.859548
10	2.228139	1.812461
15	2.131450	1.753050
20	2.085963	1.724718
50	2.008559	1.675905
100	1.983972	1.660234
200	1.971896	1.652508
Normal	1.959964	1.644854

Interpretation of confidence interval

- 95% confidence interval
  - 100 samples with n elements
  - sample distribution with expectation value μ
  - estimate mean and CI

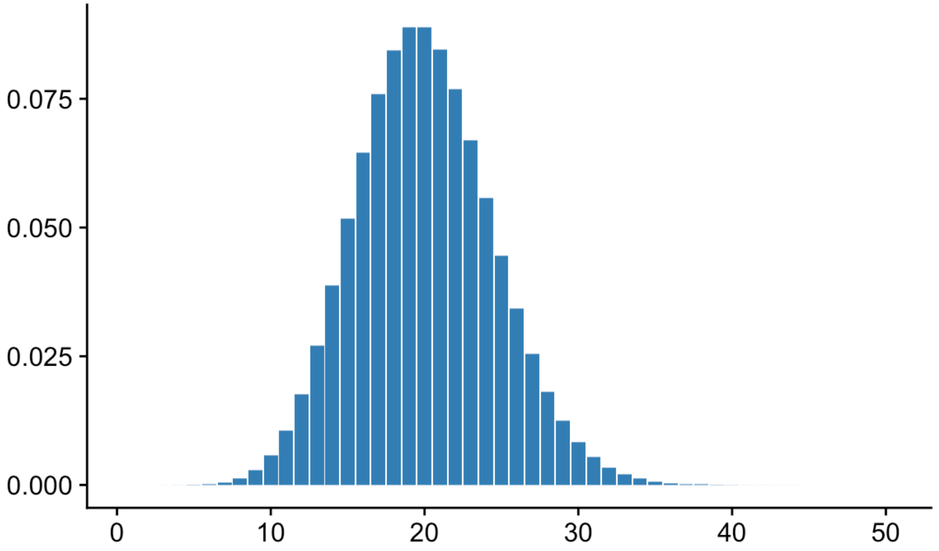
- True expectation value lies in 95% of the cases within the confidence interval



Confidence interval for Poisson distributions

- I find k=20 raisins in my raisin bun → what's in the recipe?
- Number of raisins *N* per bun is a **random variable**, described by a Poisson distribution

$$N \sim Pois(\lambda = k) \qquad Pois(k) \overset{k \gg 1}{\longrightarrow} \mathcal{N}(k, \sqrt{k})$$



$$\mu \in [m_n - t_{95}\frac{\sigma}{\sqrt{n}} \; ; \; m_n + t_{95}\frac{\sigma}{\sqrt{n}}]$$

(Sample size n=1 !)

$$\mu \in [k - t_{95}\sqrt{k} \; ; \; k + t_{95}\sqrt{k}]$$

$$11.2 \leq \mu \leq 28.8$$

Confidence interval Poisson distribution

How to reduce the confidence interval?

- **eat more buns!** n=10 samples with 1 bun
- Compute the mean over the sample $\bar{k} = 20.2$

28 17 22 19 18 ...

$$\mu \in [m_n - t_{95}\frac{\sigma}{\sqrt{n}} \; ; \; m_n + t_{95}\frac{\sigma}{\sqrt{n}}]$$

$$\mu \in [\bar{k} - t_{95}\frac{\sqrt{\bar{k}}}{\sqrt{10}} \; ; \; \bar{k} + t_{95}\frac{\sqrt{\bar{k}}}{\sqrt{10}}]$$

$$17.4 \leq \mu \leq 22.9$$