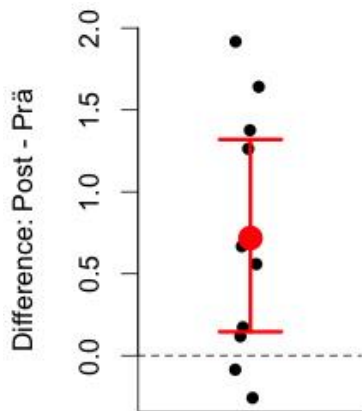


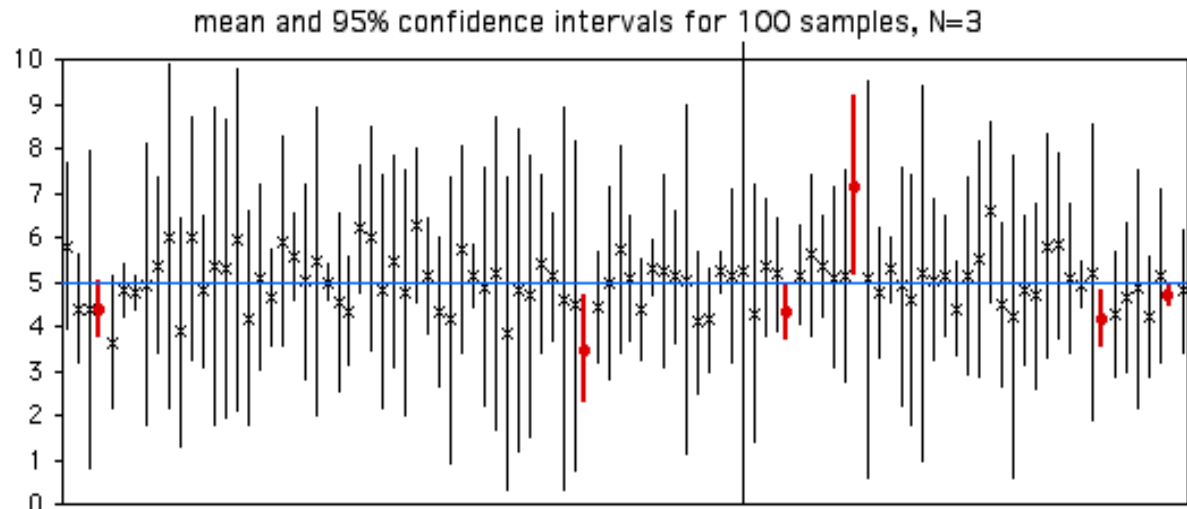
# Biostatistics Week 3

## Inferential statistics

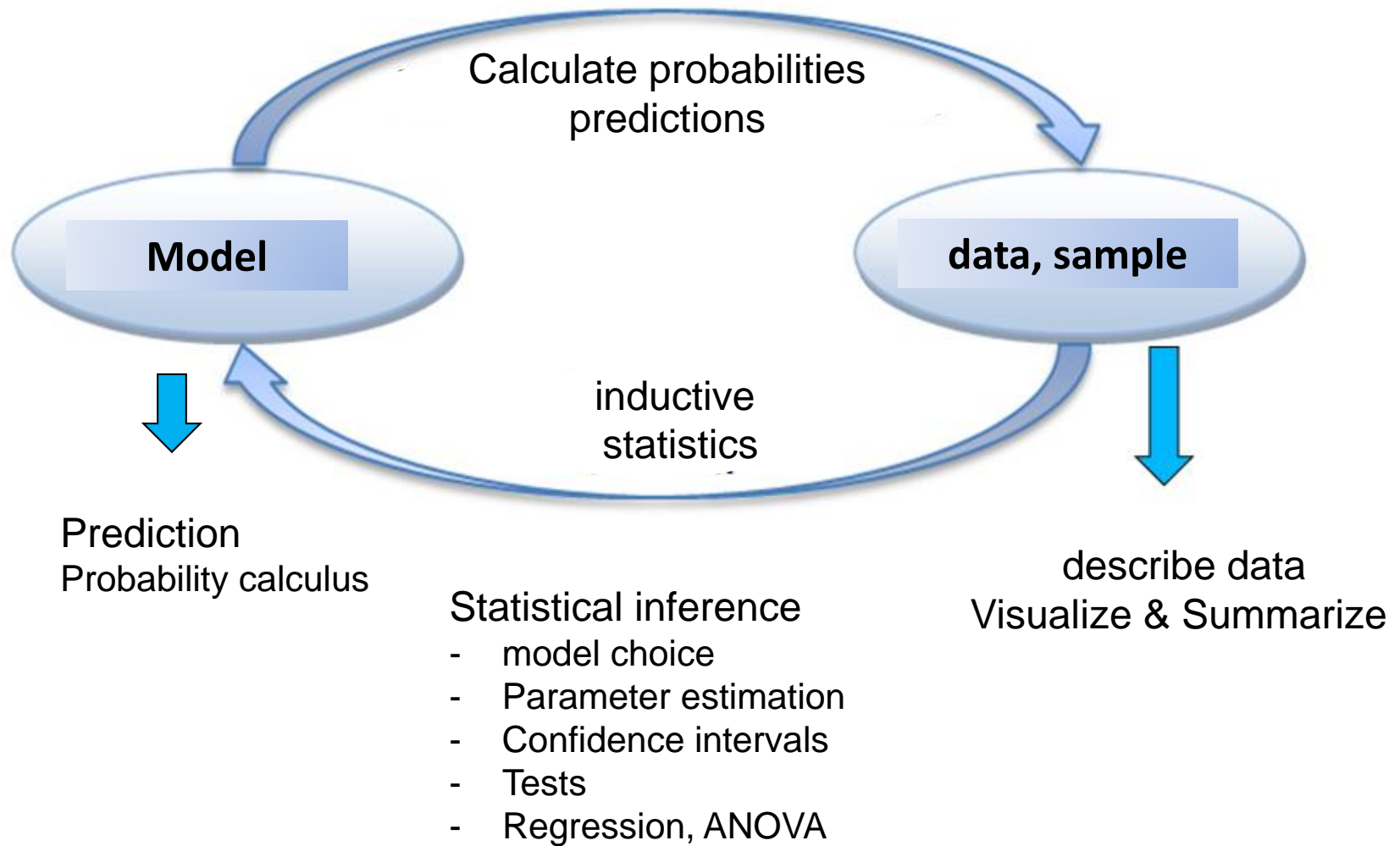
- What is inferential statistics about?
- The zoo of distribution models & model choice
- Expected value and variance and how to estimate them
- Confidence interval, significance ( $\neq$  relevance)



x

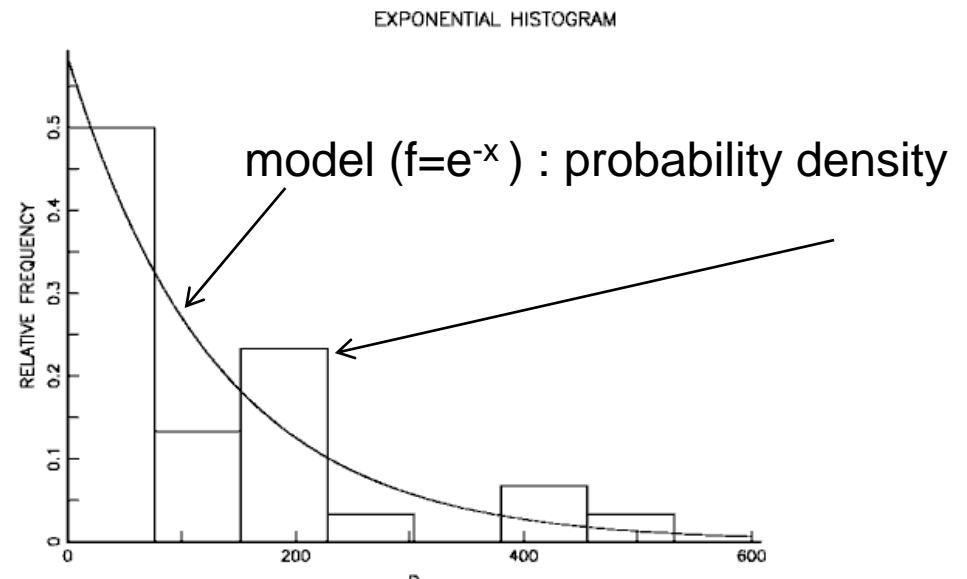
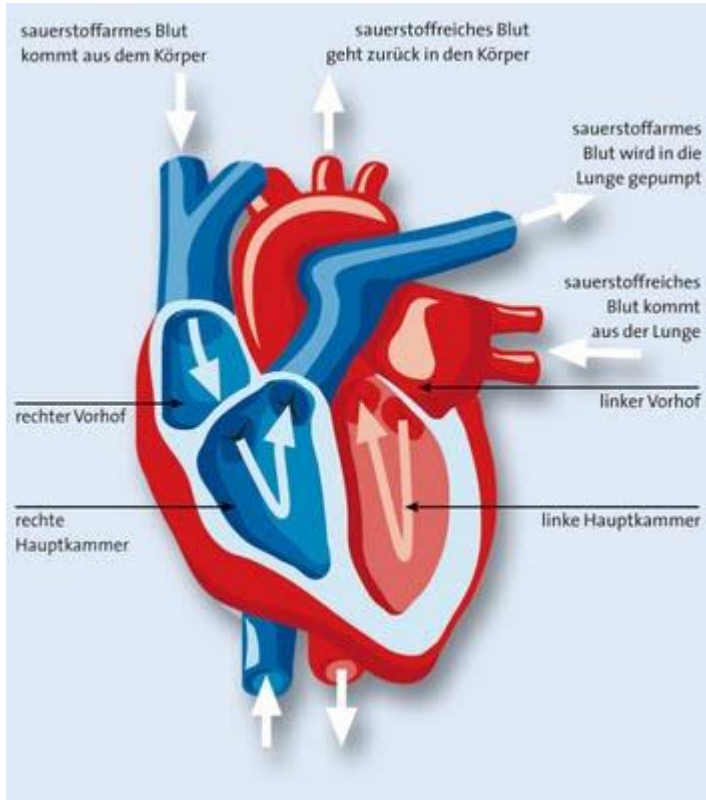


# Statistics connects data with models



# What is a model?

## Model for the human heart

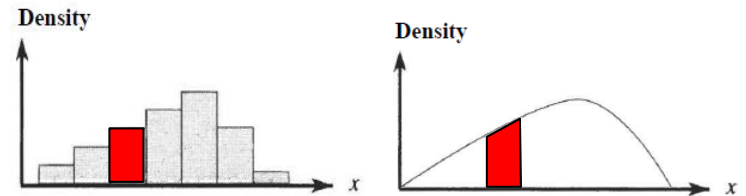


# The world of data and the world of models

data/reality	model
sample	population
discrete data/features (numeric or categorical)	discrete random variable (numeric)
continuous data/features (numeric)	continuous random variable (numeric)
observation	Random variable
relative frequency	probability (P)
histogram (scaled)	Density continuous distribution
bar plot of frequencies (scaled) (of rel. frequency at discrete features)	Probability distribution discrete distribution
average $\bar{x}$	expected value $\mu$
sample variance $s^2$	variance $\sigma^2$

# What can be meant by probability?

1. The Probability to get a value in a certain range can be given as **limit of relative frequencies**.
2. Probability can be determined from a **probability model** – we have a model for dice in our mind which tells us that the probability for each side is equal.
3. An **opinion resp. experience or expectation**, can be expressed as probability. E.g. the probability of a airplane crashing in a nuclear power station is given by expert opinion.
4. In mathematical statistics probability is given by the **Kolmogorov's formulation**, sets are interpreted as events and probability itself as a measure (probability) on a class of sets.



Axiom 1  $0 \leq P(A) \leq 1$

Axiom 2  $P(S) = 1$

Axiom 3  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

# Example for a distribution model: the discrete uniform distribution

X: Result when throwing a dice (X is a random variable)

Probability distribution in table representation:

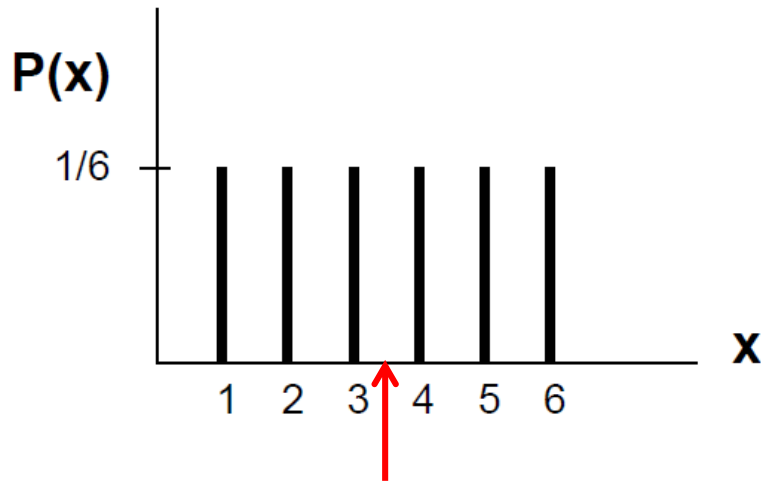
X	1	2	3	...	6
p	1/6	1/6	1/6	...	1/6

model



generates rv X

Probability distribution as graph:



expected value  $E(X)=3.5$

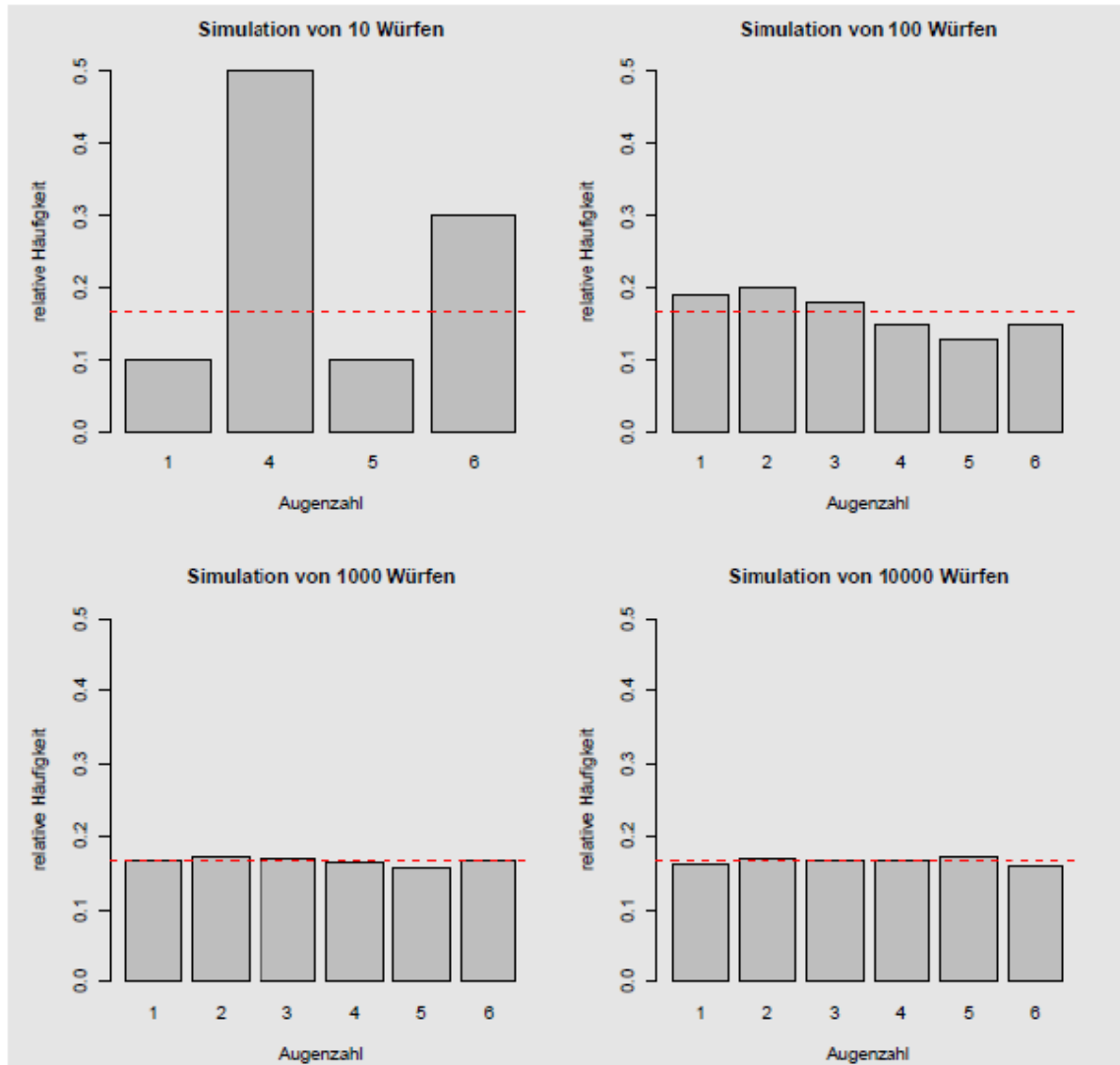
Probability distribution  
as function:

$$P(X = k) = \frac{1}{r}, k \in \{1, 2, \dots, r\}$$

$r$ : # possible results

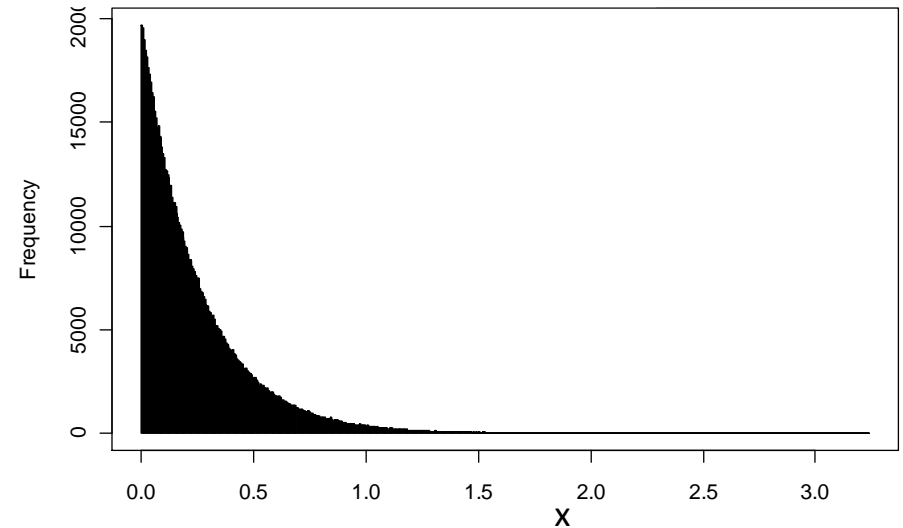
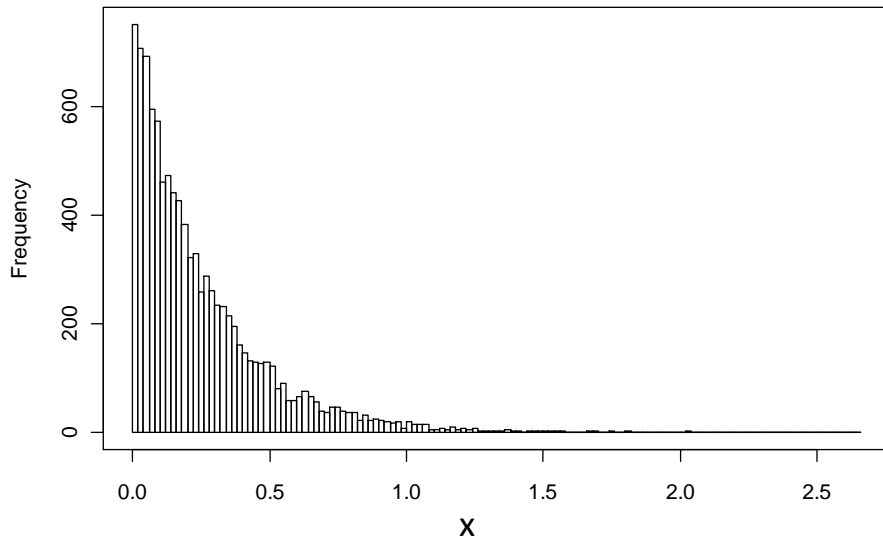
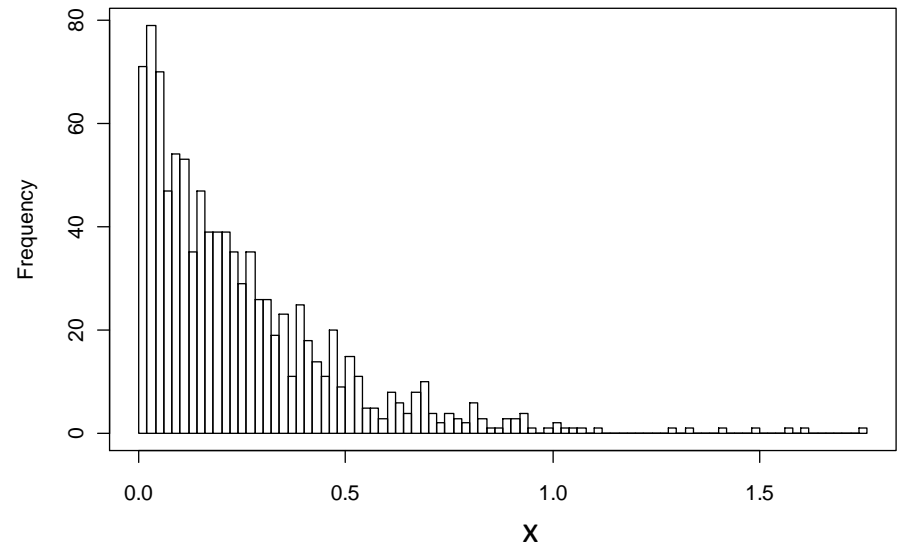
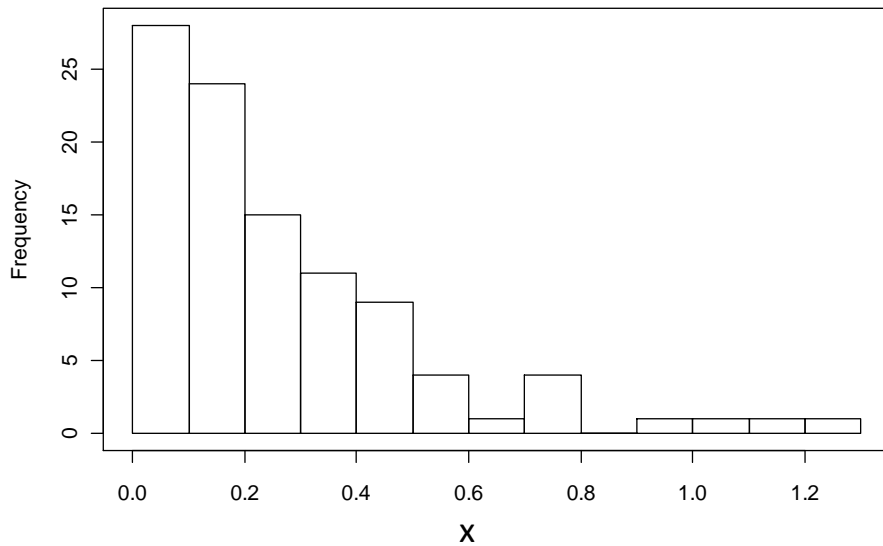
$$\sum_{k=1}^r P(X = k) = 1$$

# Estimate the probability distribution and expected value via determining the relative frequencies by experiment or simulation



We need >10'000 throws of a dice or runs of simulations before getting a reliable estimate of the distribution or the expected value.

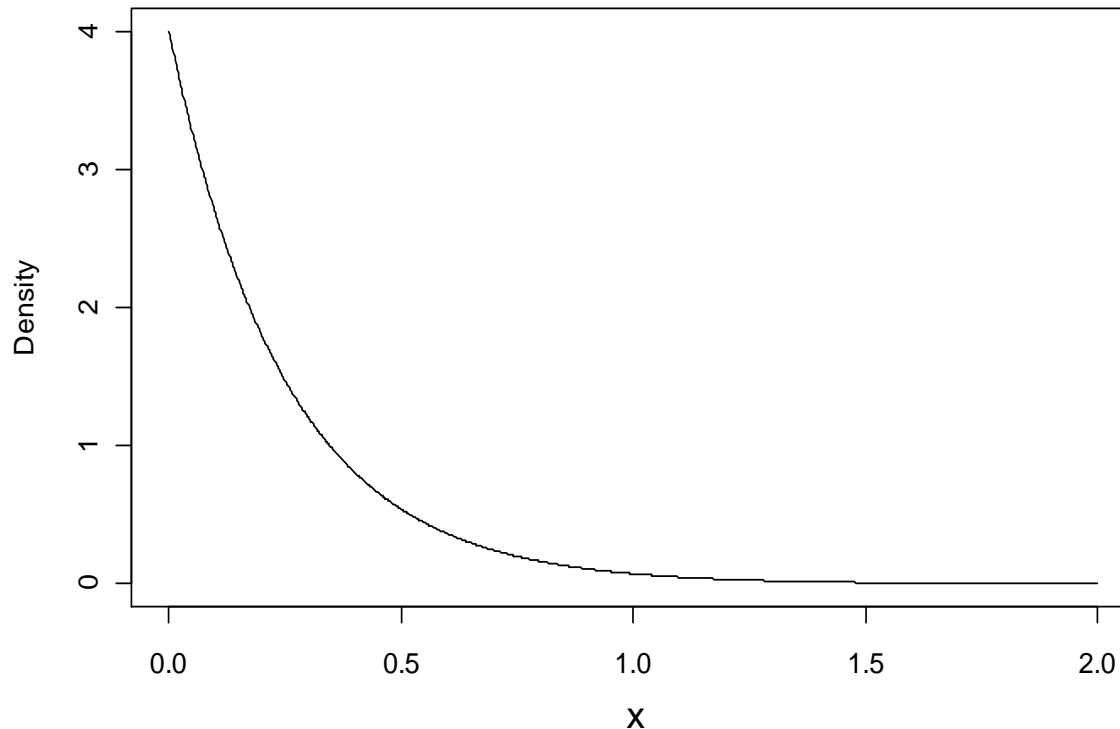
# The scaled histogram of waiting times for an increasing number of observations



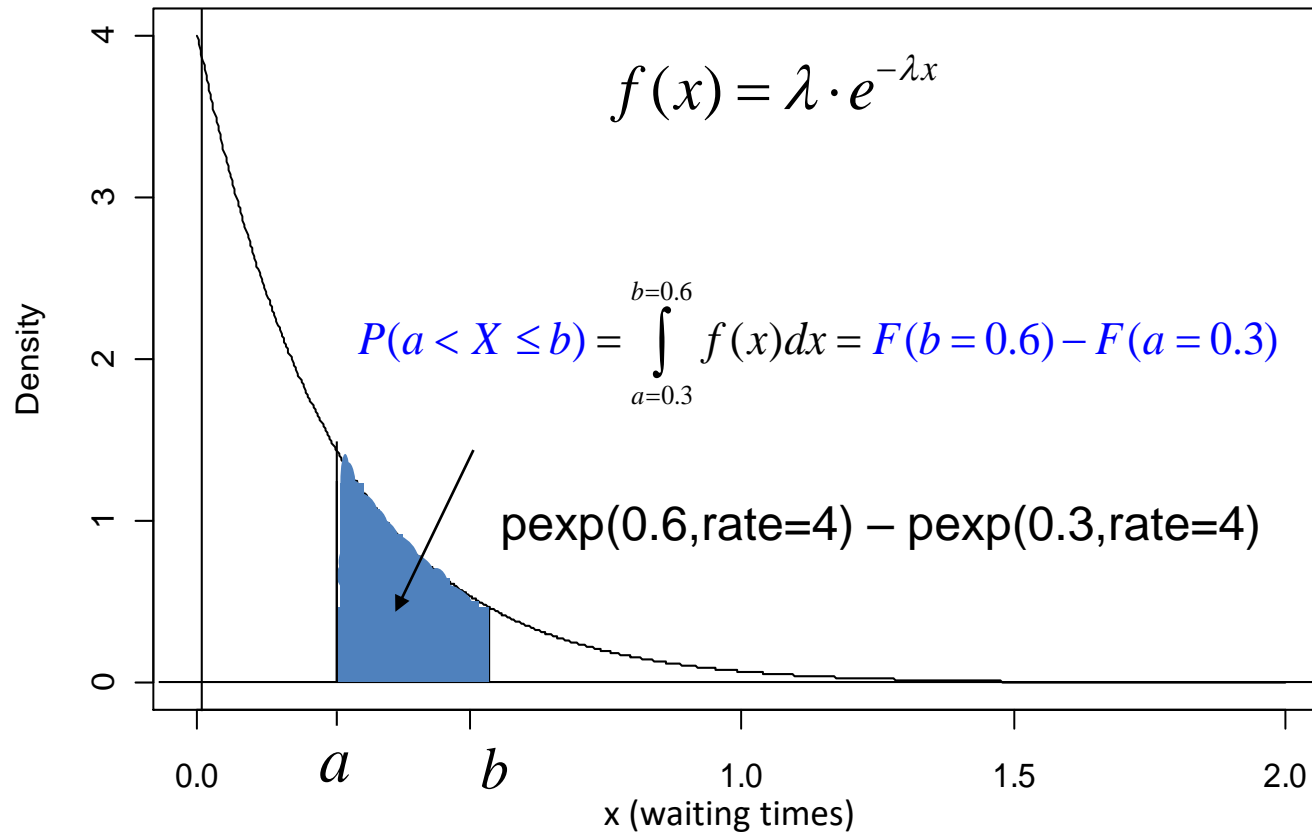


# The limit-case of infinity observations

If we have infinity observations the hull of a scaled histogram can be modeled by smooth function, called the **density function**, with an **area under the curve of one**.



# The probability density function (pdf)



The probability of getting a result between  $a$  and  $b$  is equal to the area under the density function above the interval  $[a, b]$ . The calculation of the probability is made by integrating the density function in interval  $[a, b]$

# Properties of a density function

The density function  $f$  of a continuous variable  $X$  is a piece-wise continuous function with

$$f(x) \geq 0$$

and

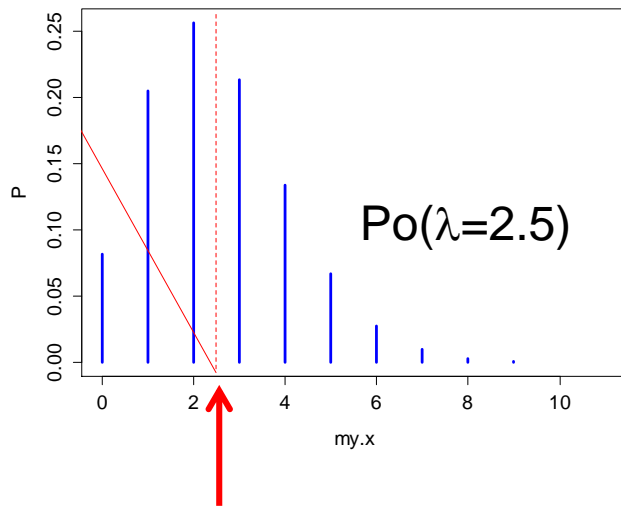
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

# The expected value = population mean

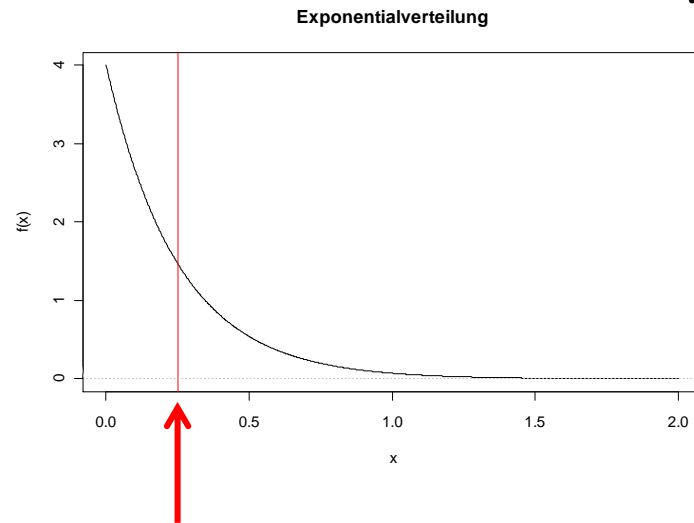
- The expected value of a random variable is the average, which we would get with an infinite big sample
- It measures the location of the random variable
- It corresponds to the center of mass of the density – balance point (see red line)
- It often determines the parameter of the model
- The expected value can also be calculated from the model

$$EW(X) = \sum_{i=1}^n x_i \cdot P(X = x_i)$$

$$EW(X) = \int x \cdot f(x) dx$$

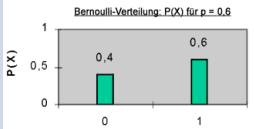
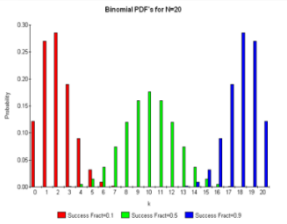
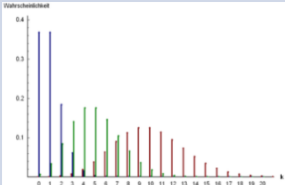


expected value  $E(X)$

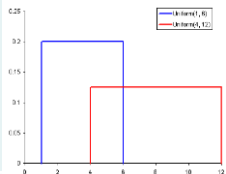
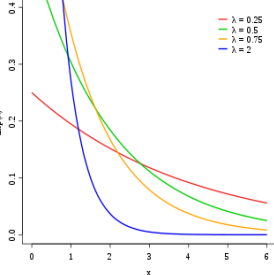
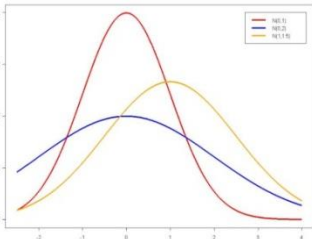


expected value  $E(X)$

# The most famous discrete distributions/models

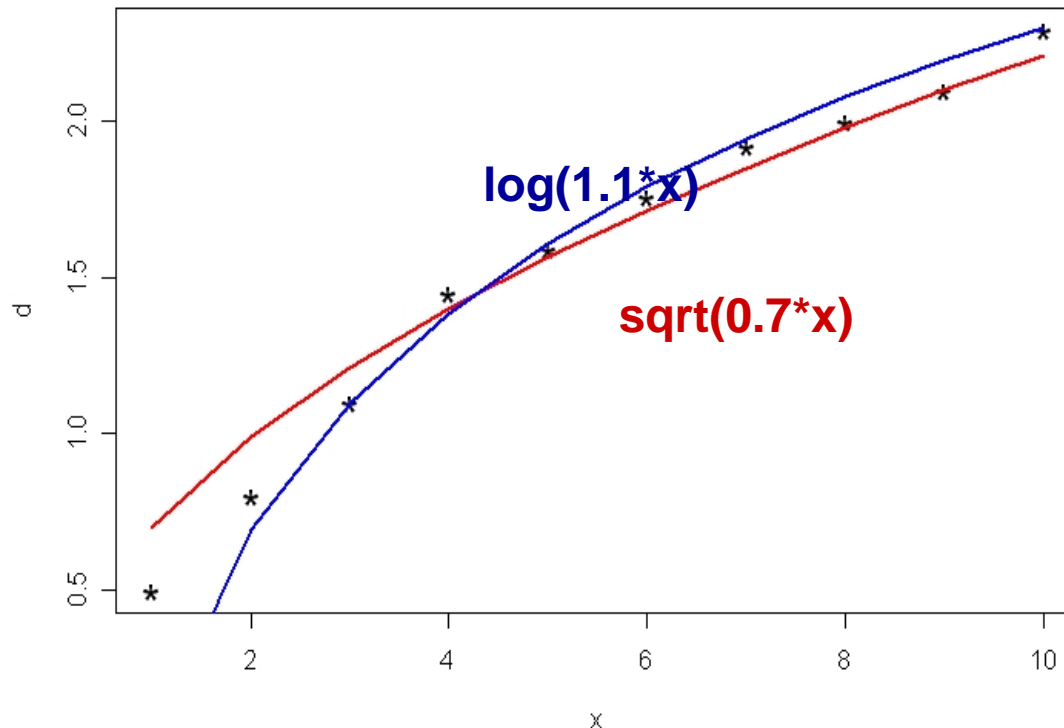
name of the distribution	possible values $x$ $P(X=k)$	expected value $\mu$ variance $\sigma^2$	application
Bernoulli $X \sim \text{Bern}(\pi)$ 	$x \in \{0, 1\}$  $P(X = 1) = \pi$ $P(X = 0) = 1 - \pi$	$\mu = E(X) = \pi$ $\sigma^2 = \text{Var}(X) = \pi * (1 - \pi)$	$X$ : indicates if an event occurs or not
Binomial $X \sim B(n, \pi)$ 	$x \in \{0, 1, \dots, n\}$  $P(X = k) = \binom{n}{k} \cdot \pi^k \cdot (1 - \pi)^{n-k}$	$\mu = E(X) = n * \pi$ $\sigma^2 = \text{Var}(X) = n * \pi * (1 - \pi)$	$X$ : number of successes in $n$ independent Bernoulli trials
Poisson $X \sim \text{Po}(\lambda)$ 	$x \in \{0, 1, \dots\}$  $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$	$\mu = E(X) = \lambda$ $\sigma^2 = \text{Var}(X) = \lambda$	$X$ : number of events in a certain interval or time-bin

# The most important continuous distributions/models

Name of the distribution (parameter)	domain density f distribution F	expected value variance	application
Uniform V. $X \sim U(a,b)$ 	$\mathbb{R}$ $f(x) = \frac{1}{b-a}, \quad \text{for } a \leq x \leq b, \\ \text{otherwise } f(x) = 0$ $F(x) = \frac{x-a}{b-a} \quad \text{für } a \leq x \leq b$	$E(X) = \frac{a+b}{2}$ $Var(X) = \frac{(b-a)^2}{12}$	if all events have the same probability or if the probability is not known at all
Exponential V. (rate $\lambda$ ) 	$\mathbb{R}_0^+$ $f(x) = \lambda \cdot e^{-\lambda x}$ $F(x) = 1 - e^{-\lambda x}$	$E(X) = \frac{1}{\lambda}$ $Var(X) = \frac{1}{\lambda^2}$	waiting times, time to fail (or decay)
Normal V. $X \sim N(\mu, \sigma^2)$ 	$\mathbb{R}$ $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x'-\mu)^2}{2\sigma^2}} dx'$	$E(X) = \mu$ $Var(X) = \sigma^2$	typical measurements (affected symmetrically by various factors), Asymptotic approximation for other distributions

## Model choice

Which model fits the data better?



Logarithms or Square-Root function?

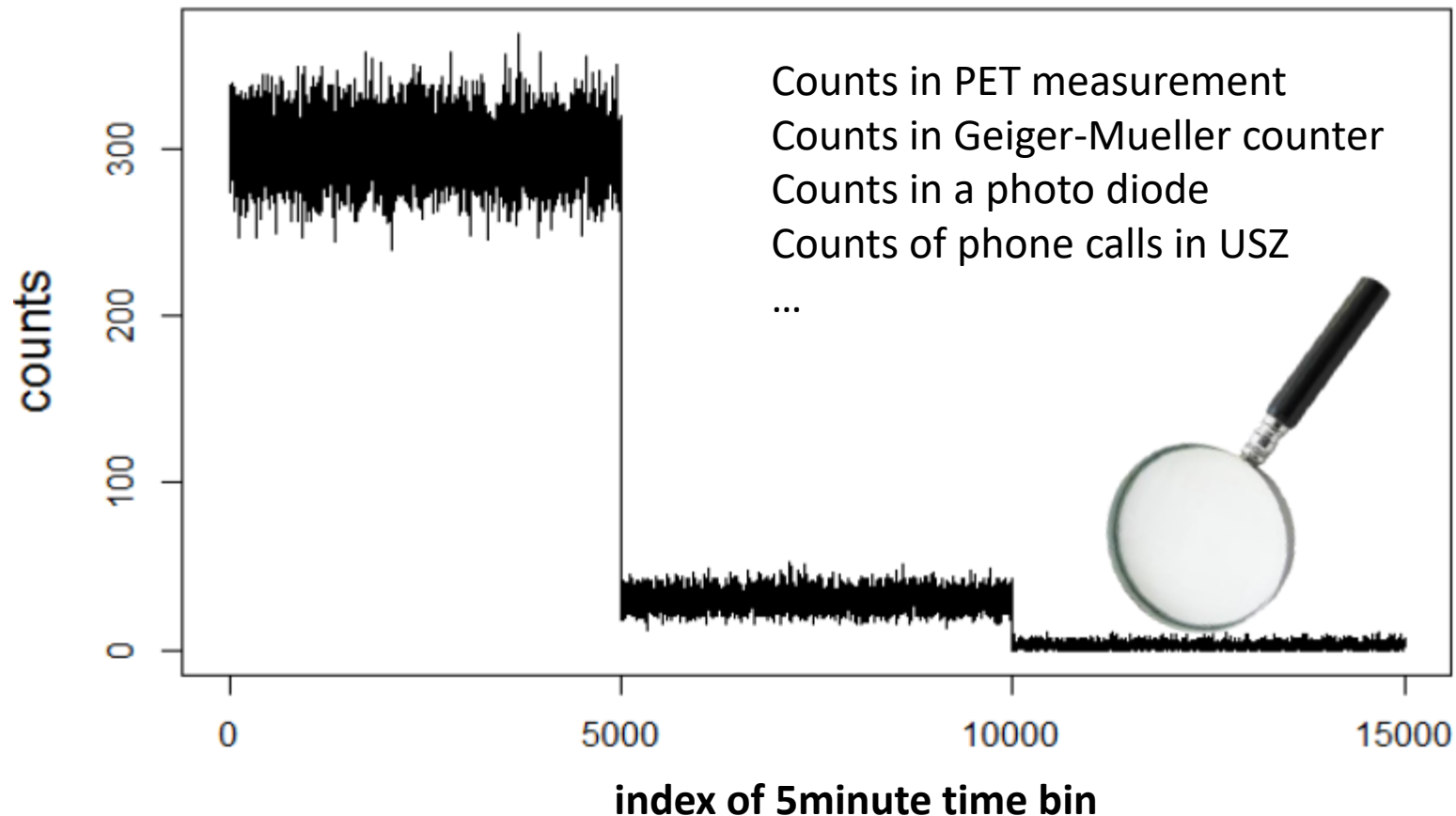
We should always check if the model is appropriate and fits the data(visually, by residual analysis, qq-Plot...)

## The Poisson model is appropriate when the measure of interest is the number of events per unit

Poisson Distribution for count data:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

$E(X) = \text{Var}(X) = \lambda = \text{mean number of events per unit (e.g. time bin)}$

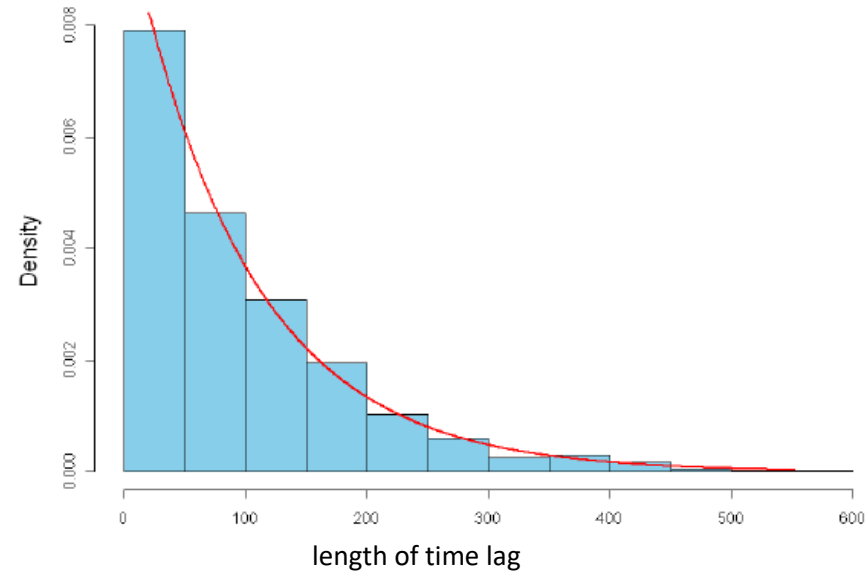
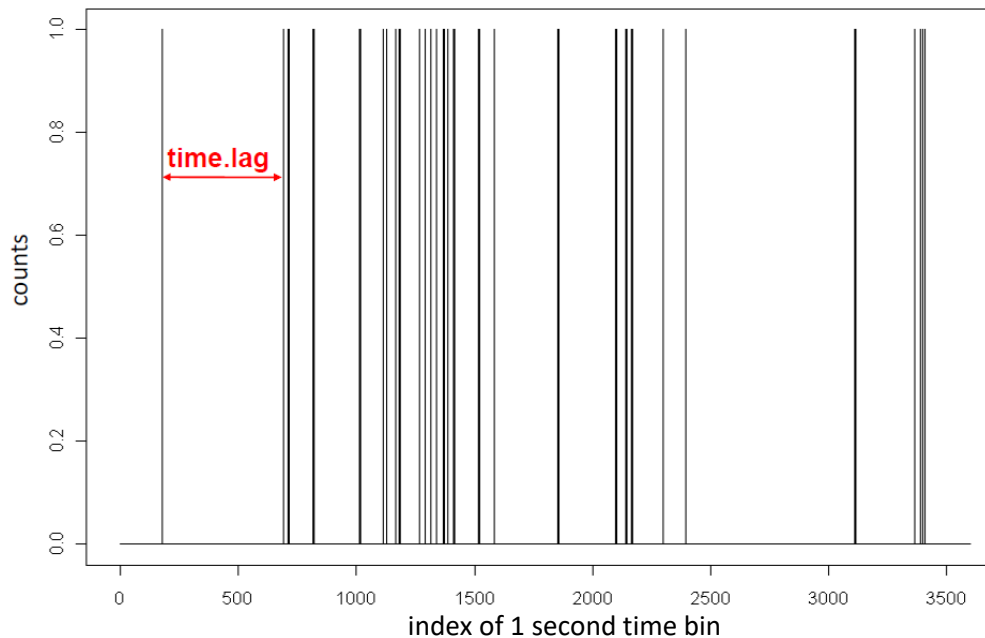




The Exponential model is often appropriate when we measure time to event, e.g. the time between 2 Poisson-events

$$f(x) = \lambda \cdot e^{-\lambda x}$$

$$\lambda = \frac{1}{\text{mean time lag}}$$



# The parameter estimation problem

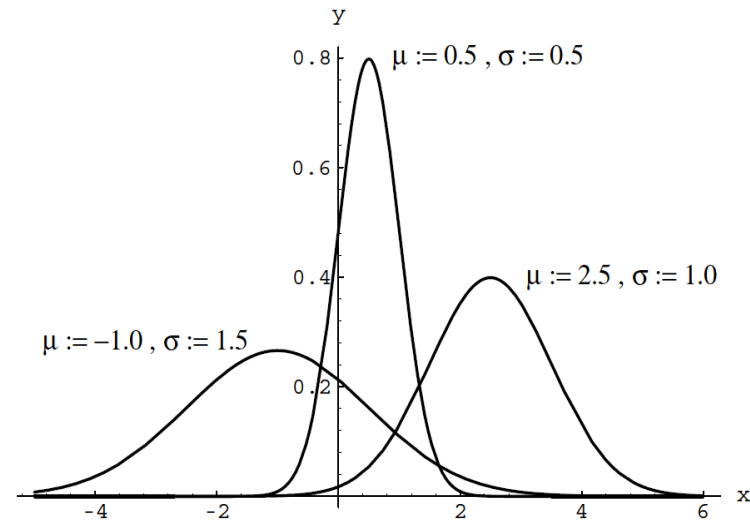
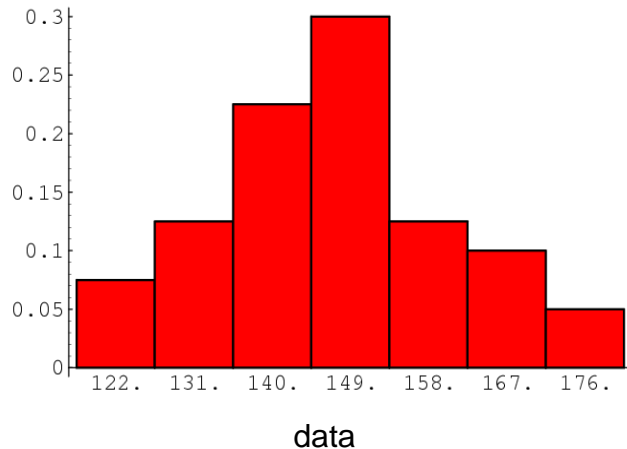
Location and the shape of a distribution is determined by the values of its parameters

We assume we know the appropriate model class.

How to estimate the parameter value from the data?

Which **parameter value** range is plausible?

Is a given parameter value compatible with the data?



## Parameter estimation for the most important distributions

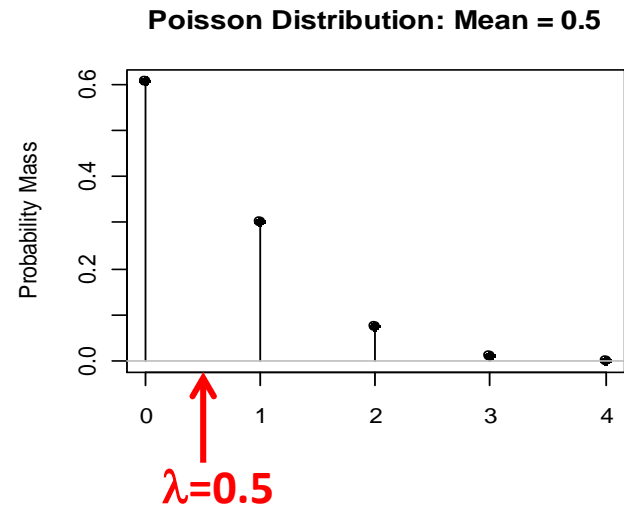
Distributionfamily V $X \sim V(\text{Parameter-Set})$	Relation Parameter-E(X)-Var(X)	Parameter-estimator as function of the data
Exponential $X \sim \text{Exp}(\lambda)$	$E(X) = \frac{1}{\lambda}$ $\text{Var}(X) = \frac{1}{\lambda^2}$	$\hat{\lambda} = \frac{1}{E(\hat{X})} = \frac{1}{\bar{x}}$
Normal $X \sim N(\mu, \sigma^2)$	$E(X) = \mu$ $\text{Var}(X) = \sigma^2$	$\hat{\mu} = E(\hat{X}) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ $\hat{\sigma}^2 = \text{var}(\hat{X}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
Binomial $X \sim B(n, p)$	$E(X) = n \cdot p$ $\text{Var}(X) = n \cdot p \cdot (1 - p)$	$\hat{p} = \text{average \#successes per } n \text{ trials}$
Poisson $X \sim \text{Po}(\lambda)$	$E(x) = \lambda$ $\text{Var}(X) = \lambda$	$\hat{\lambda} = E(\hat{X}) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

## The expected value is often a parameter of the model

Poisson Distribution for count data:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

$$E(X) = \sum_{k=1}^{\infty} k \cdot P(X = k) = \lambda$$



We use the sample mean to estimate the population mean = expected value (the central limit theorem (CLT) gives justification – see later in this lecture).

The ^ (hat) above the parameter indicates, that the parameter is not known but estimated.

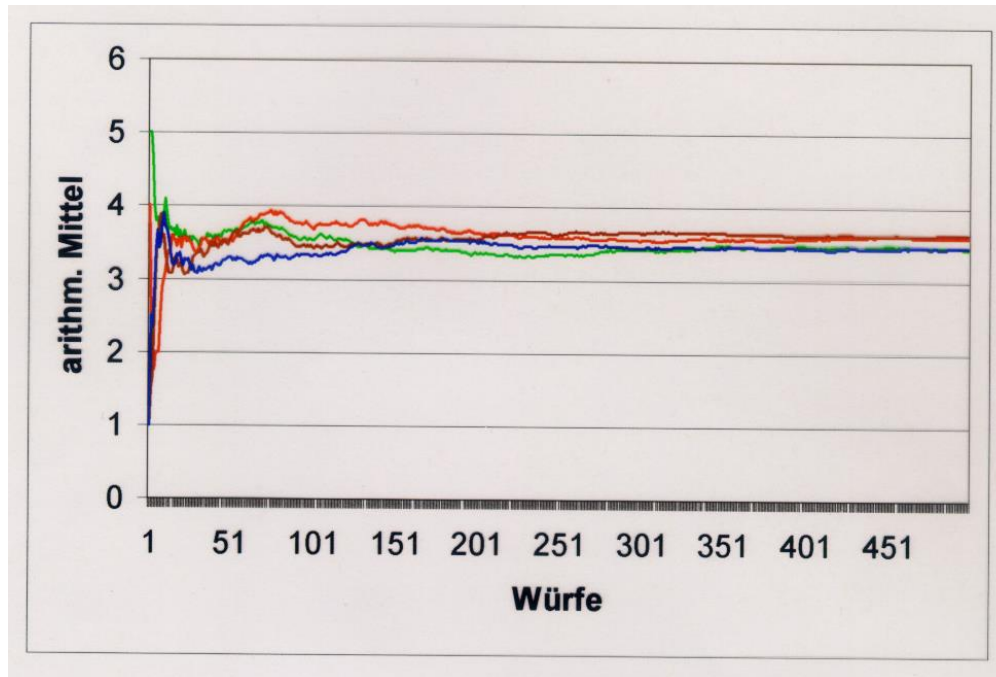
$$\bar{X} \xrightarrow{n \rightarrow \infty} E(X)$$

$$\hat{E}(X) = \bar{X}$$

# Expected value: The law of large numbers

The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.:

$$\bar{X} \xrightarrow{n \rightarrow \infty} E(X)$$



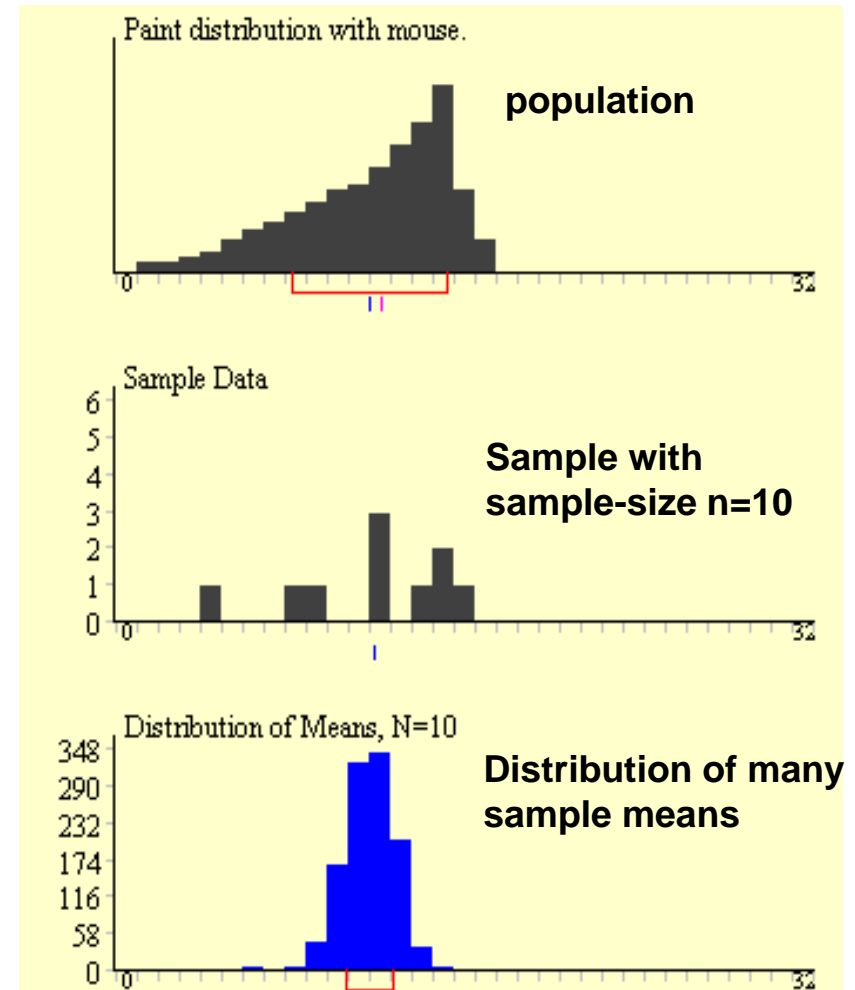
The expected value a dice throw is 3.5.

# The estimated expected value is random since it is based on a random sample

## Sample Variation & Central Limit Theorem

- Because of the sample variation also derived statistics like the mean value varies from sample to sample
- The sample mean is an unbiased estimator for the population mean  $E(X)$ .
- CLT: The sample mean is normally distributed around the population mean and the variation decreases with increasing sample size.

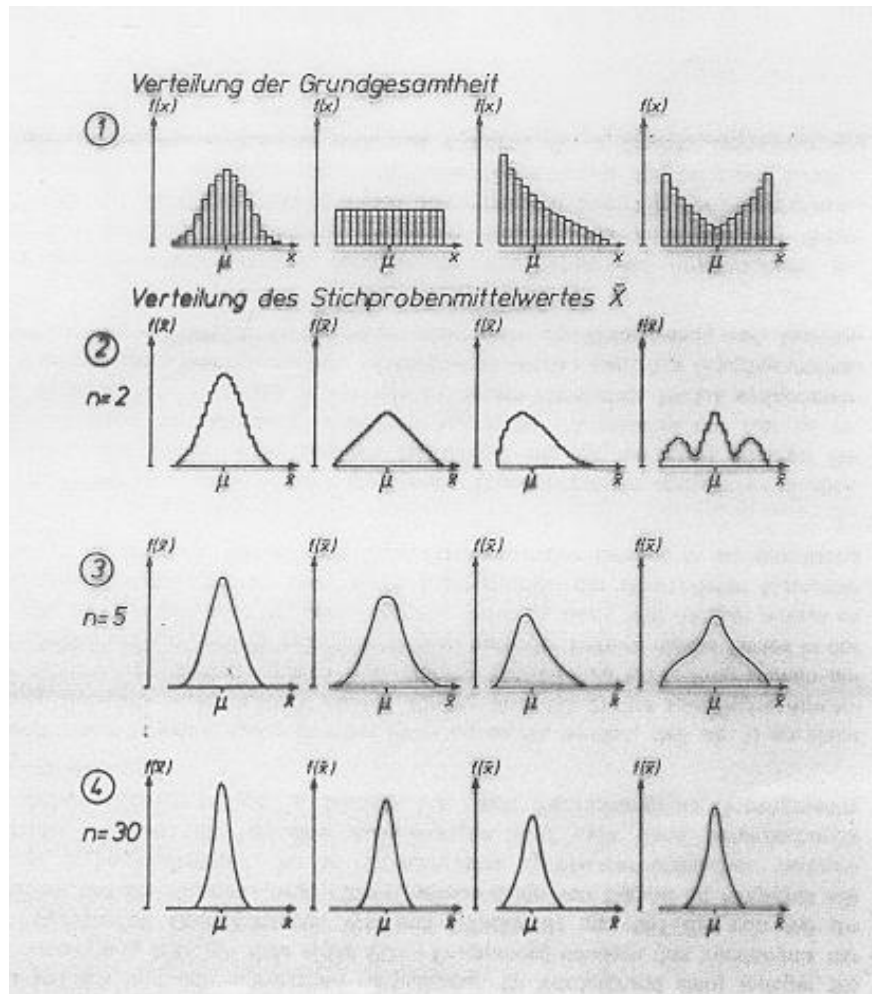
$$\bar{X} \overset{a}{\sim} N\left(\mu_x, \frac{\sigma_x^2}{n}\right) = N\left(E(X), \frac{Var(X)}{n}\right)$$



Warning: There are exceptions to the CLT, e.g.  $t_1$  also called Cauchy distribution or Lorenz curve in Spectroscopy which “has no mean”.

# The distribution of the mean -> Central Limit Theorem

The arithmetic mean of a sufficiently large number of independent random variables, will be approximately normally distributed



← distribution of the population

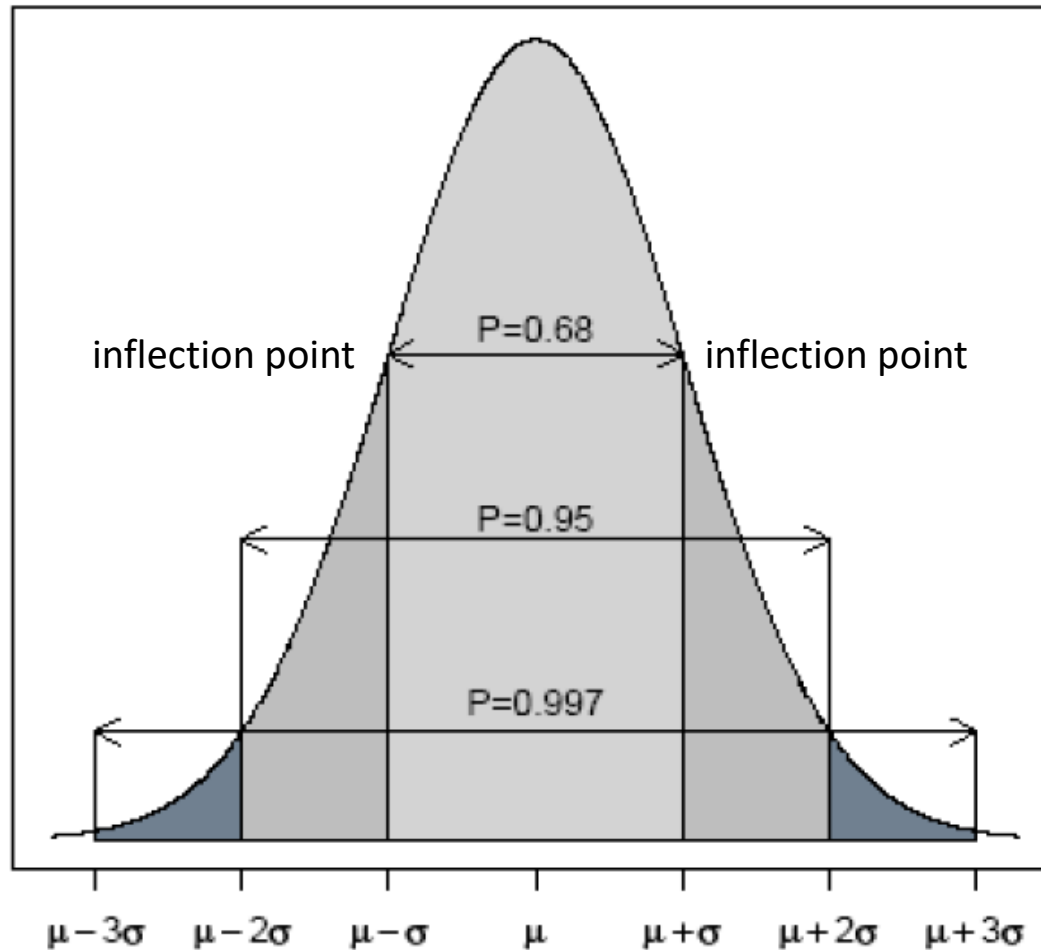
← distribution of the mean ( $n=2$ )

← distribution of the mean ( $n=5$ )

← distribution of the mean ( $n=30$ )

## Density of the Normal distribution

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



Rule of thumb:

A random value  $X \sim N(\mu, \sigma^2)$  has 95% of its probability mass within the following interval:

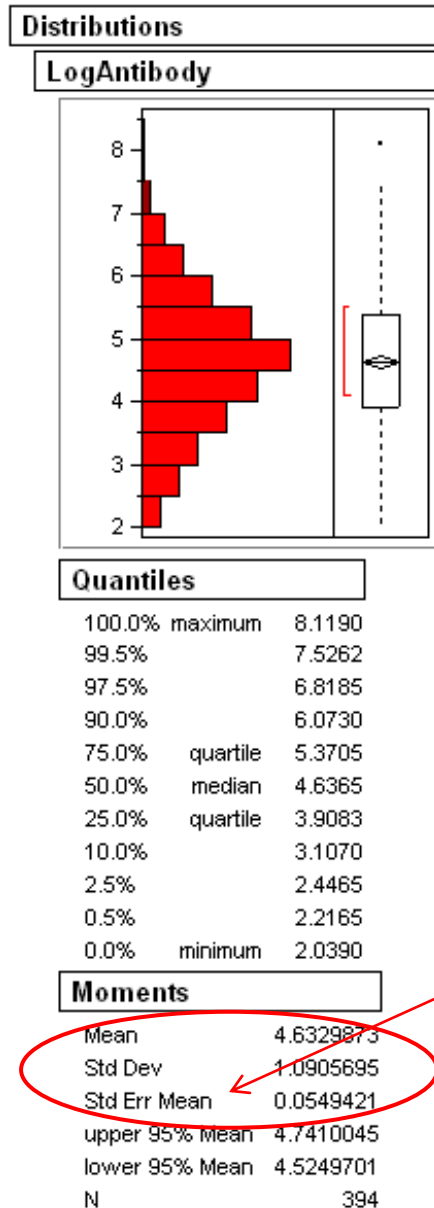
$$[\mu - 2\sigma, \mu + 2\sigma]$$

Or equivalently:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$$



## At which intervals do people usually look?



The expression of antibodies is measured on a log scale in n=394 cells – at intervals for the expression do we look?

PI (prediction interval) or reference range, which covers 95% of the individual observations

$$\bar{x} \pm 2 \cdot sd(x)$$

Reference range: [102.8;909.6]

Confidence Interval (CI) for the expected value  $E(X)$  covers with 95% probability the true value  $E(X)$ .

$$\bar{x} \pm 2 \cdot \frac{sd(x)}{\sqrt{n}} = \bar{x} \pm 2 \cdot se(\bar{x})$$

$$se(\bar{x}) = \frac{sd(x)}{\sqrt{n}}$$

CI: [92.1, 114.8]

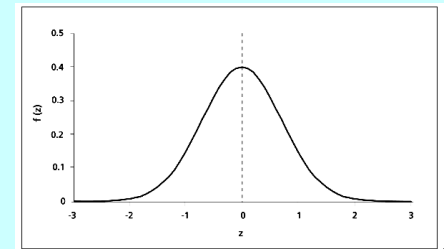
To get to the original scale, we can exp-transform the limits of the intervals

# Distribution of the standardized mean in case of normal distributed observations

$X_1, X_2, \dots, X_n \sim N(\mu_x, \sigma_x^2)$  i.i.d.

Variance  $\sigma_x^2$  is known.

$$T = \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \sim N(0,1)$$



$n \rightarrow \text{big}(> 25) \quad t_{df=n-1} \rightarrow N(0,1)$

$X_1, X_2, \dots, X_n \sim N(\mu_x, \sigma_x^2)$  i.i.d.

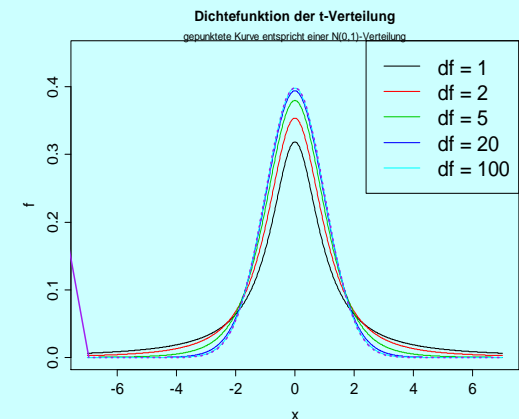
Variance  $\sigma_x^2$  is unknown and is estimated from the data

$$s_x^2 = \hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$T = \frac{\bar{X} - \mu_x}{\frac{s_x}{\sqrt{n}}} \sim t_{n-1}$$

$$se(\bar{x}) = \frac{sd(x)}{\sqrt{n}}$$

se: standard error of the mean  
variation of the estimator



Remark: Since beside the mean also the variance is derived from the random sample we have some additional variation when determining T and the distribution of T gets broader and is given by the  $t_{df=n-1}$

## Construct an exact confidence interval for the expected value if values are normally distributed

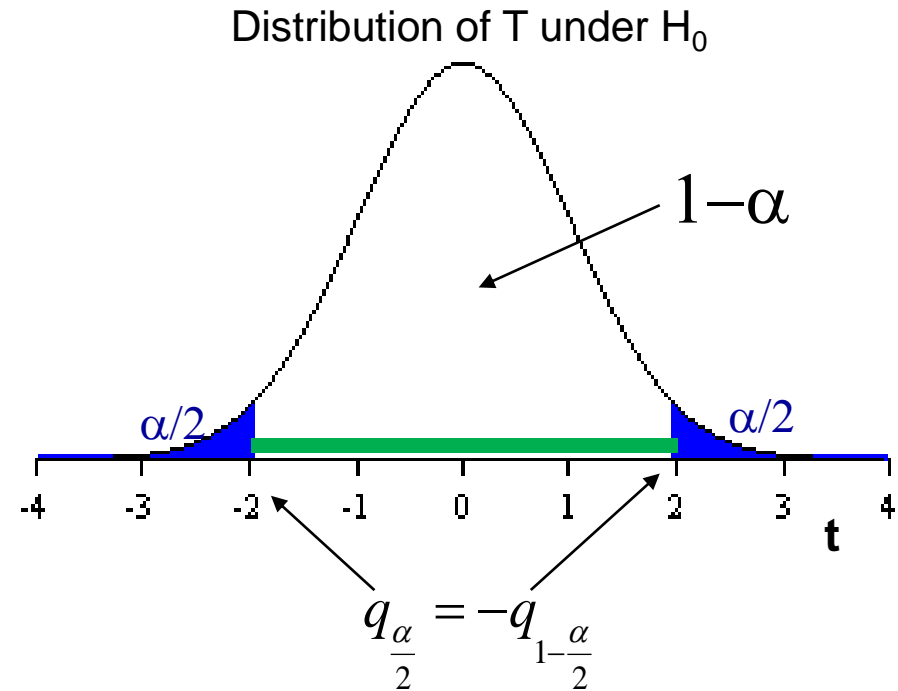
$$X_i \text{ i.i.d. } \sim N(\mu, \sigma^2), E(X) = \mu_x, \text{Var}(X) = \sigma_x^2$$

$$\Rightarrow T = \frac{\bar{X} - \mu_x}{\frac{s_x}{\sqrt{n}}} \sim t_{df=n-1}$$

$$P\left(q_{\frac{\alpha}{2}}^t \leq \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \leq q_{1-\frac{\alpha}{2}}^t\right) = 1 - \alpha$$

$$P\left(-\frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t \leq \bar{X} - \mu_x \leq \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t\right) = 1 - \alpha$$

$$P\left(\bar{X} - \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t \leq \mu_x \leq \bar{X} + \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t\right) = 1 - \alpha$$



95% CI for  $\mu_x$   $\left[ \bar{X} - \frac{\sigma_x}{\sqrt{n}} \cdot q_{0.975}^{t_{n-1}} ; \bar{X} + \frac{\sigma_x}{\sqrt{n}} \cdot q_{0.975}^{t_{n-1}} \right]$

# The exact 95% CI for the expected value if values are normally distributed

$$\bar{x} \pm t_{n-1} q_{97.5\%} \cdot \frac{sd(x)}{\sqrt{n}} \stackrel{n>25}{\approx} \bar{x} \pm z q_{97.5\%} \cdot \frac{sd(x)}{\sqrt{n}}$$

$se(\bar{x})$  : standard error of the mean

Please note that for the exact CI the quantiles of the t-distribution are used. The t-distribution has a parameter df (degree of freedom), which must be set on n-1, where n is the number of observations in the sample. If n gets large (>25) the quantiles of the t-distribution can be approximated by the quantiles of the N(0,1) distribution.

This CI has to be used when the standard deviation  $sd(x)$  is estimated by a i.i.d normally distributed sample:  $X_i \text{ iid } \sim N(\mu, \sigma^2)$ .

# Construct an approximate confidence interval for the expected value it is not assumed that observations $x$ come from a Normal distribution

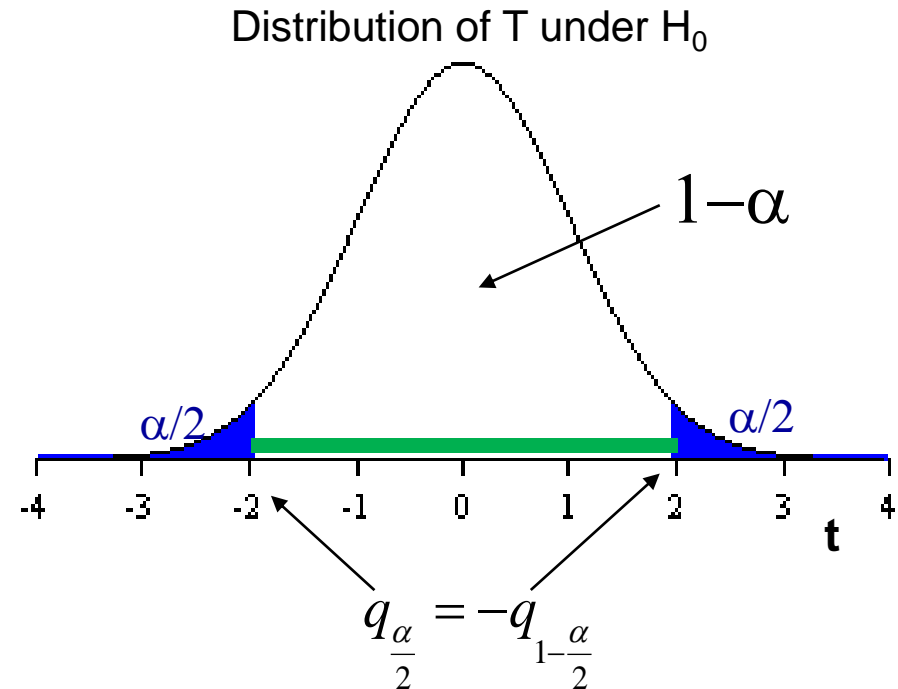
$X_i$  i.i.d.  $i \in 1, \dots, n$ ,  $n > 25$ ,  $E(X) = \mu_x$ ,  $Var(X) = \sigma_x^2$

$$\stackrel{CLT}{\Rightarrow} \bar{X} \overset{a}{\sim} N\left(\mu_x, \frac{s_x^2}{n}\right) \quad \stackrel{CLT}{\Rightarrow} T = \frac{\bar{X} - \mu_x}{\frac{s_x}{\sqrt{n}}} \overset{a}{\sim} N(0,1)$$

$$P\left(q_{\frac{\alpha}{2}}^Z \leq \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \leq q_{1-\frac{\alpha}{2}}^Z\right) \approx 1 - \alpha$$

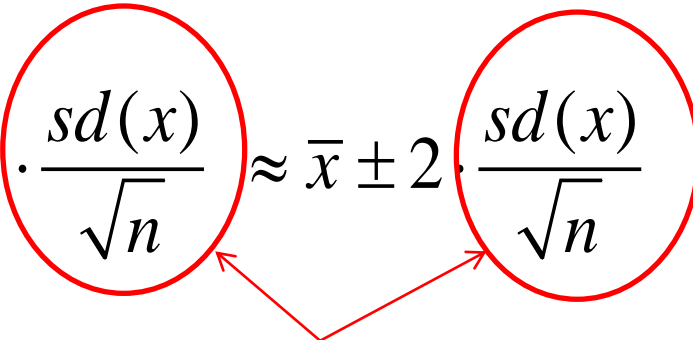
$$P\left(-\frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^Z \leq \bar{X} - \mu_x \leq \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^Z\right) \approx 1 - \alpha$$

$$P\left(\bar{X} - \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^Z \leq \mu_x \leq \bar{X} + \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^Z\right) \approx 1 - \alpha$$



95% CI for  $\mu_x$   $\bar{X} \pm \frac{s_x}{\sqrt{n}} \cdot q_{0.975}^Z \approx \bar{X} \pm 2 \cdot \frac{s_x}{\sqrt{n}}$

**The approximative 95% CI for the expected value**  
it is not assumed that observations  $x$  come from a Normal distribution

$$\bar{x} \pm z_{q_{97.5\%}} \cdot \frac{sd(x)}{\sqrt{n}} \approx \bar{x} \pm 2 \cdot \frac{sd(x)}{\sqrt{n}}$$


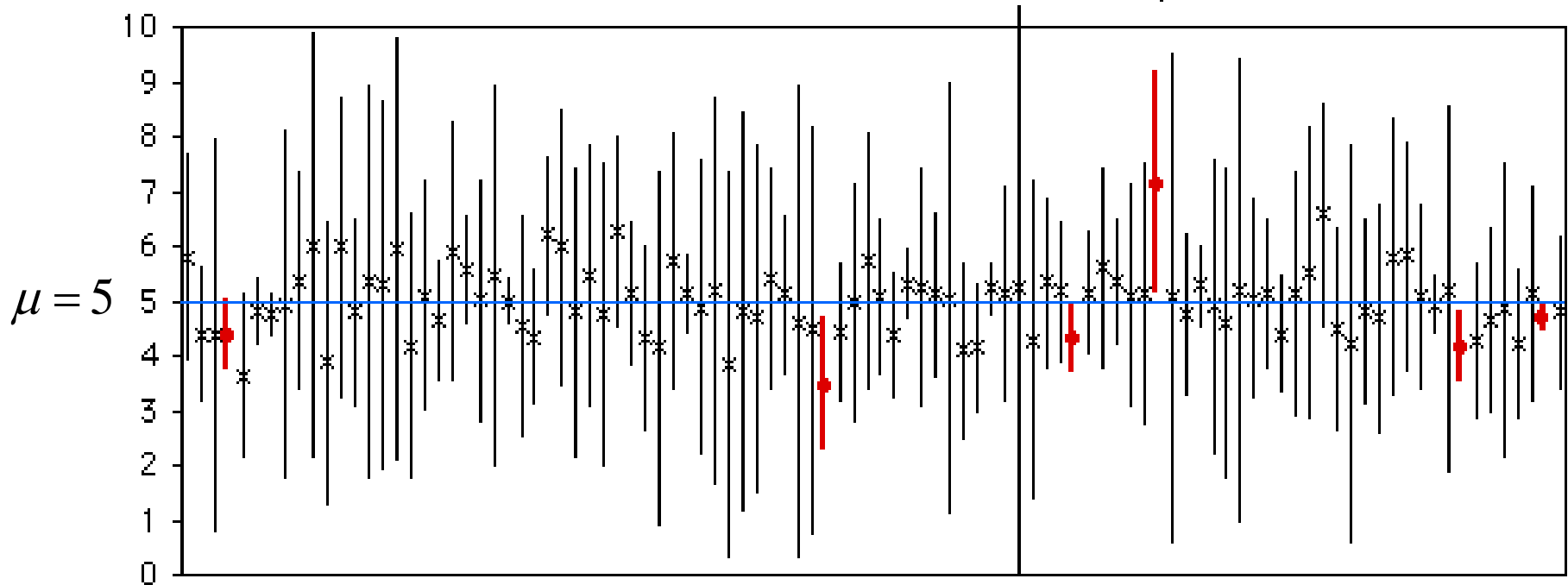
$se(\bar{x})$  : standard error of the mean

- This is the 95% confidence interval for the expected value  $\mu_x = E(X)$ .
- In the light of the data, all plausible values of the population's mean lay in this CI.
- The width of the CI is proportional to the standard deviation of the data.
- The width of the CI is proportional to the inverse of the square root of  $n$  – hence it gets smaller if we have more observations
- The width of the CI is proportional to the quantile which gets smaller if we relax the significance level from  $\alpha=5\%$  to a larger value, e.g.  $\alpha=10\%$ .

# The CI is as random as the sample

$$\bar{x} \pm 2 \cdot \frac{sd(x)}{\sqrt{n}}$$

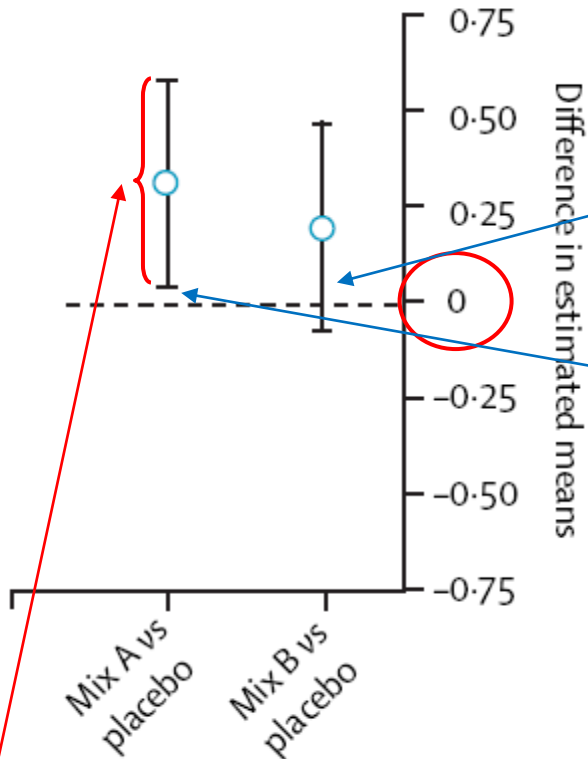
mean and 95% confidence intervals for 100 samples, N=3



95 out of 100 95%-CI for  $\mu$  do cover the true population parameter  $\mu=5$  when simulating 100 random samples from a population following  $N(\mu=5, \sigma^2)$ .  
With a 95%-CI we have a risk of 5% that our random sample was not typical for the population and the true population parameter is not contained by the CI.

# Interpretation of a confidence interval

Example from paper on hyperactivity form McCann et al.



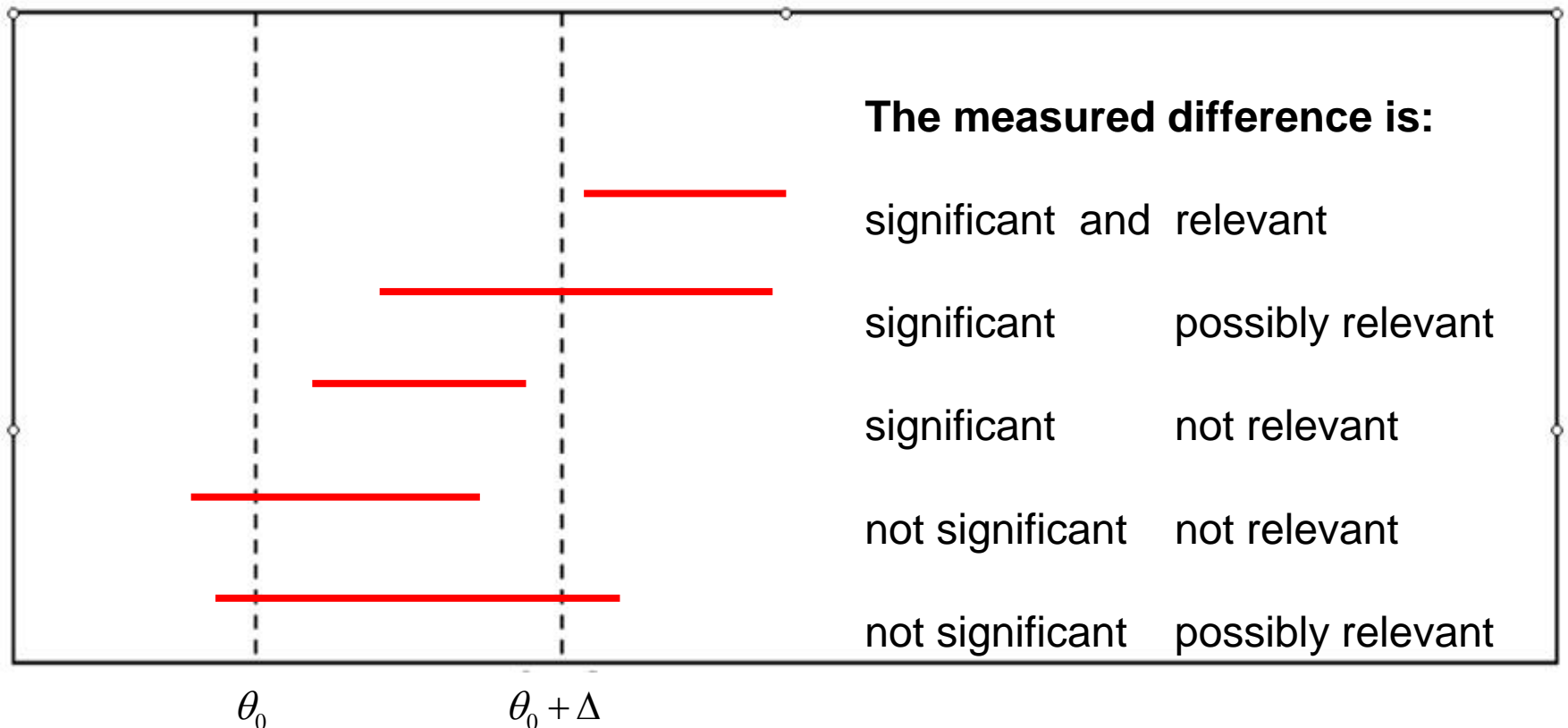
- The CI covers all plausible values for the true mean-difference – here the true treatment effect
- If 0 is covered by the CI it is plausible that the treatment effect is 0 – we have no evidence against  $H_0$ , that the treatment has no effect.
- If 0 is **not** covered by the CI, we say that the treatment effect is **significantly** different from 0.
- To have a reasonable chance (80%) to claim a relevant treatment effect to be significant we must plan the sample size to be large enough to be able to find a the effect to be significant if existing.

Here we see a 95% CI of the difference of the mean hyperactivity under placebo and under treatment with Mix A indicating a significant effect of Mix A.

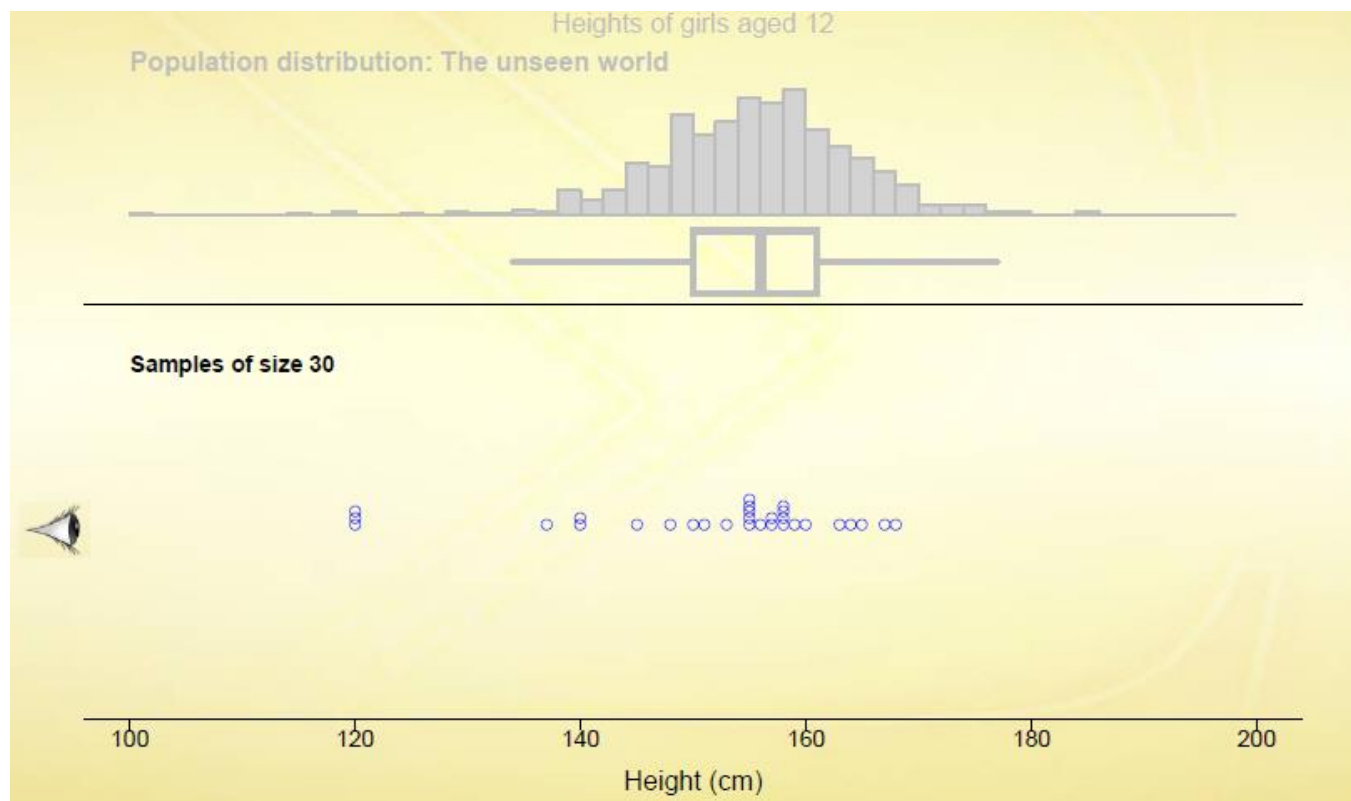


With a confidence interval we can decide:  
Is there a significant difference to a postulated value  $\theta_0$ ?  
Is the difference relevant ( $>\Delta$ )?

Draw CIs that correspond to the description on the right



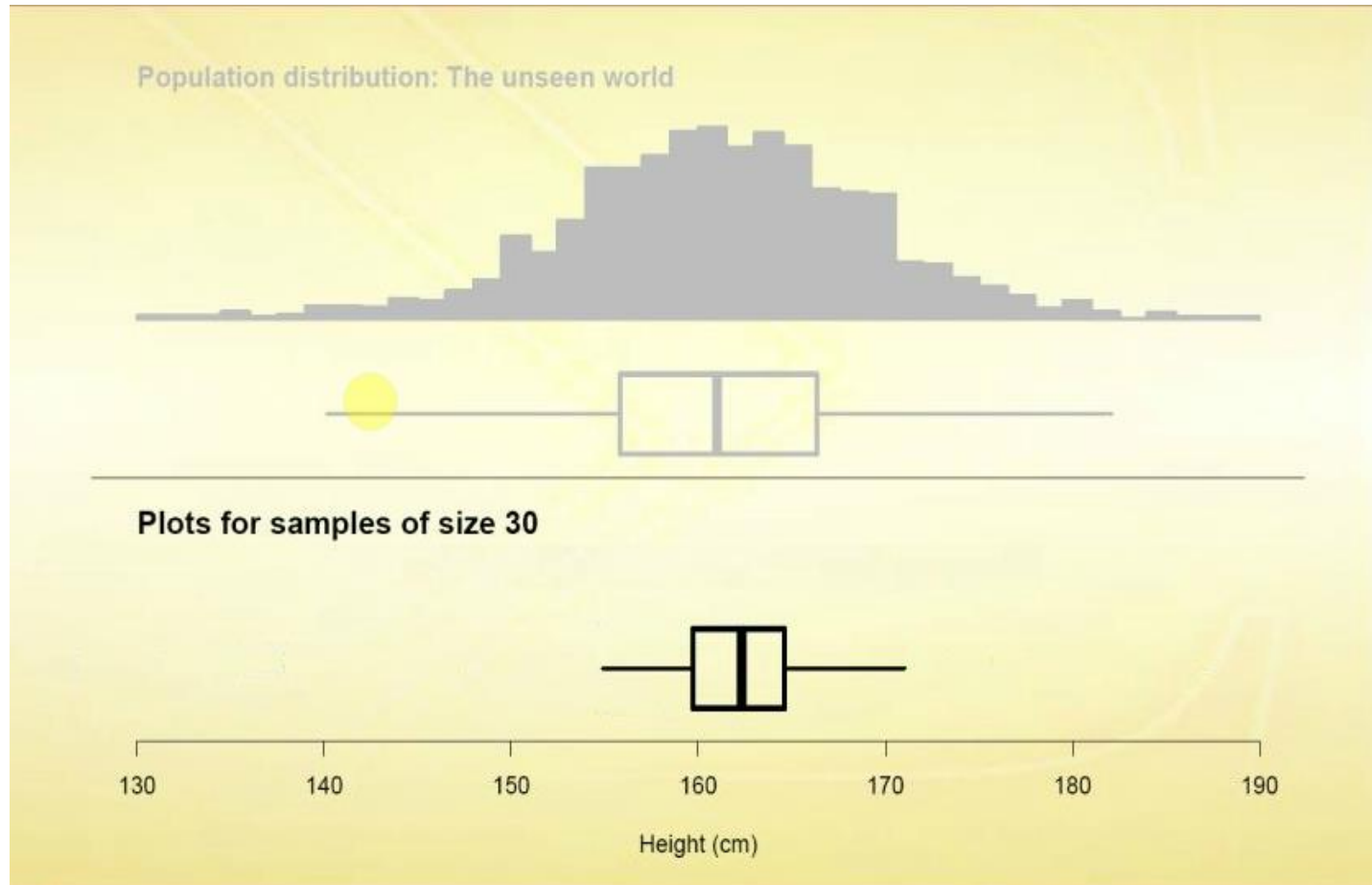
# Where is the center of the population?



$n=30$

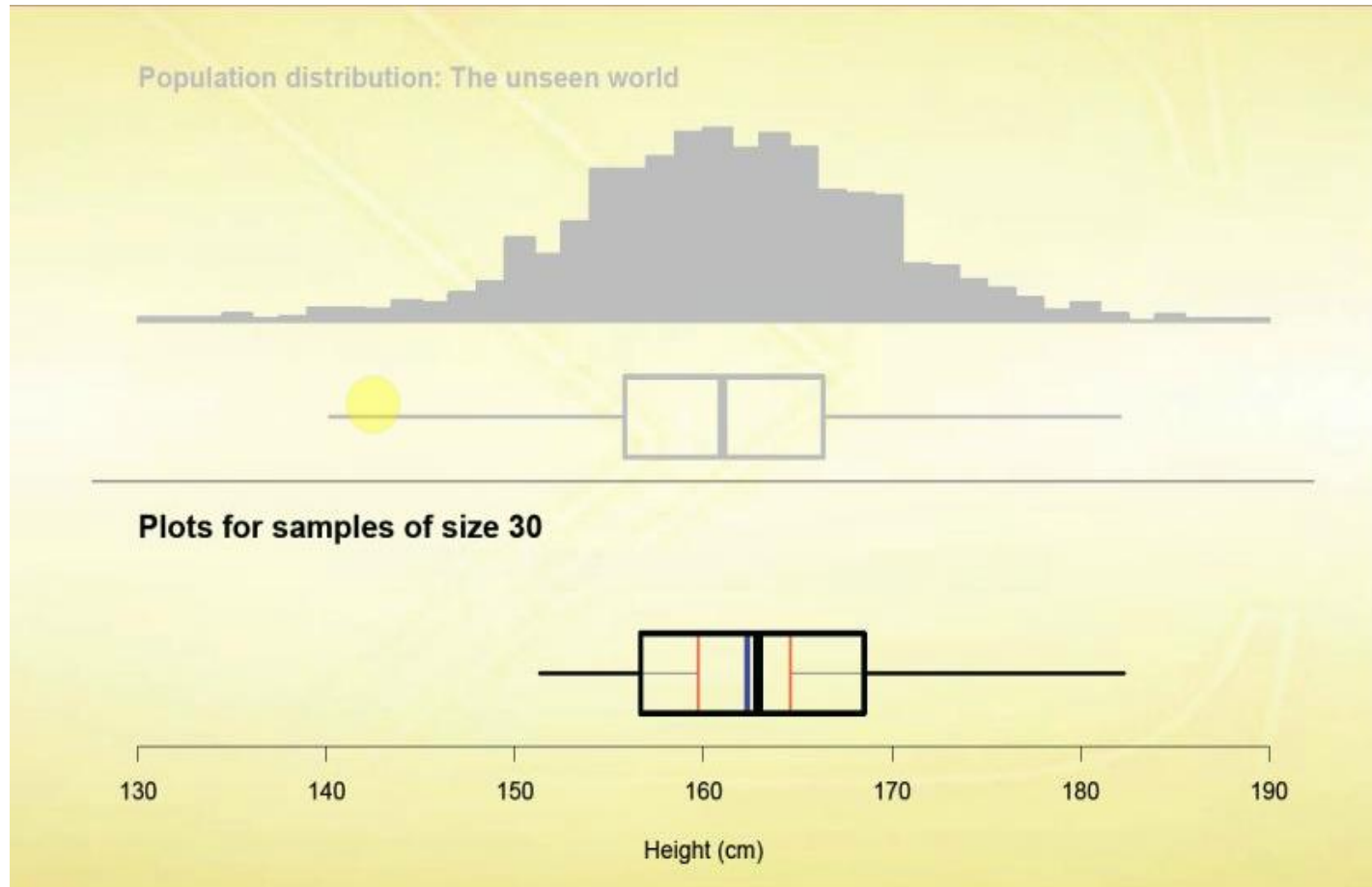
[CtsVar\\_1samp\\_Dots30.pdf](#)

# Where is the center of the population?



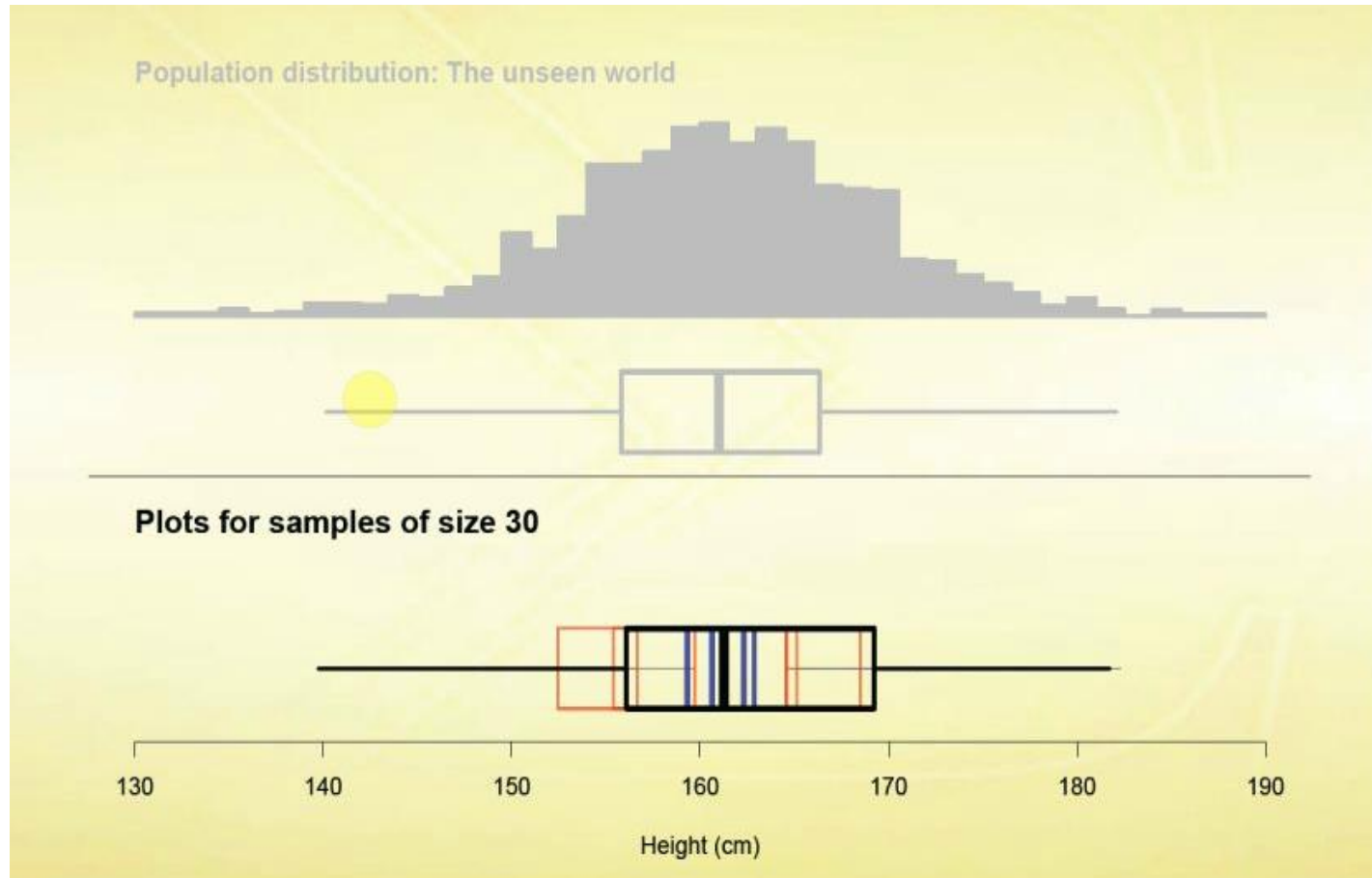
Visualize boxplot with memory

# Where is the center of the population?



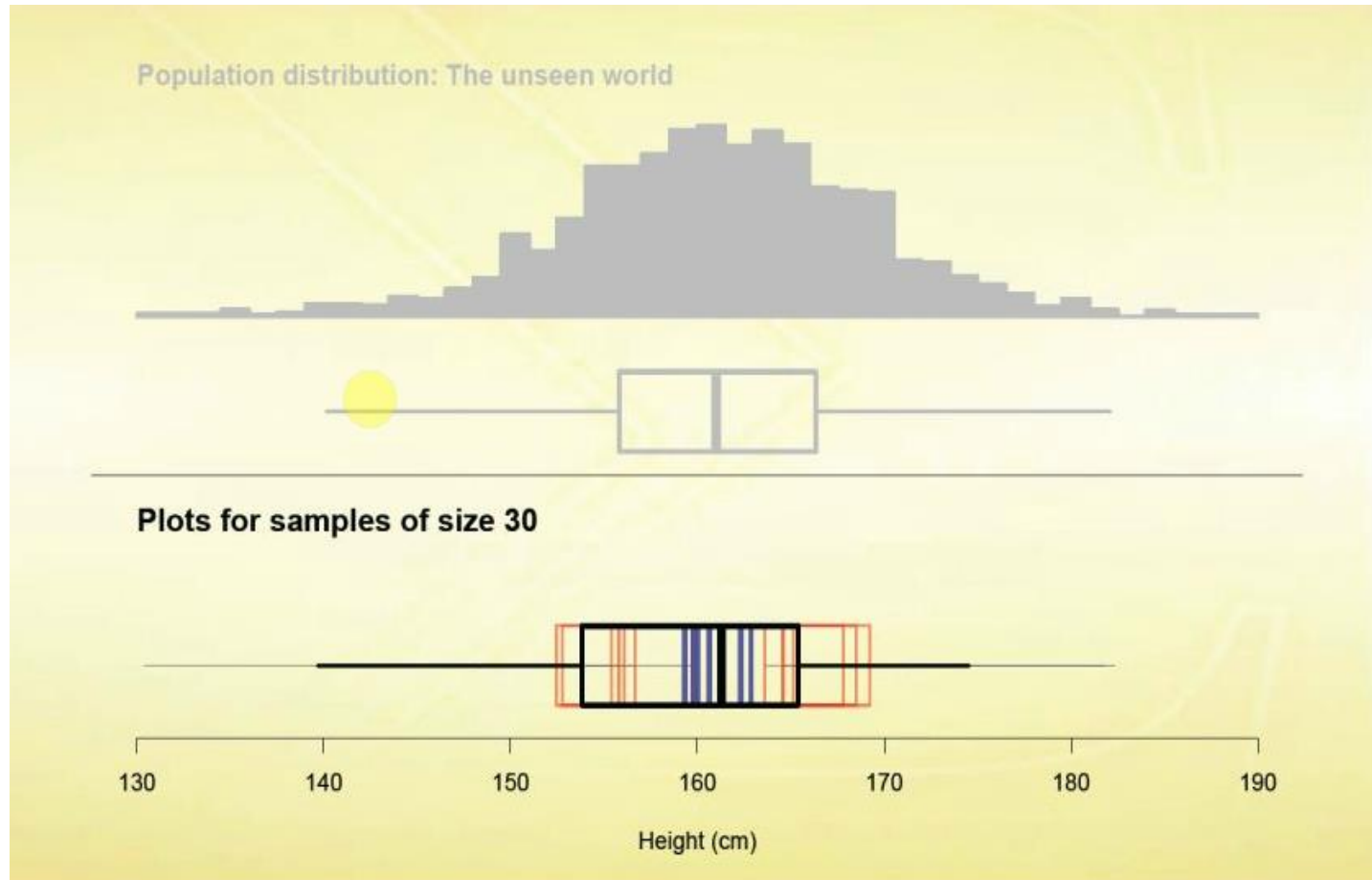
Visualize boxplot with memory

# Where is the center of the population?



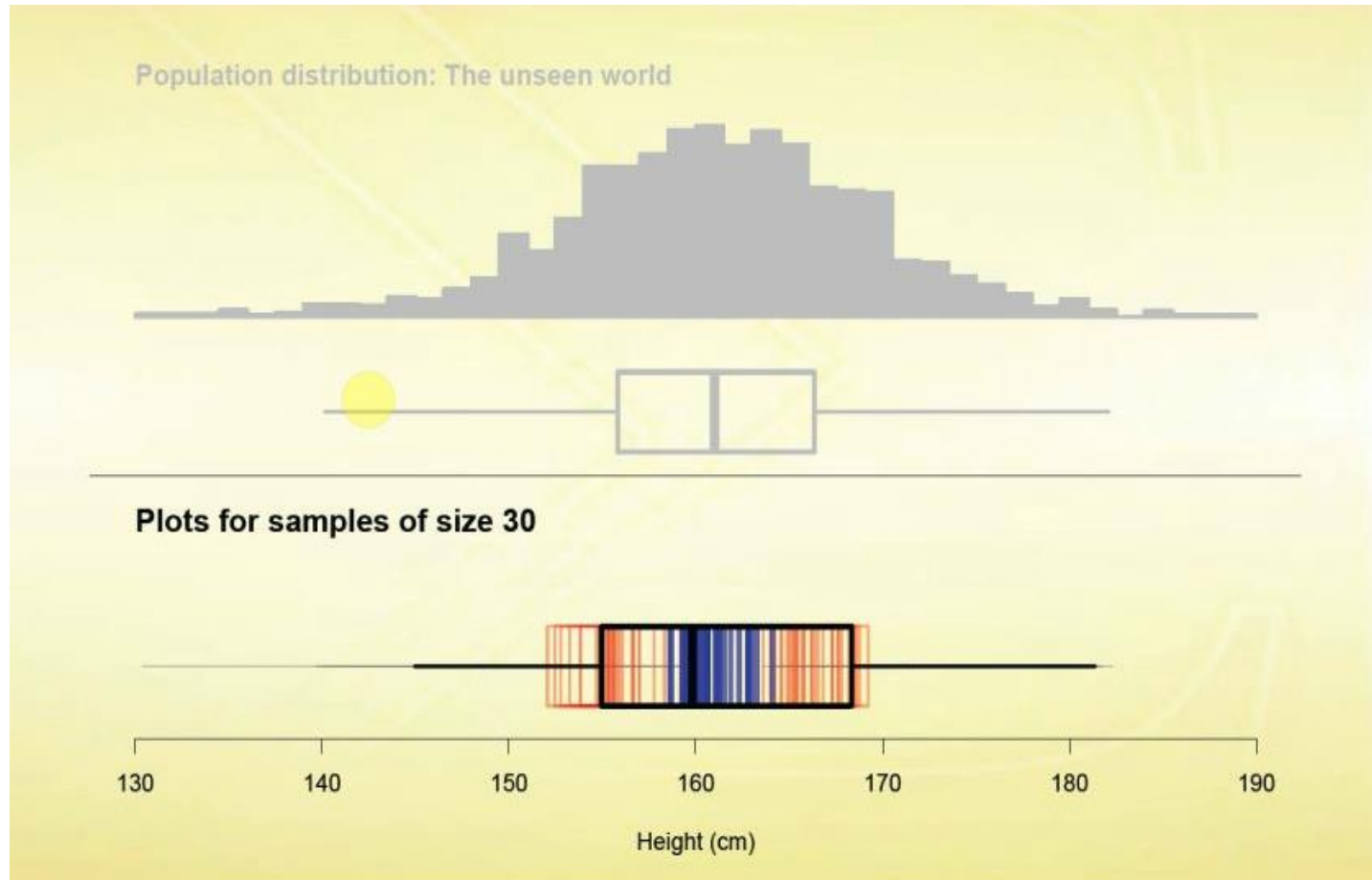
Visualize boxplot with memory

# Where is the center of the population?



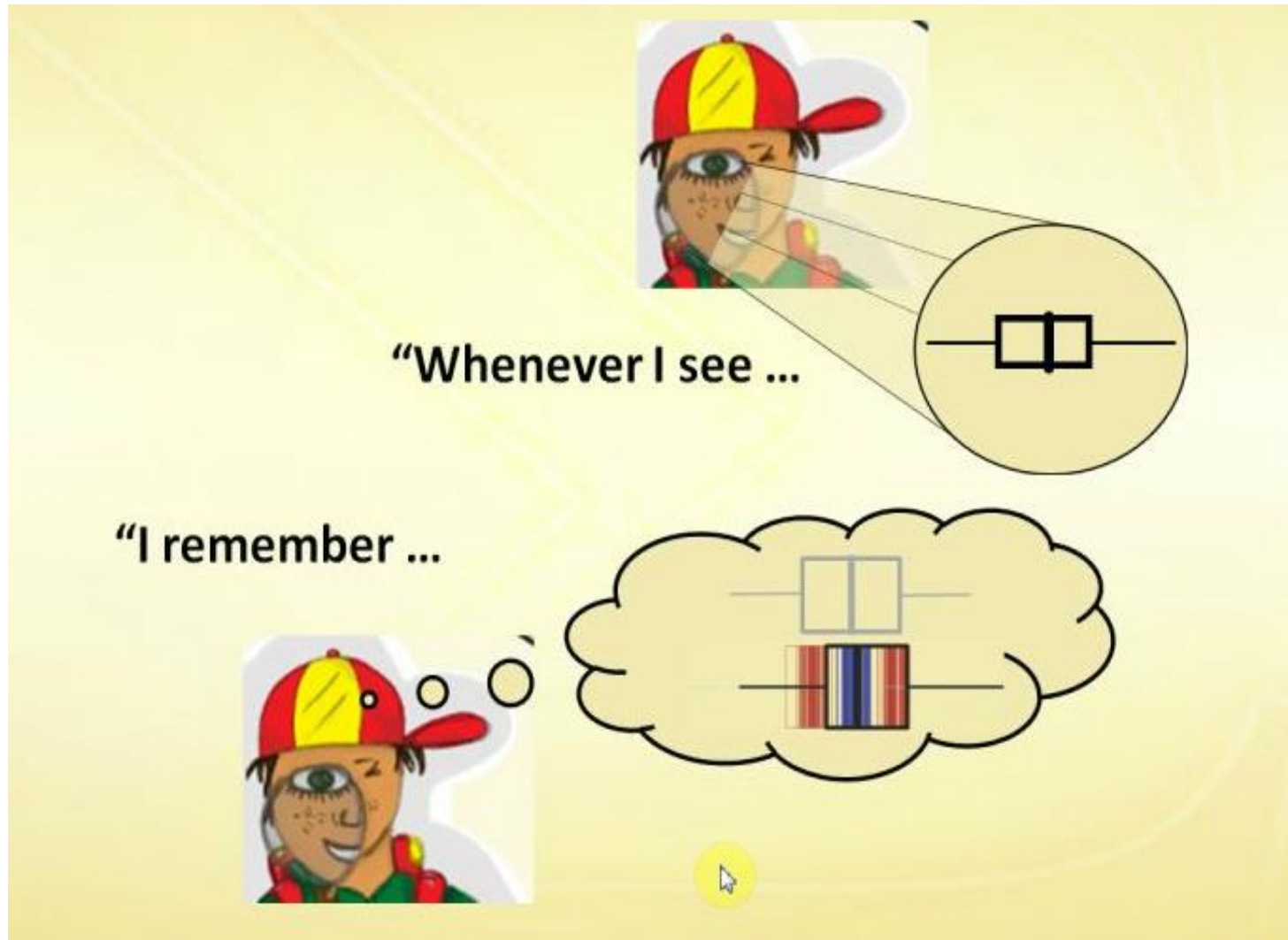
Visualize boxplot with memory

# Where is the center of the population?



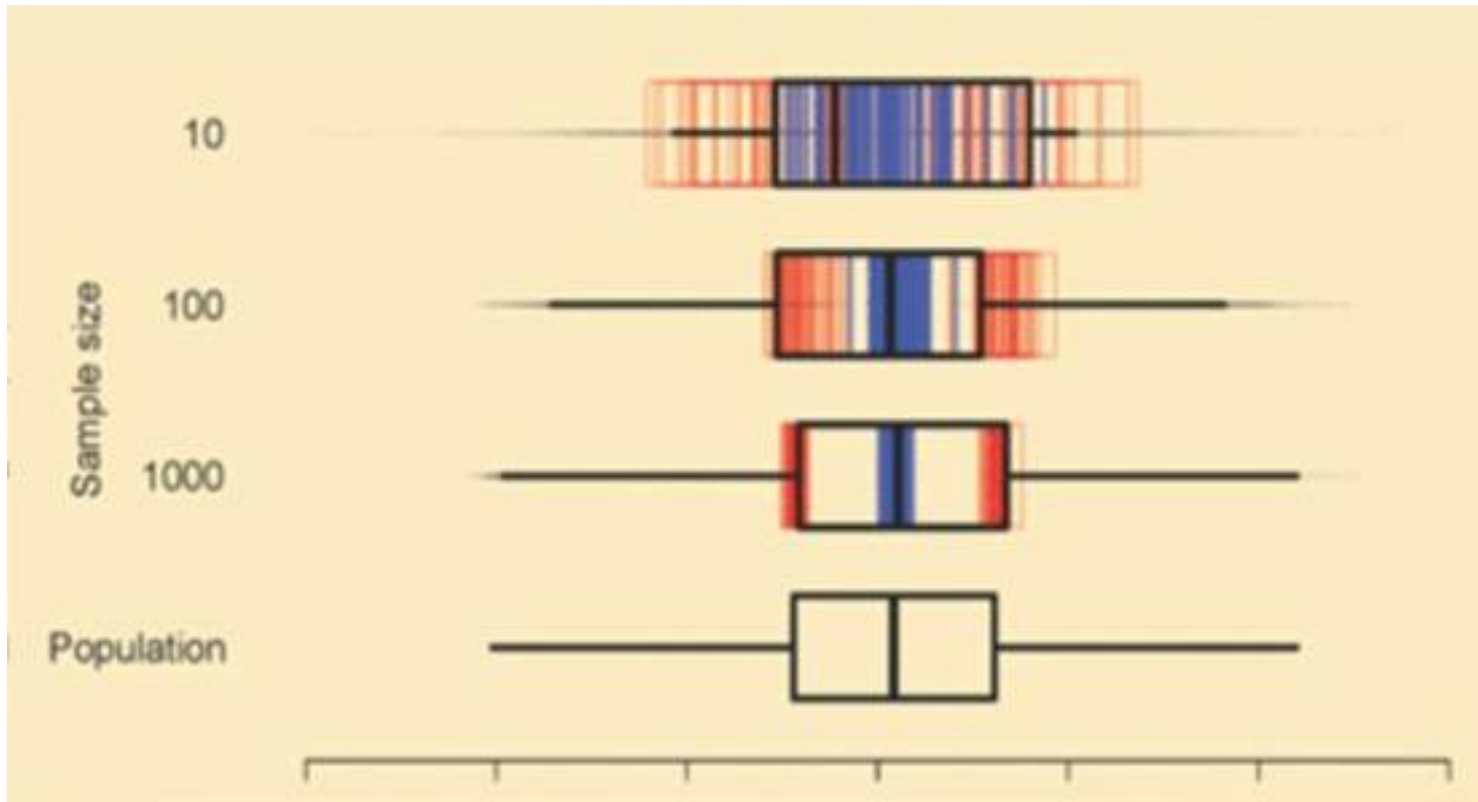
Vizualize boxplot with memory

# Where is the center of the population?





Where is the center of the population?  
We get more certain with increasing sample size

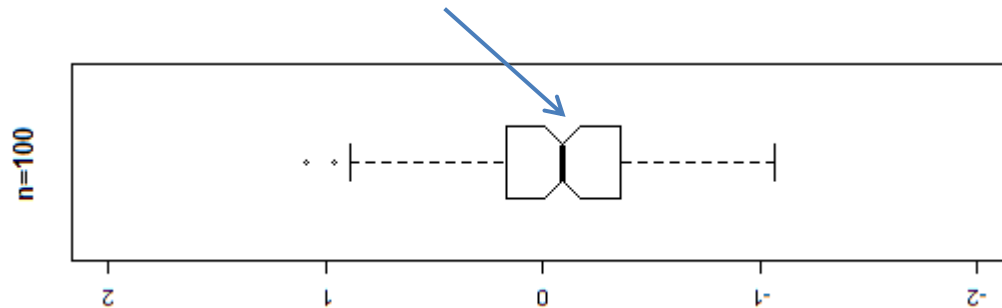


# How sure can I be about the true parameter value?

## Goal:

We would like to determine from our sample/observations an interval, which covers the true parameter value with a probability of 95%.

```
boxplot(x, notch=TRUE)
```



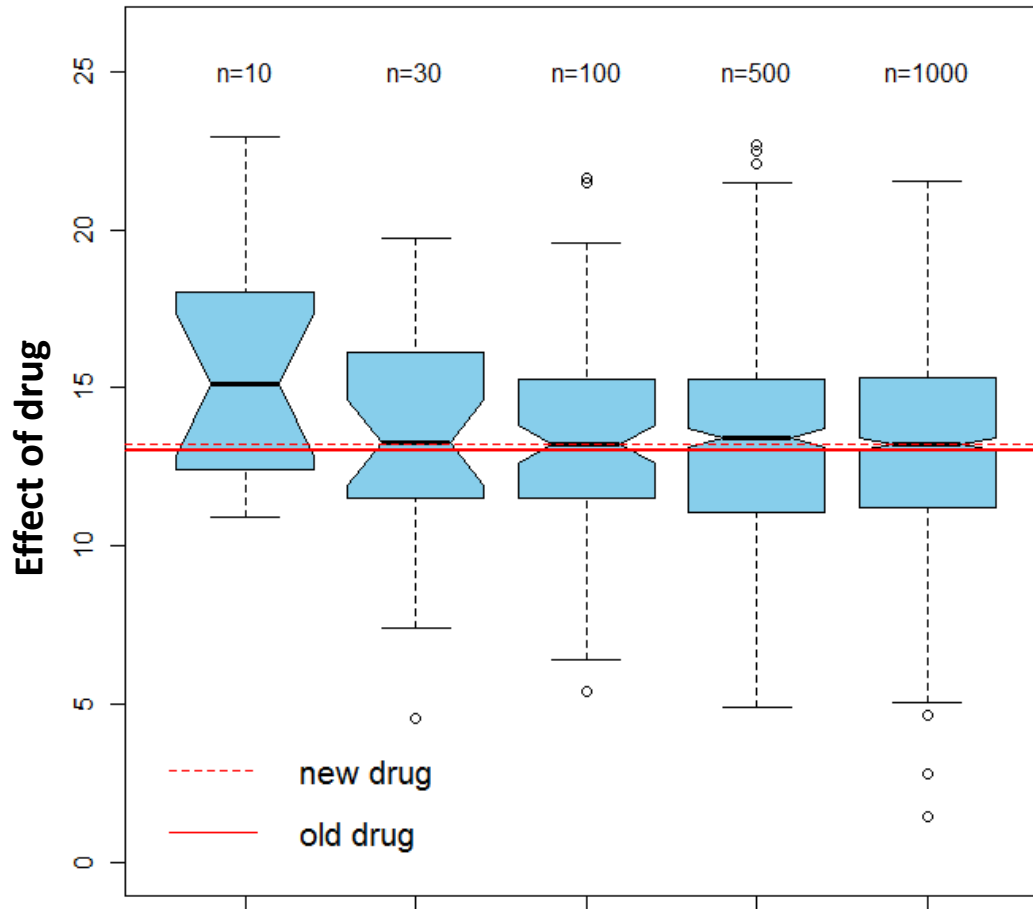
$\pm 1.58 \text{ IQR} / \sqrt{n}$

The notch covers the  
population median  
«quite certain»

# Significance does not imply relevance

## Everything gets significant if the sample size is large enough

Sample with different sample sizes drawn from a Normal distribution with expected value of 13.1



$$H_0: \mu_0 = \mu_{\text{old-drug}} = 13$$

$$H_A: \mu > 13$$

Assume true median of the new drug is 13.1 which would be no relevant improvement compared to old drug value 13

With increasing sample size the 95% confidence interval for the true median gets smaller, while  $\alpha$  stays the same and the power increases to find a significant difference to the old mean of 13.

- To ensure relevance of a significant test one should formulate a relevant  $H_A$ .
- Non-significance could be caused by a too small sample.