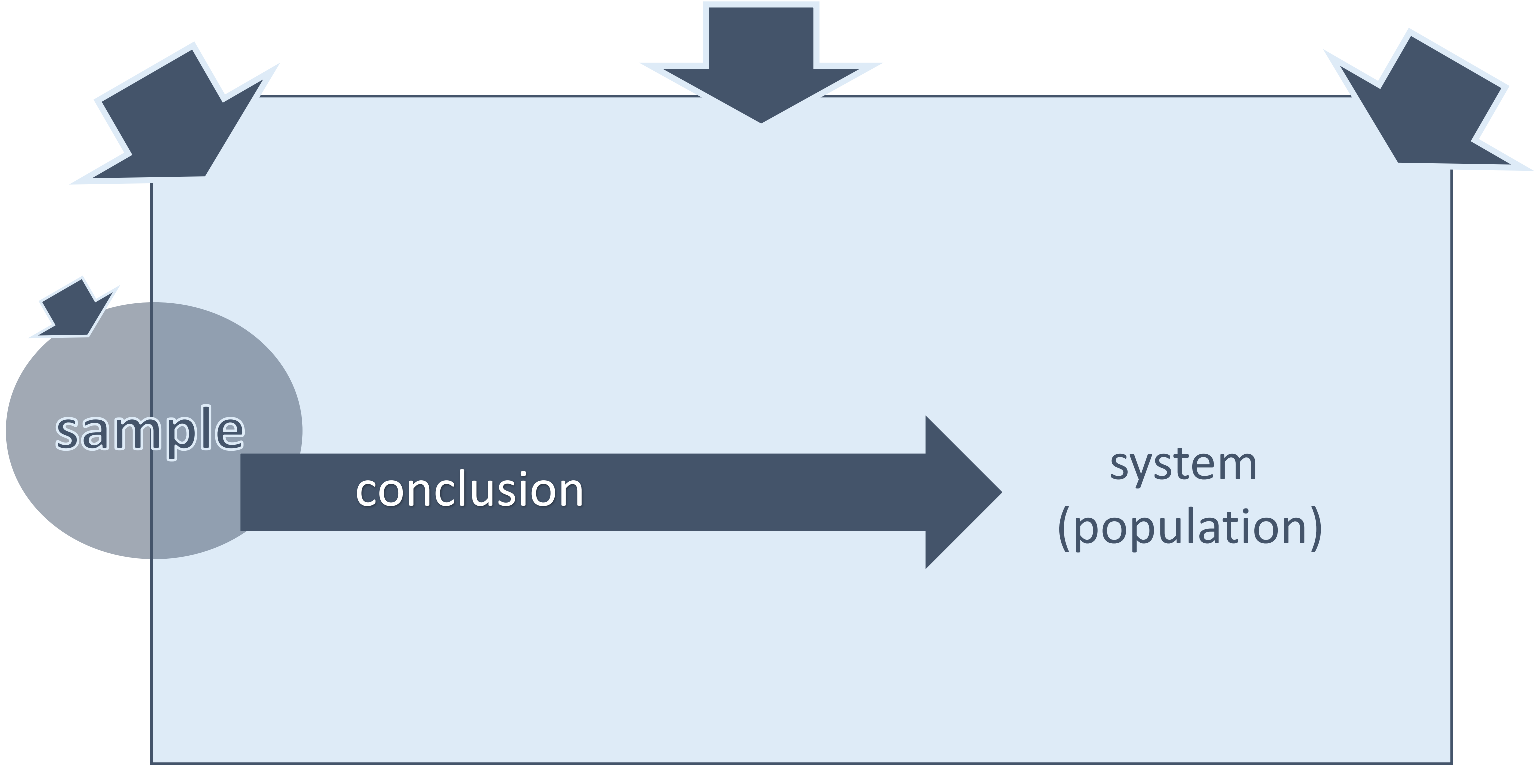
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Introduction to biostatistics and FSharp.Stats

CSB COMPUTATIONAL SYSTEMS BIOLOGY

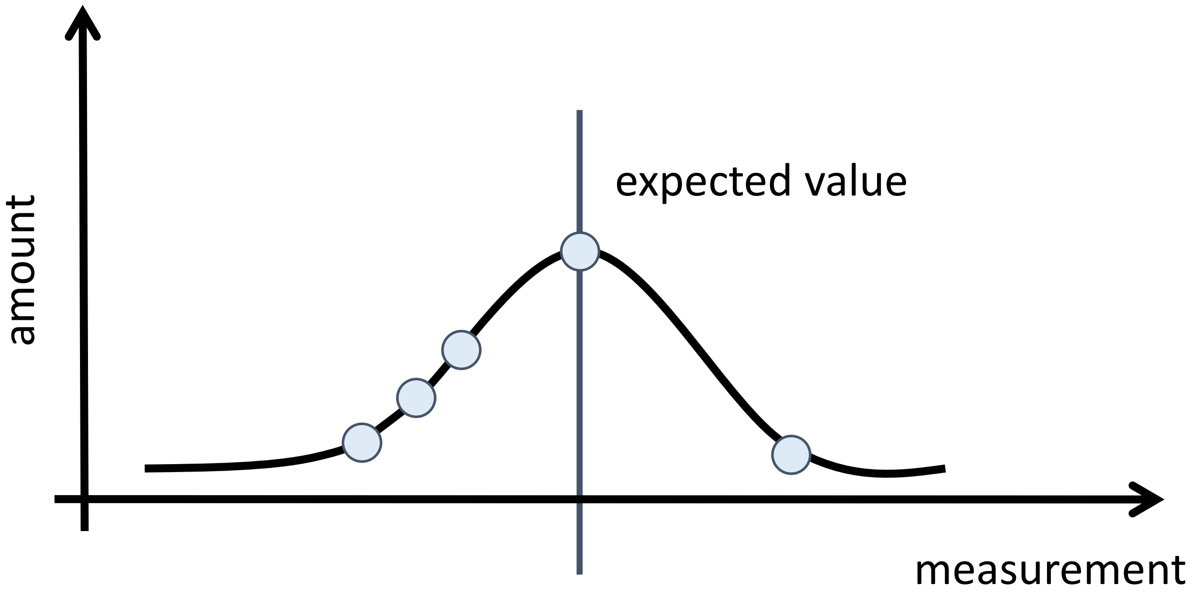
Benedikt Venn
Computational Systems Biology
Kaiserslautern University of Technology

General goal in statistics

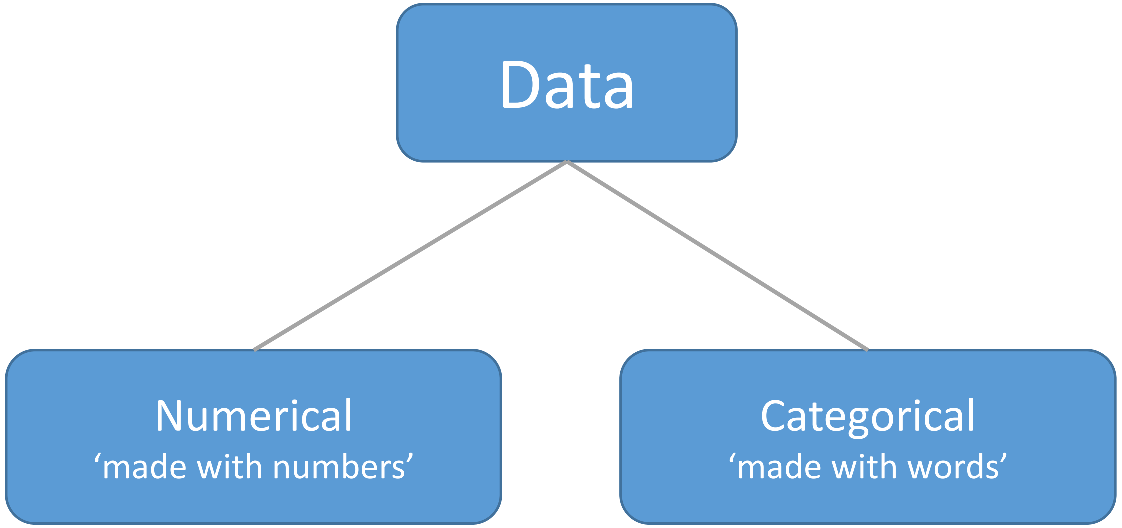


- Drawing conclusions from sample to population

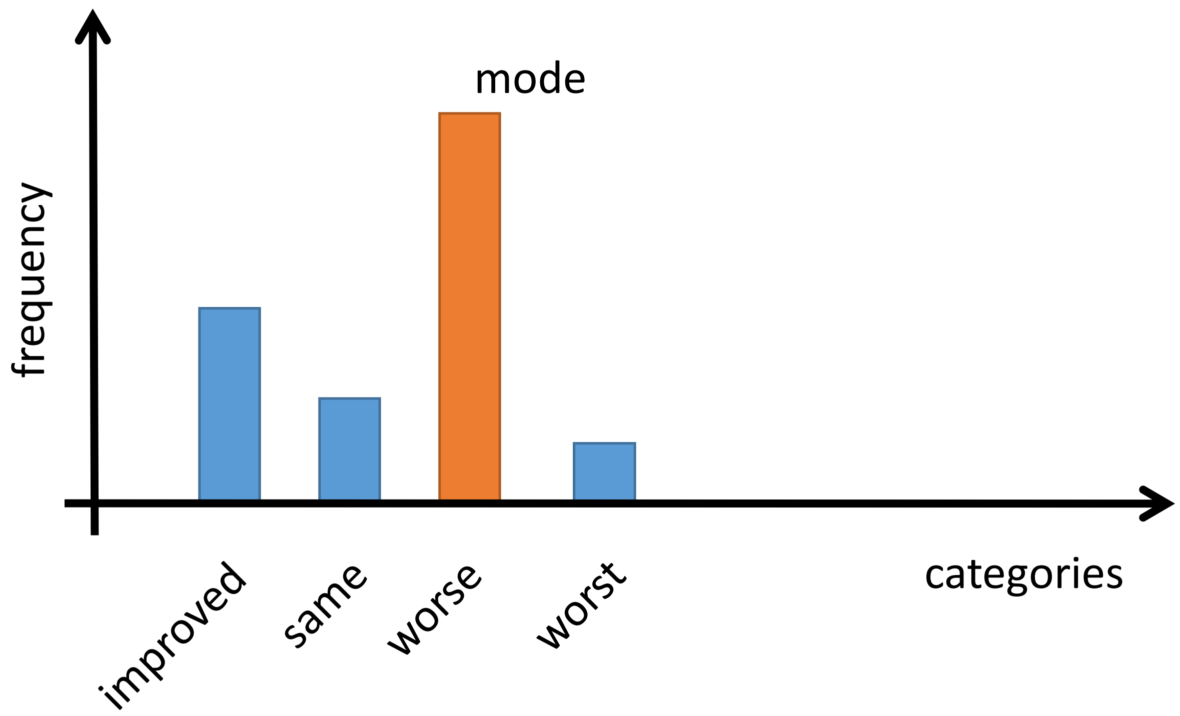
Central tendency

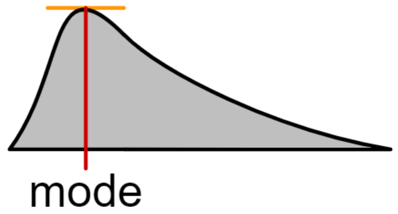


- Finding the expected value by measures of the central tendency using (type L) point estimators


Data
Numerical
'made with numbers'
Categorical
'made with words'

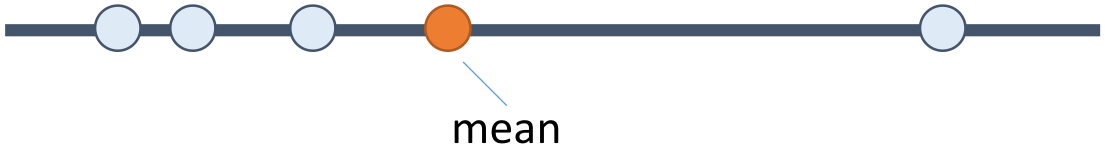
Measures of central tendency: mode



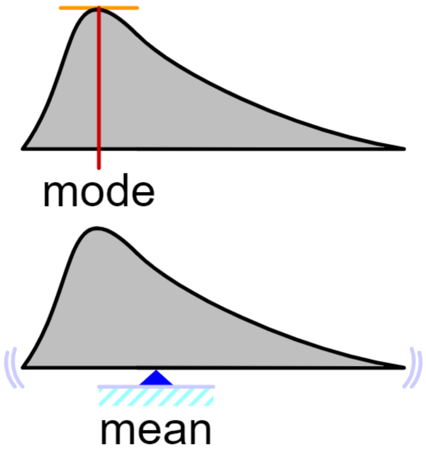


- The mode is the most frequently occurring category

Measures of central tendency: mean



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



- The mean is not robust against outliers (equally influenced by all values)

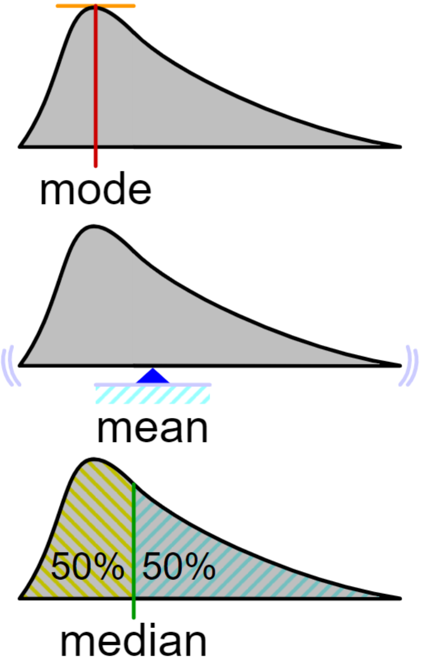
```
open FSharp.Stats

let x = [|11.0; 13.0; 14.5; 18.0; 10.0|]

let meanOfX    = x |> Seq.mean
```

FSharp Interactive

```
val meanOfX : float = 13.3
```

Measures of central tendency: median

$$P(X \leq m) = P(X \geq m) = \int_{-\infty}^{m} f(x)\,dx = \frac{1}{2}.$$



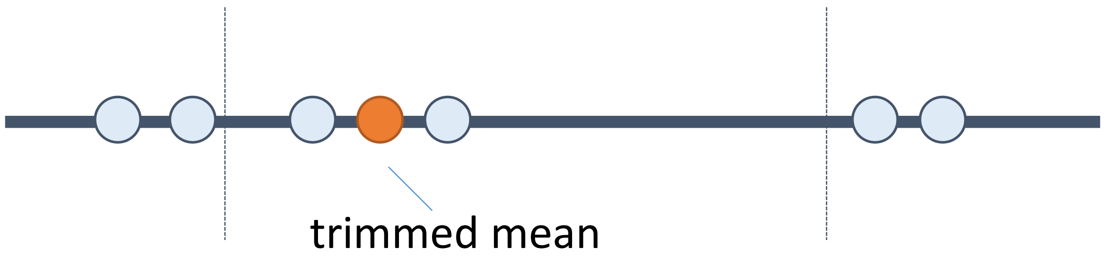
- The median is that value such that half of data points fall above it an half below it
  => more robust against outliers

```fsharp
open FSharp.Stats

let x = [|11.0; 13.0; 14.5; 18.0; 10.0|]

let medianOfX = x |> Seq.median
```

FSharp Interactive

```
val medianOfX : float = 13.0
```

Trimmed mean

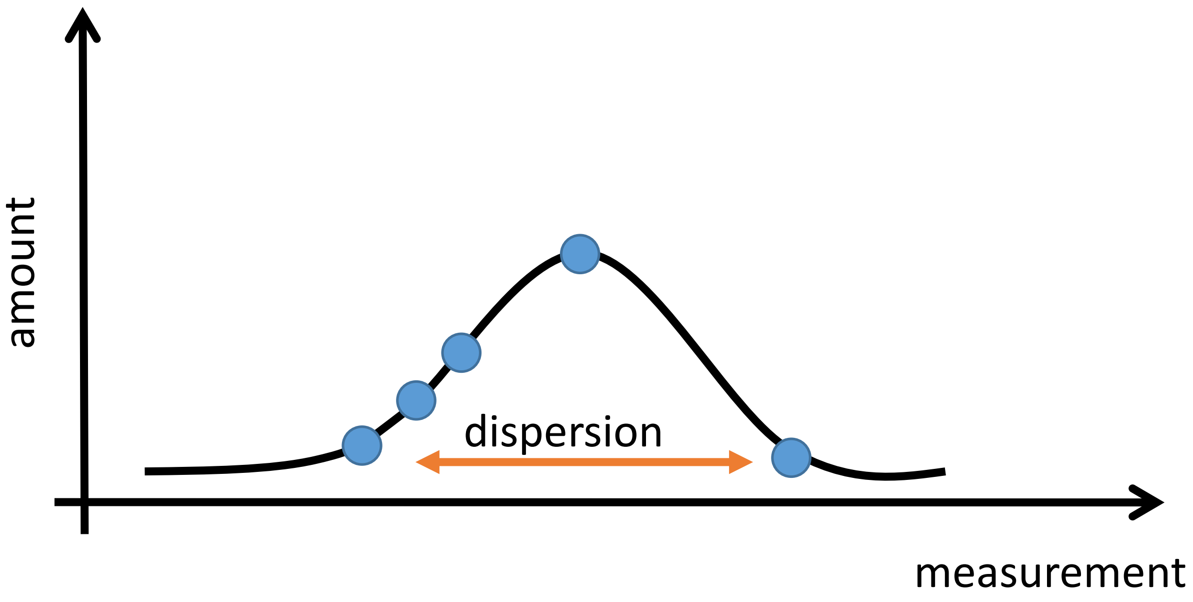


- A trimmed mean involves the calculation of the mean after discarding given parts of a sample at the high and low end
- Typically 5% to 25% of the values are discarded at both ends

Describing distributions

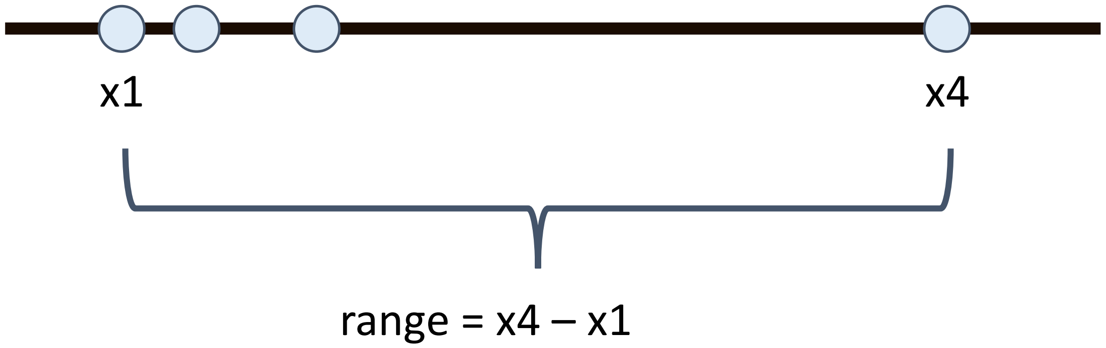
- Central tendency
  - mode
  - mean
  - median
  - trimmed mean
- Dispersion
  - range
  - mean (absolute) deviation
  - variance & standard deviation
  - coefficient of variation

Estimating dispersion



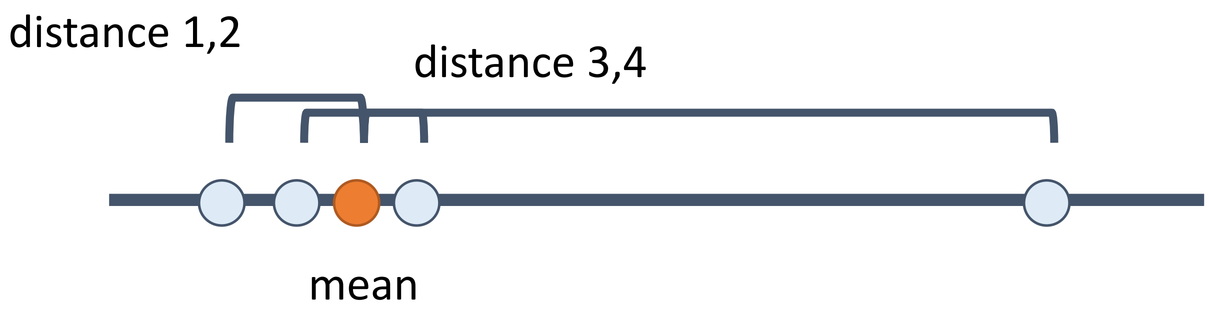
- Estimating the spread/dispersion of the data distribution

The range



- The range is the difference between the highest and lowest value => not robust against extrema

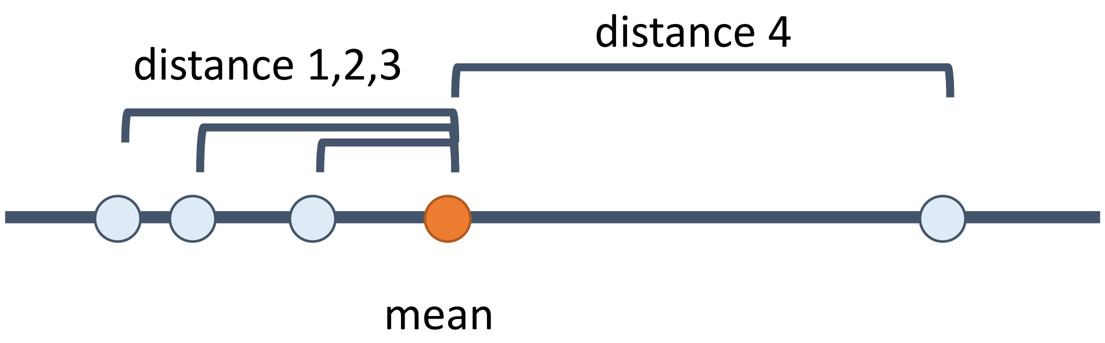
Mean deviation of a sample



$$MD = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$$

- The sum of the absolute amount of deviations from the mean divided by their number

Variance and Standard Deviation of a sample



- Variance: Sum of all squared distances divided by their number

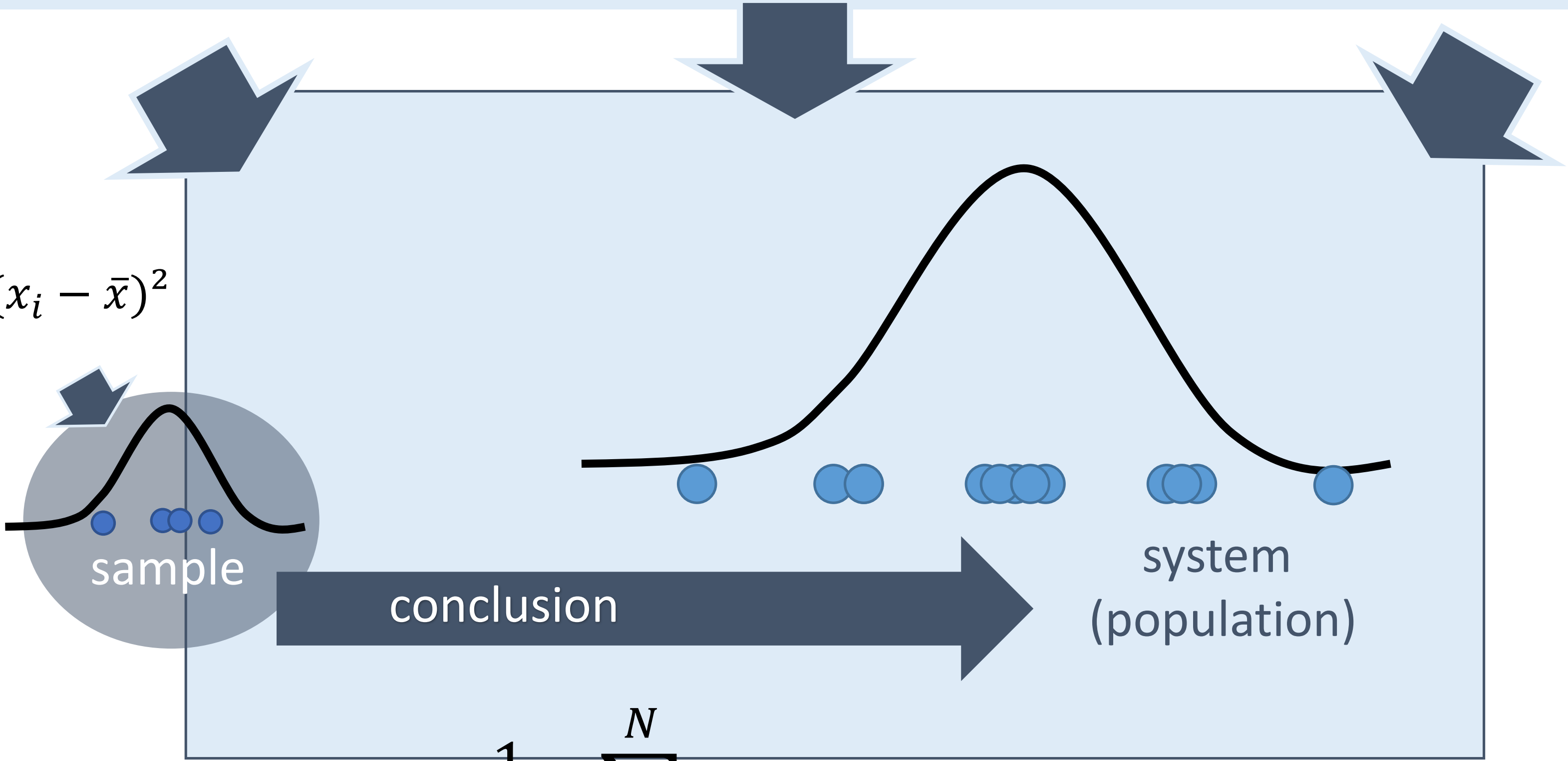
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Standard Deviation is the square root of the variance to get back to the original units

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

The Variance and Standard Deviation of a population

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



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

- Variance: Sum of all distance quadrates divided by the degrees of freedom (N-1)

The Variance and Standard Deviation of a population
- Bessel's correction -

sample variance

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

population variance

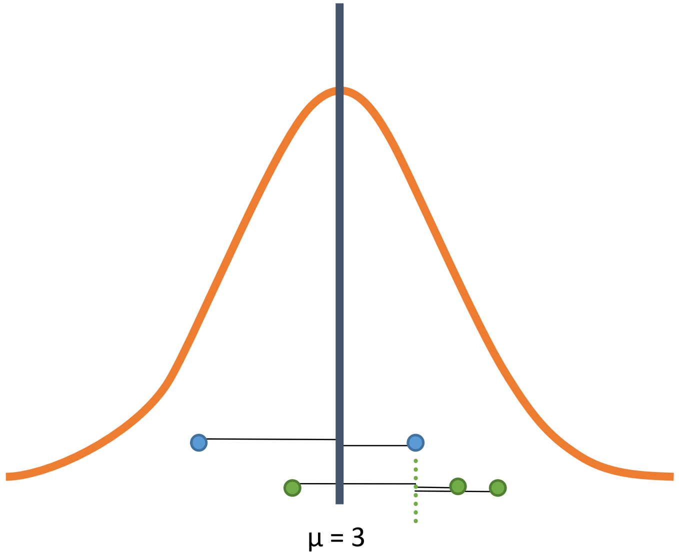
$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

3 independent observations from population (μ = 3)

i	x_i	$x_i - \mu$
1	5	5 - 3 = 2
2	0	0 - 3 = -3
3	?	?

3 independent observations from population ($\bar{x}$ = 5)

i	x_i	$x_i - \bar{x}$
1	7	7 - 5 = 2
2	6	6 - 5 = 1
3		



Computational Systems Biology

```fsharp
open FSharp.Stats

let x = [|11.0; 13.0; 14.5; 18.0; 10.0|]

let stDevPop    = x |> Seq.stDevPopulation
let stDevSample = x |> Seq.stDev
```

FSharp Interactive

```
val stDevPop : float = 2.821347196
val stDevSample : float = 3.154362059
```

Coefficient of variation

$$c_v = \frac{\sigma}{\mu}$$

σ = standard deviation

μ = mean

- The coefficient of variation represents the ratio of the standard deviation to the mean.
  It is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other

```fsharp
open FSharp.Stats

let x = [|11.0; 13.0; 14.5; 18.0; 10.0|]

let cvOfX = x |> Seq.cv
```

FSharp Interactive

```
val cvOfX : float = 0.2371700796
```

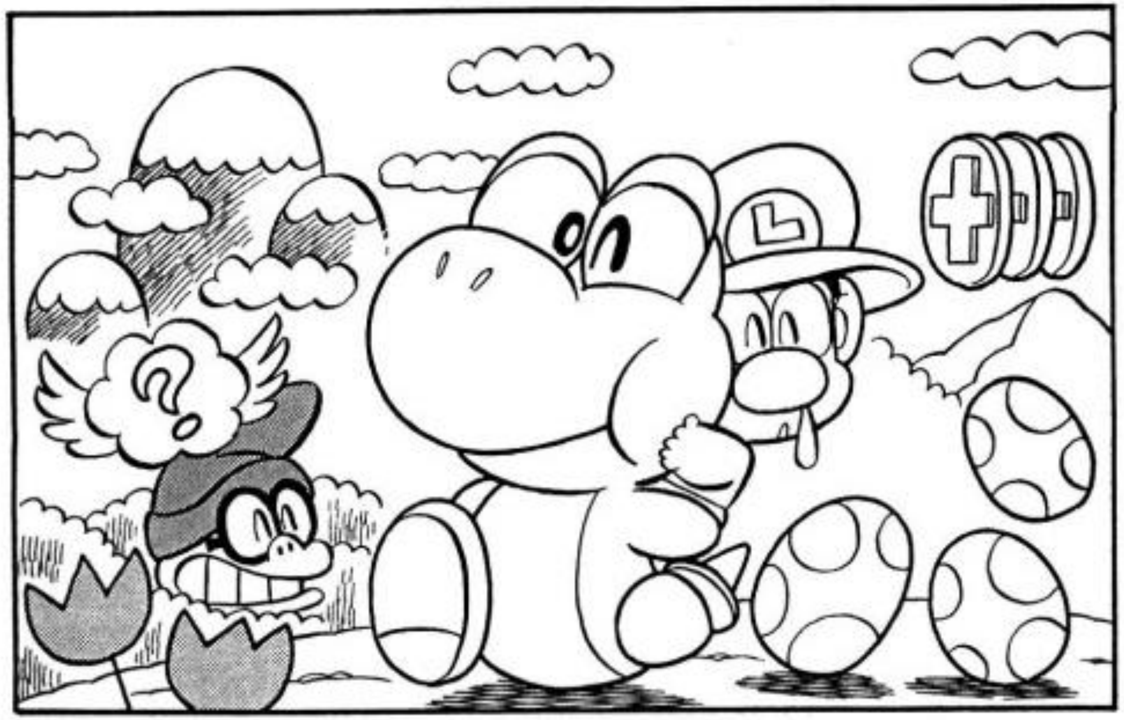
Computational
Systems Biology

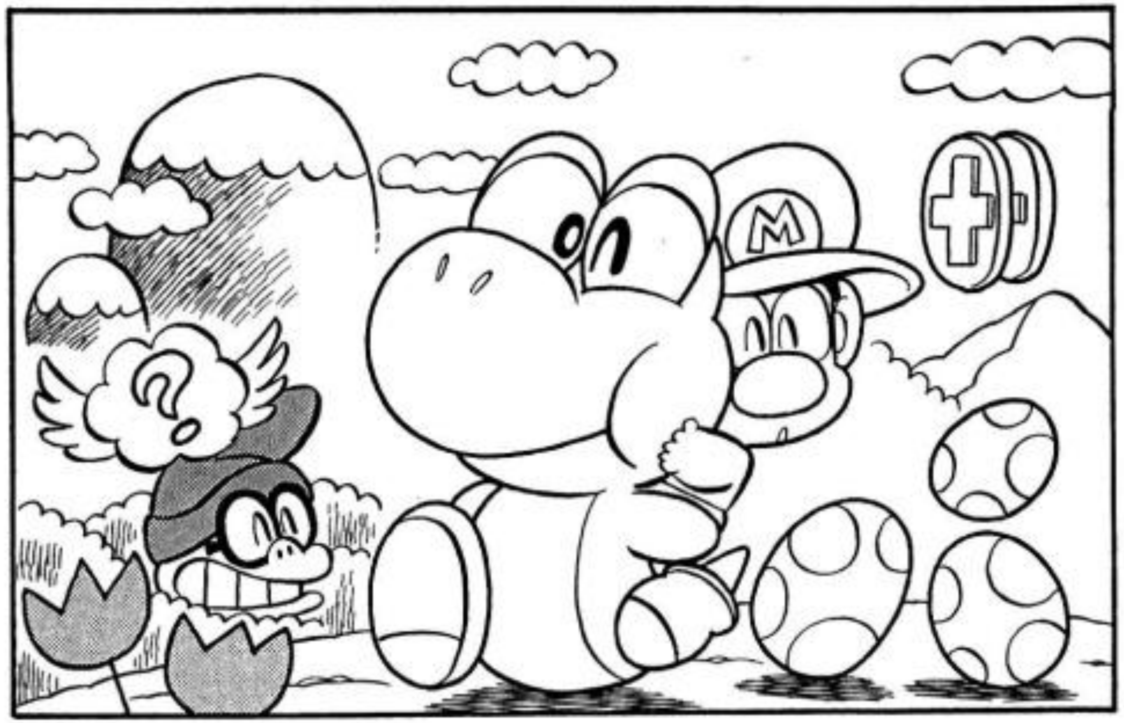
Describing distributions

- Central tendency
  - mode
  - mean
  - median
  - trimmed mean

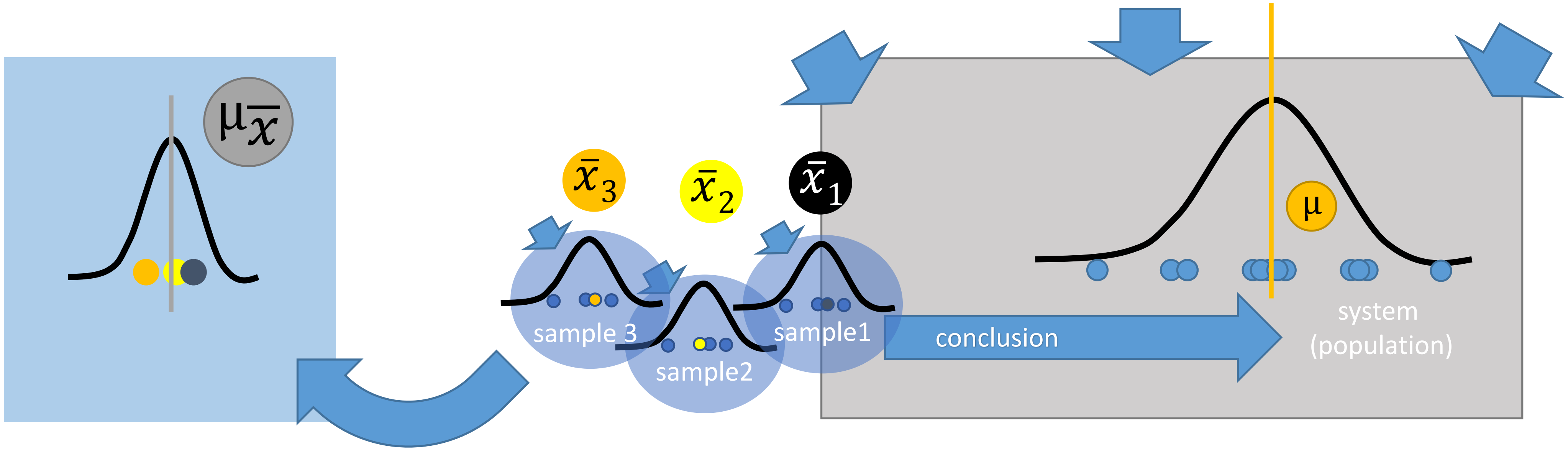
- Dispersion
  - range
  - mean (absolute) deviation
  - variance & standard deviation
  - coefficient of variation

Hypothesis testing: A framework for finding the differences



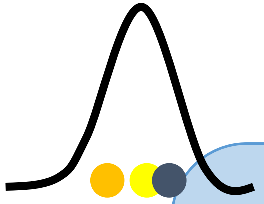


Sampling | sample | population distribution



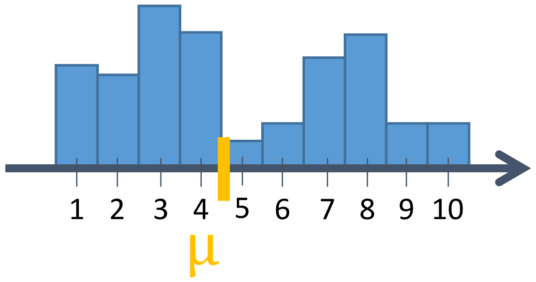
- The *sampling distribution* is the distribution of the estimated parameter values (here: expected value) of the population taken from the sample distribution

Central limit theorem



No matter how the population is distributed: the population of sample means will approximate a Gaussian distribution if the sample size is large enough

Central limit theorem ("simulation")



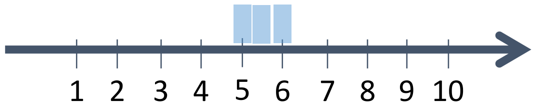
s_1=[3; 4; 7; 8] $\bar{x}_1 = 5.5$

s_2=[1; 5; 8; 10] $\bar{x}_2 = 6.0$

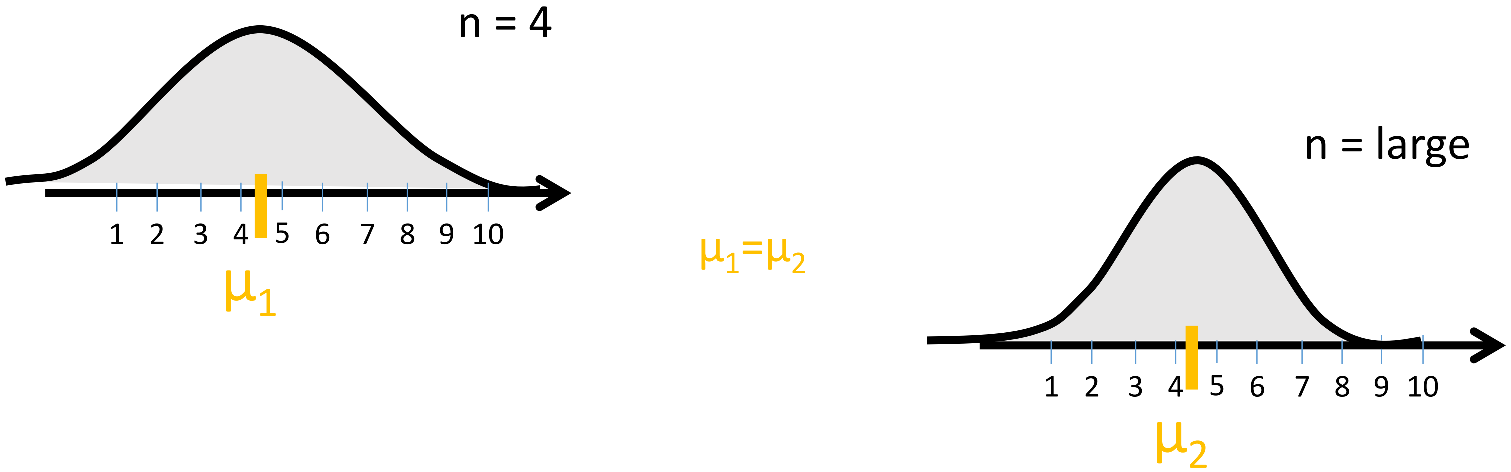
s_3=[2; 3; 6; 9] $\bar{x}_3 = 5.0$

…

s_n=[…   ]



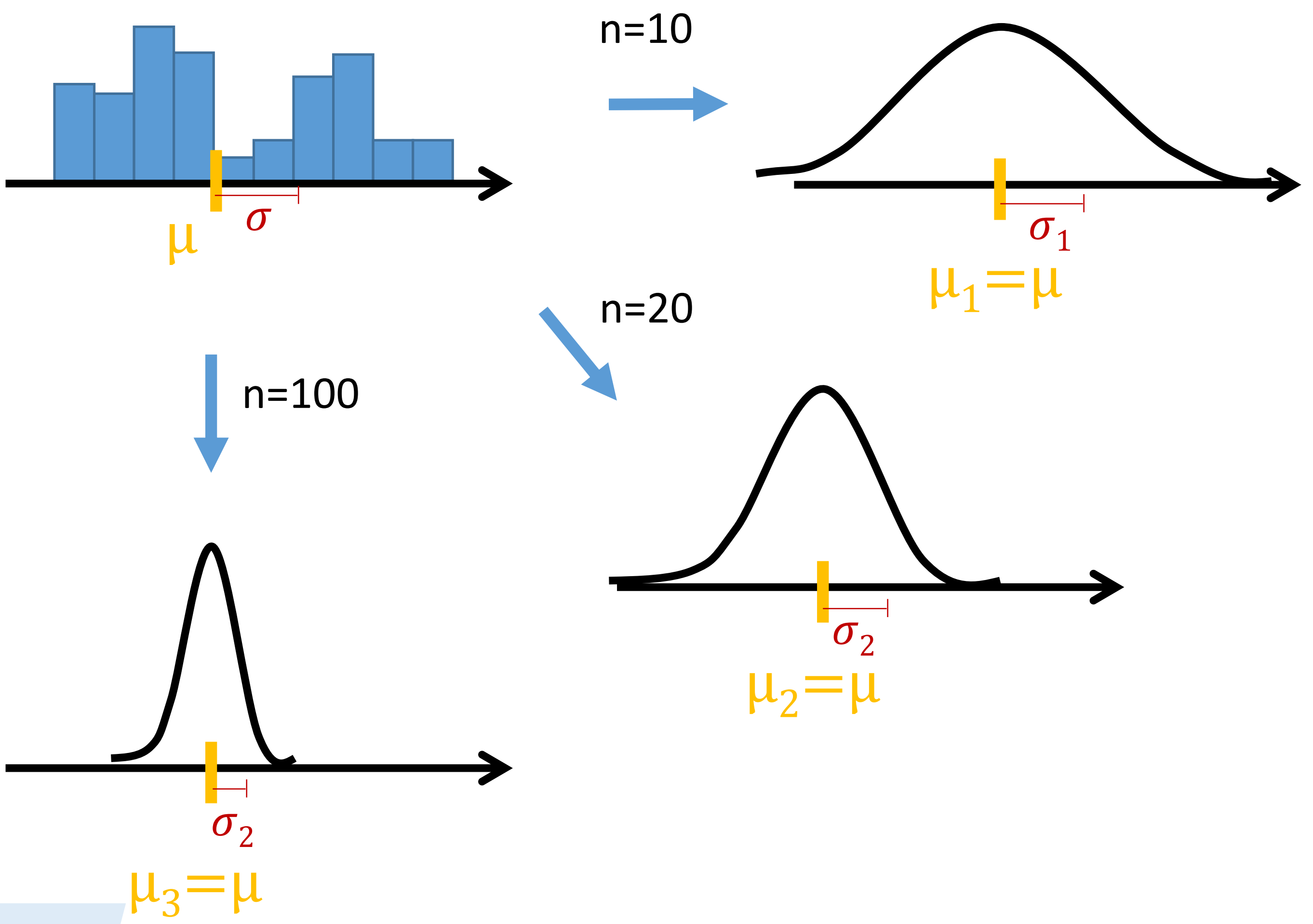
Central limit theorem ("simulation")



- Sample size ---> ∞
- Sampling distribution ---> normal distribution

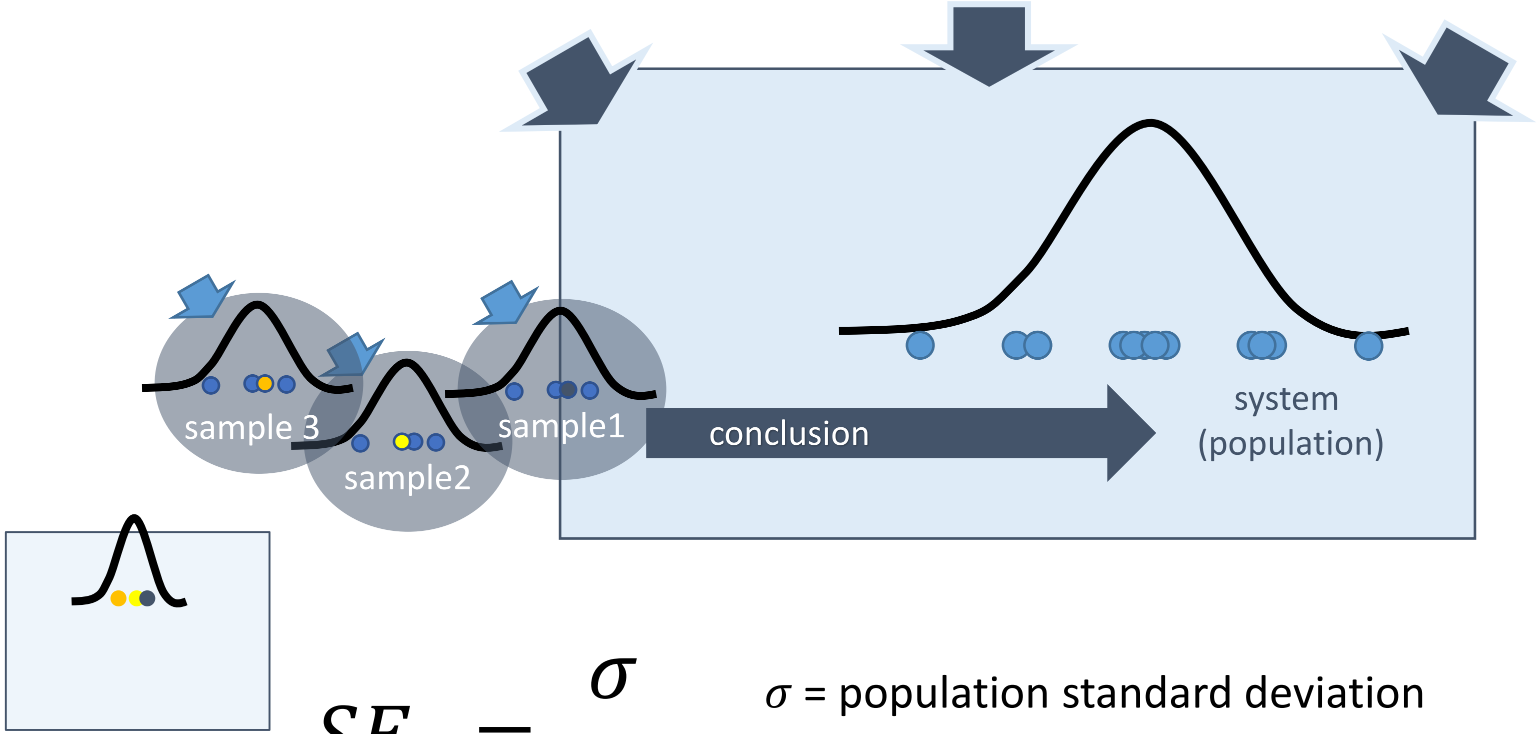
Standard error of the mean

aka: the standard deviation of the sampling distribution of the sample means



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Remark: Standard error of the mean



$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

σ = population standard deviation

n = sample size

- It defines the standard deviation of different samples means taken from the same population

Hypothesis testing

- Question: Is the effect I observe true/real or occurred by chance?

- Proof by contradiction:

  To prove A, you temporarily assume that A is false. If the assumption leads to a contradiction, you conclude that A must actually be true.

Establish two hypothesis

- Null hypothesis (H_0)

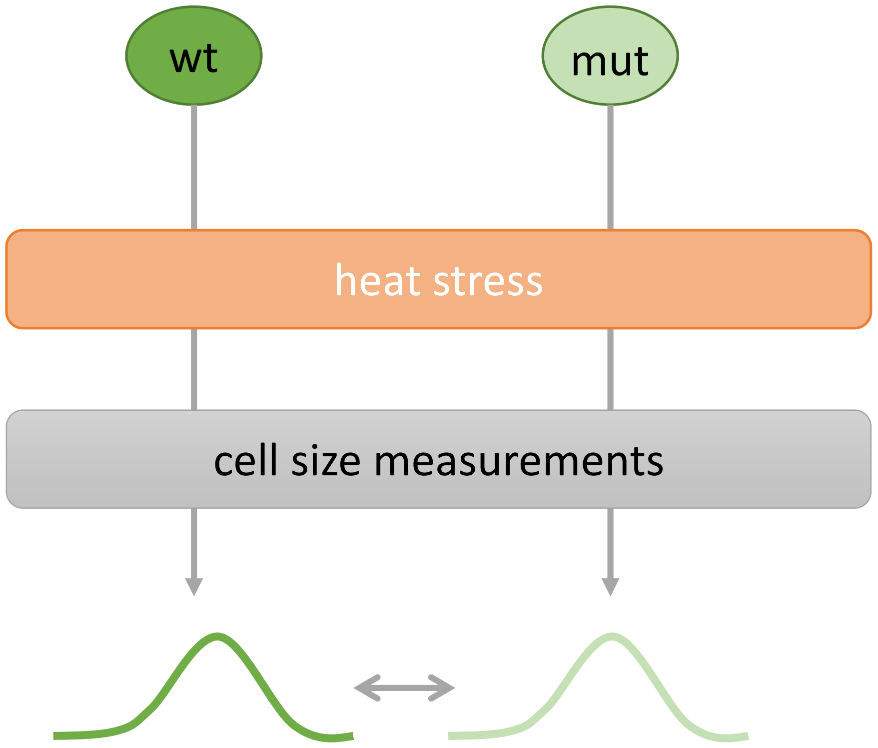
no effect

$\mu 1 = \mu 2$

- Alternative hypothesis (H_1)

effect

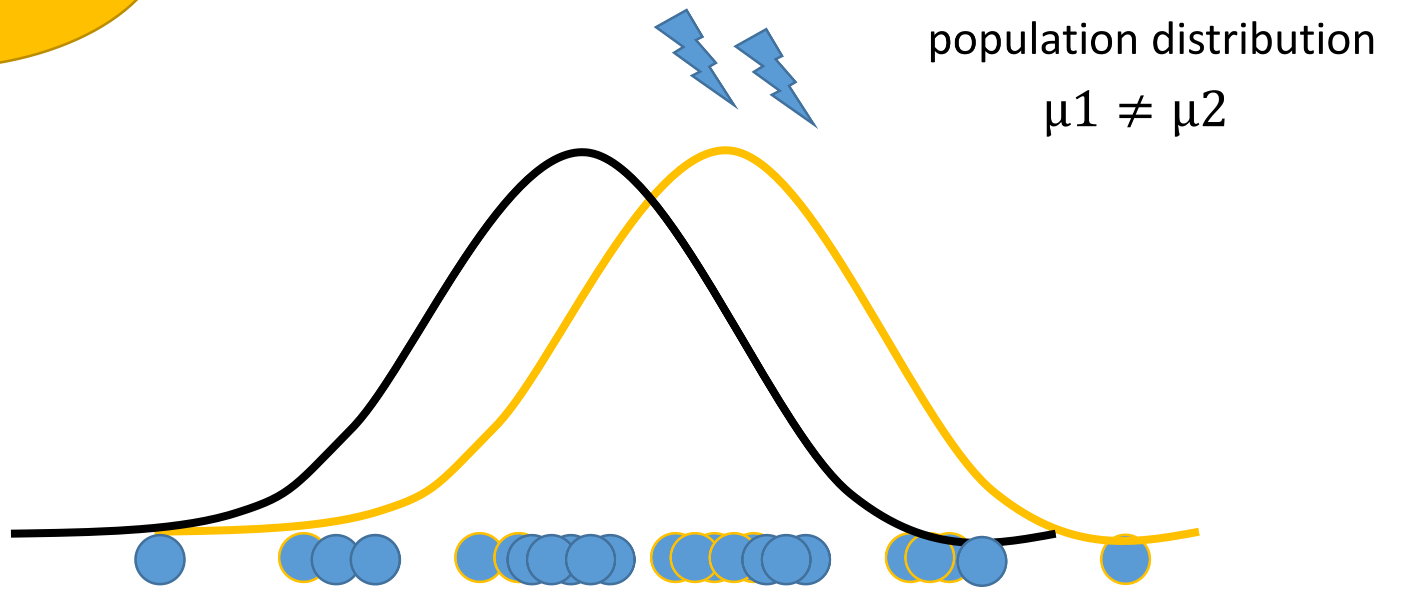
$\mu 1 \neq \mu 2$



Is the effect I observe true ?

H_1= true

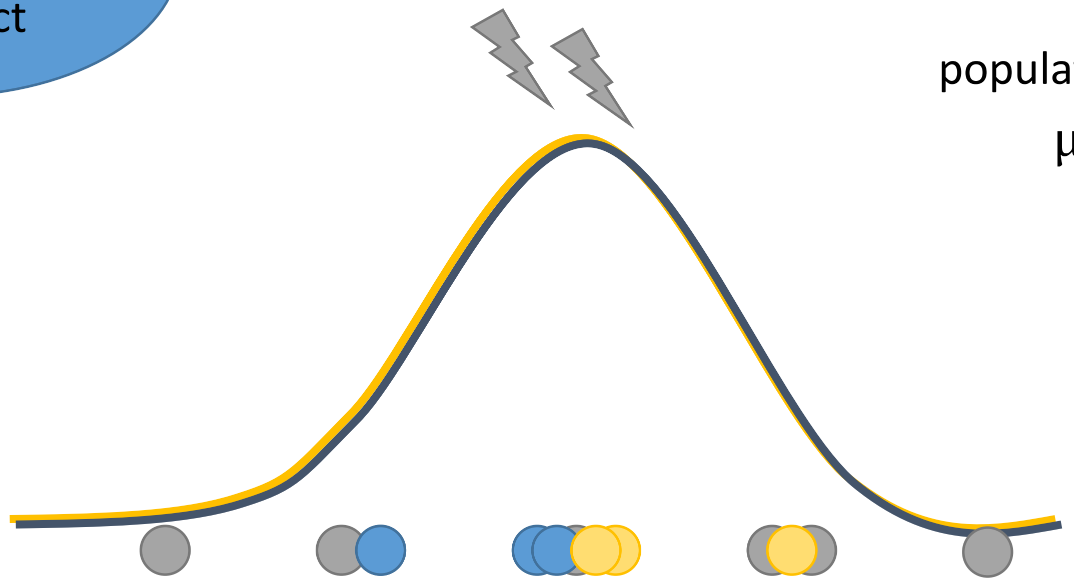
effect



- Alternative hypothesis states that the populations are different

Is the effect I observe true ?

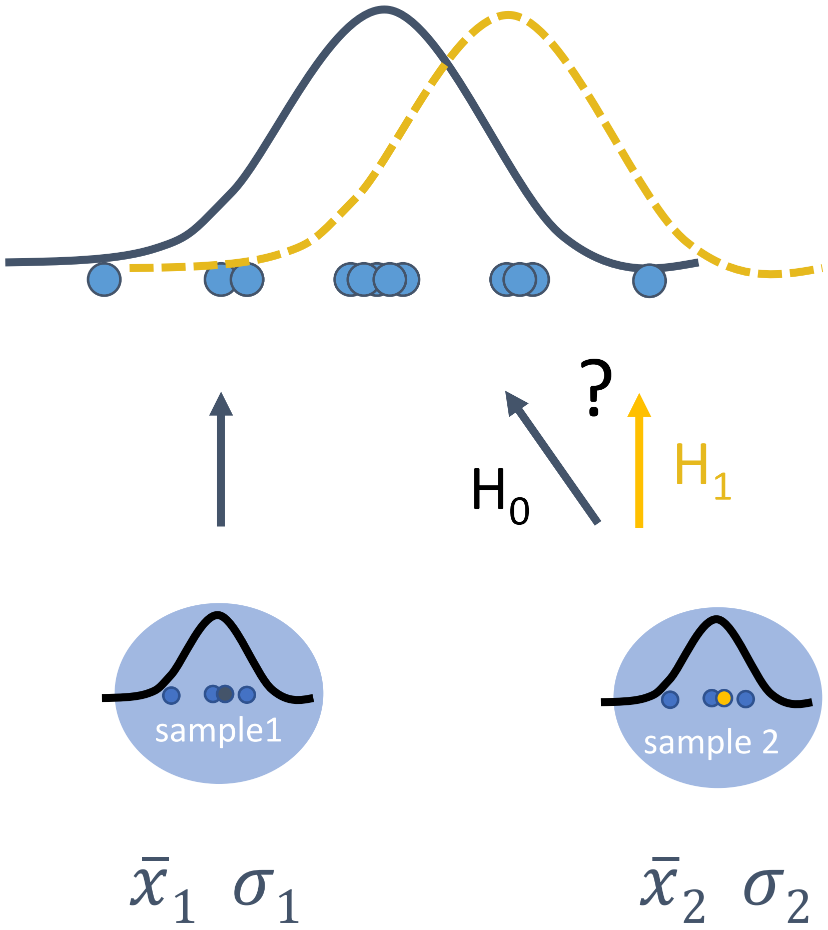
H_0= true
no effect



population distribution
$\mu1 = \mu2$

- Null hypothesis states that the populations are equal

Is the effect I observe true ?

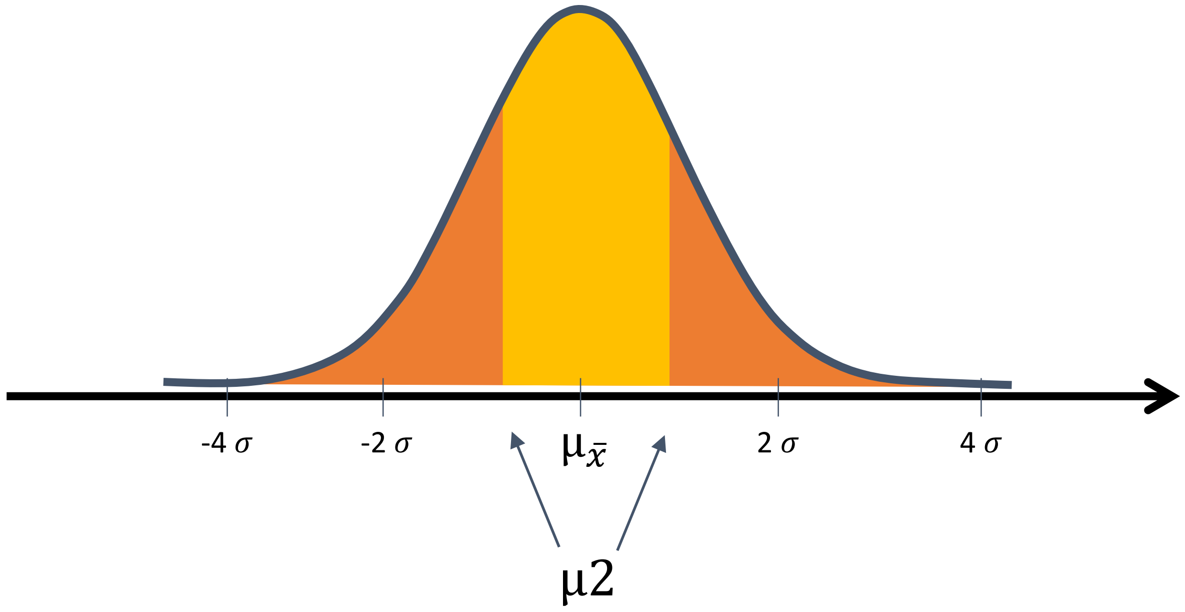


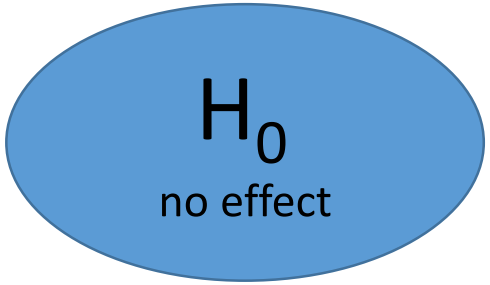
unknown

measured

calculated

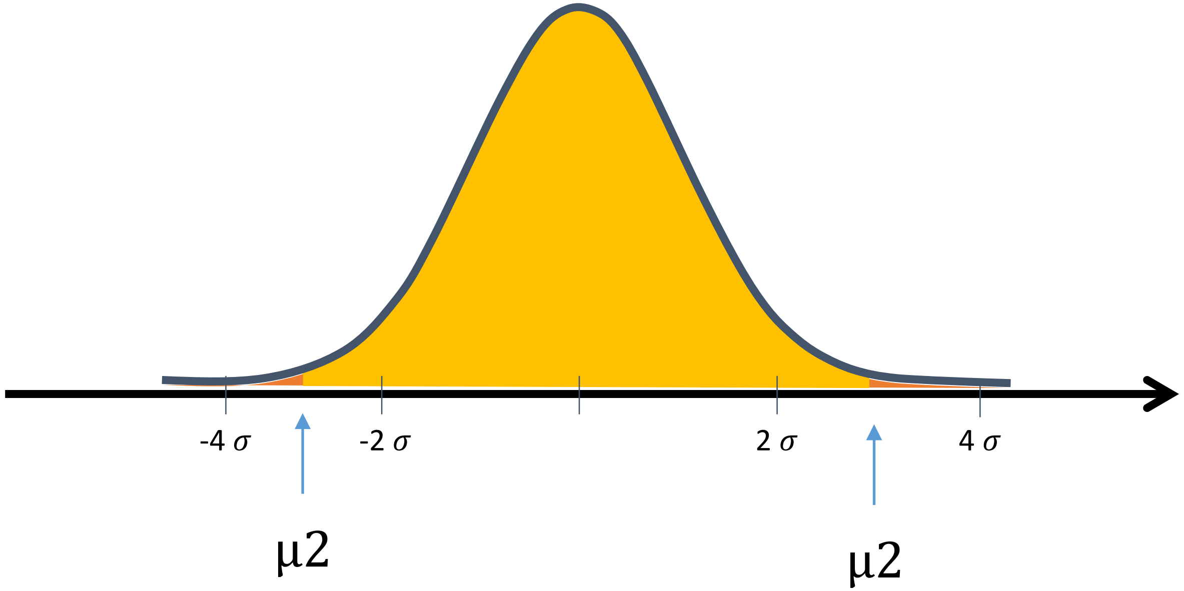
What is the probability of obtaining a value at least as extreme as the one that was observed ?

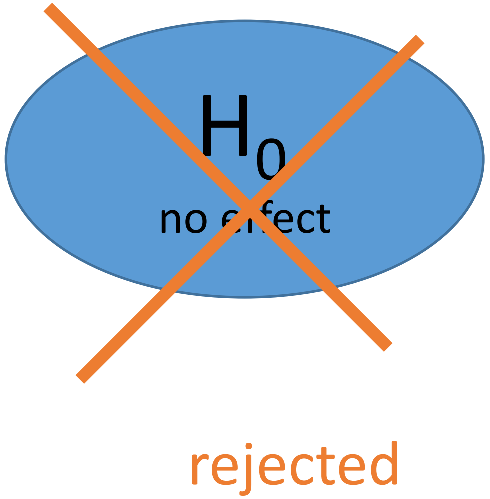




- The difference between μ1 and μ2 was most probably by chance: We take H_0 as true -> no effect

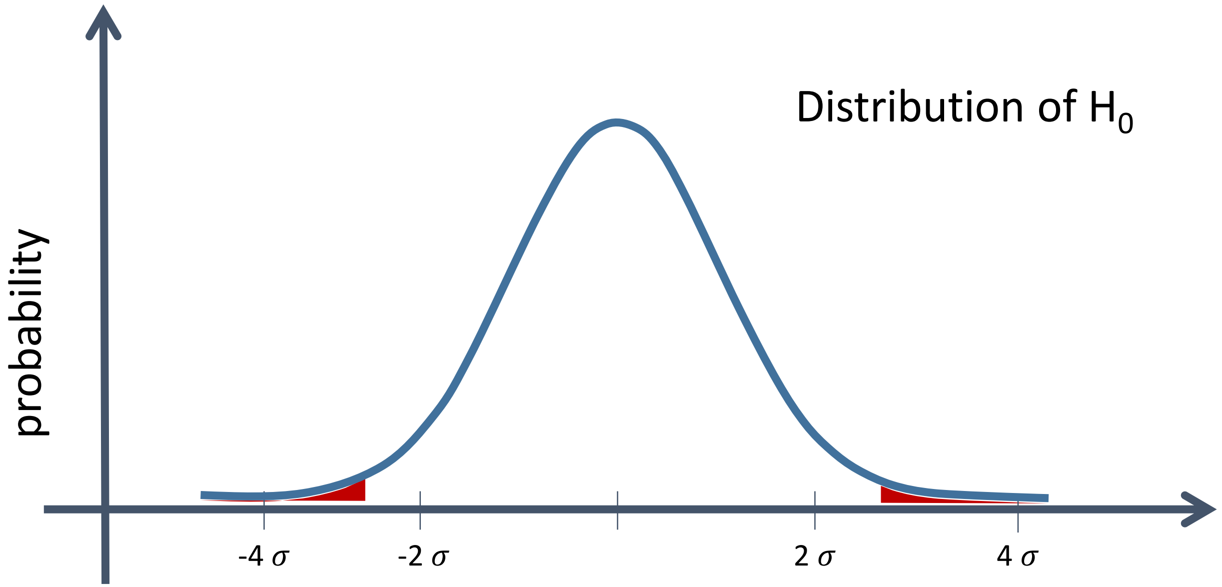
What is the probability of obtaining a value at least as extreme as the one that was observed ?





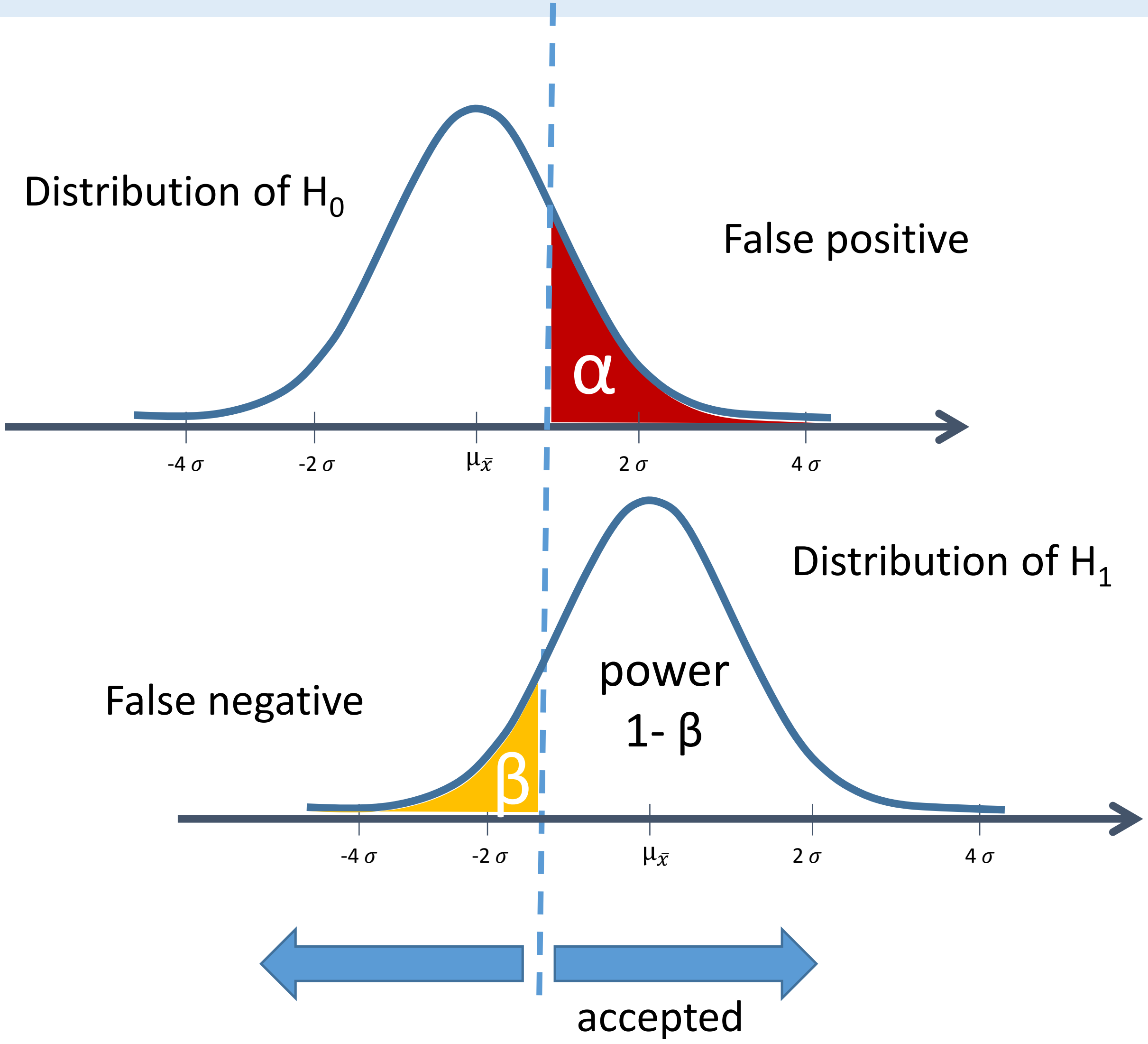
- Proof by contradiction:
  If we can reject H_0 than we assume H_1 to be true

P-Value

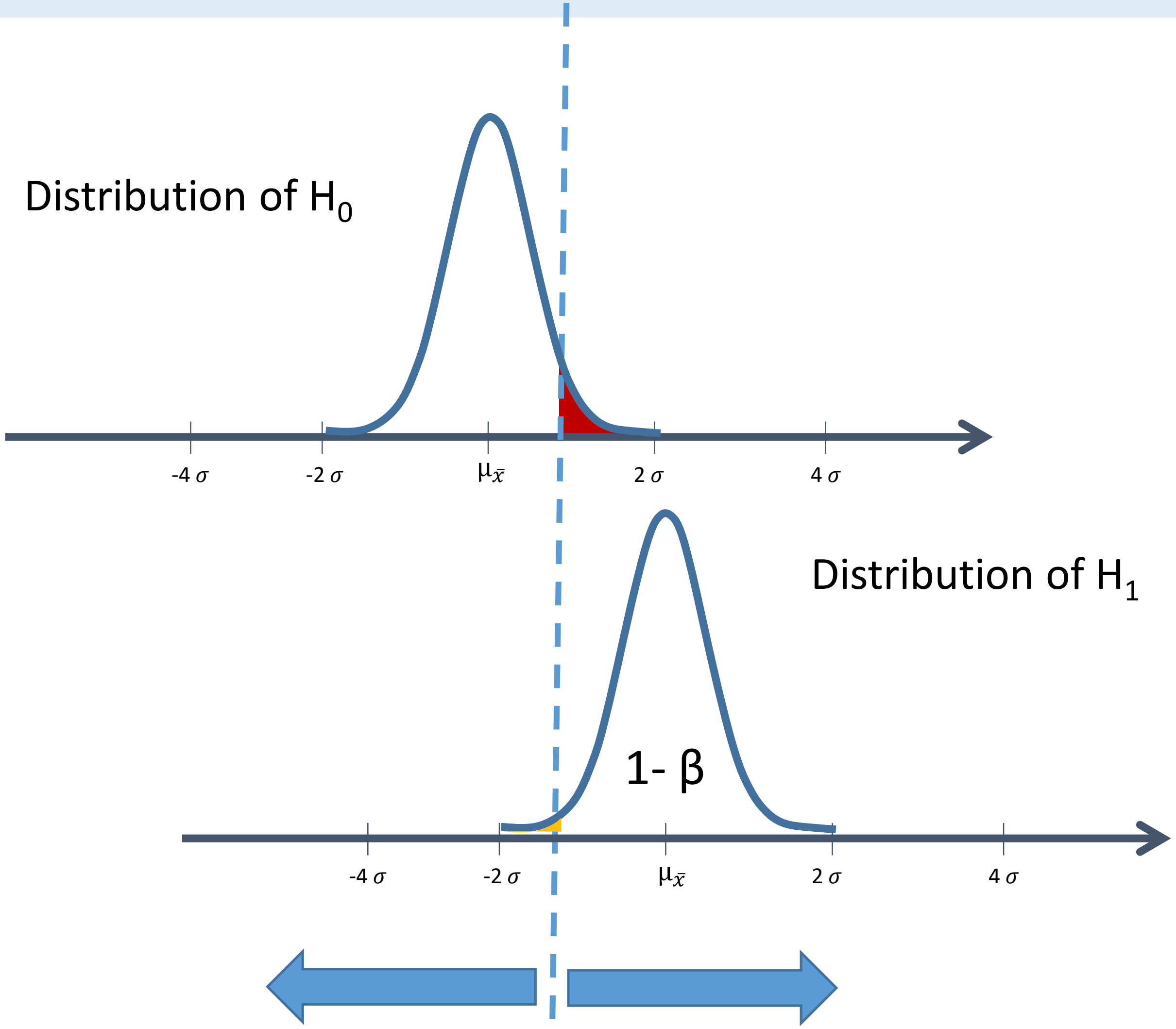


- A p-value is the probability of obtaining a value at least as extreme as the one that was observed

Power of a Test



Increase sample size

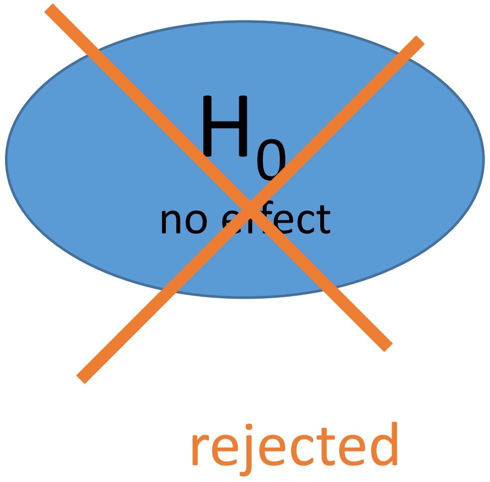


Computational
Systems Biology

Significance criterion (when to reject H_0)

- The most common approach to hypothesis testing is to choose a threshold α for the p-value and to accept as significant any effect with a p-value ≤ α

P-value	**Interpretation**
$P < 0.01$	very strong evidence against H_0
$0.01 \leq P < 0.05$	moderate evidence against H_0
$0.05 \leq P < 0.10$	suggestive evidence against H_0
$0.10 \leq P$	little or no real evidences against H_0

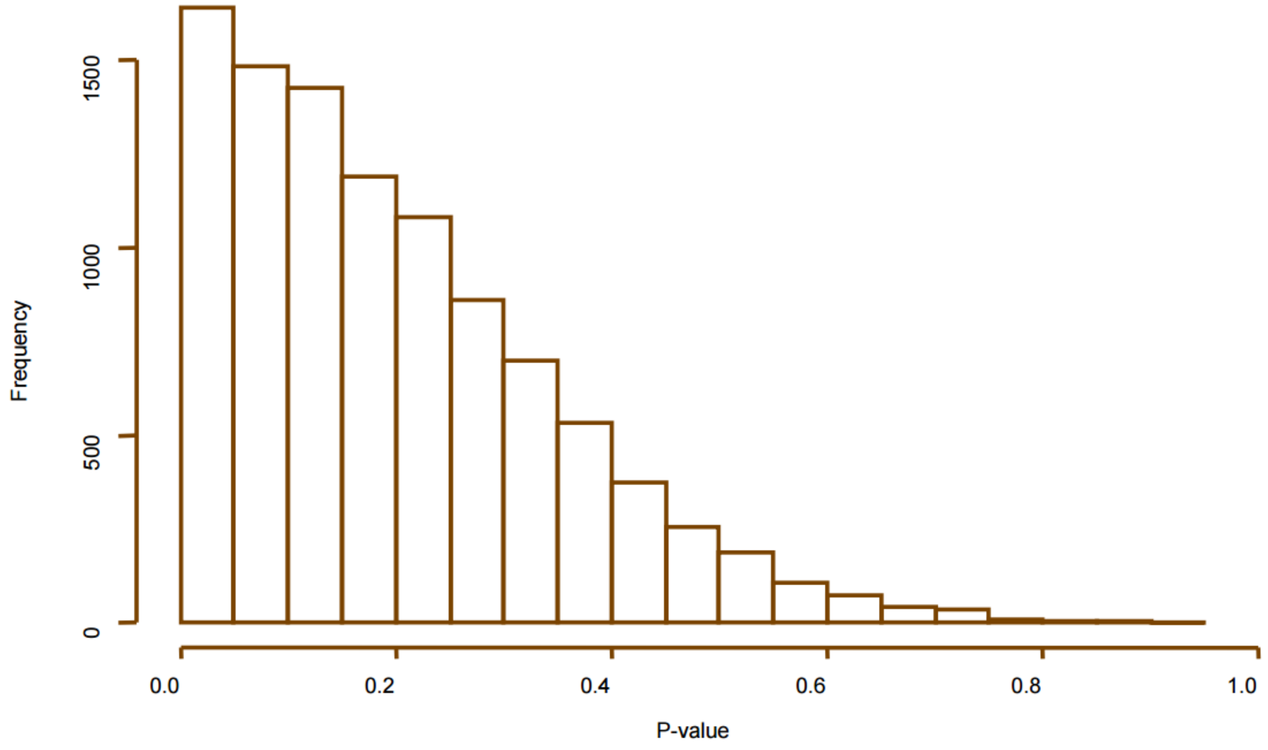


Multiple testing remarks

- The hypothesis test framework was built to perform one test only.
- What about testing multiple times?
- What does that mean for the p-value?

Estimating the proportion of truly Null Tests

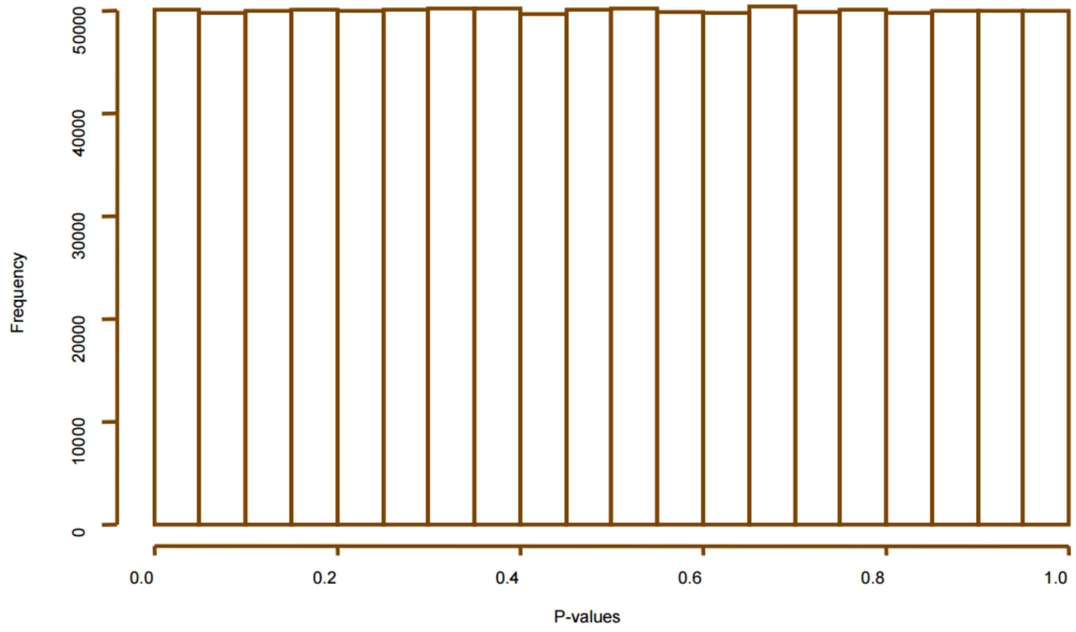
- Under the alternative hypothesis p-values are skewed towards 0



Computational
Systems Biology

Estimating the proportion of truly Null Tests

- Under the null hypothesis p-values are expected to be uniformly distributed between 0 and 1



Adaptation to multiple testing

- Family wise error rates:

$$P(\#false\ positives \geq 1)$$

- False discovery rate:

$$E\left[\frac{\#false\ positives}{\#\ total\ discoveries}\right]$$

Example:

Given: 550 out of 10 000 genes are significant at 0.05 level

- P-value < 0.05
  Expect 0.05 * 10 000 = 500 false positives

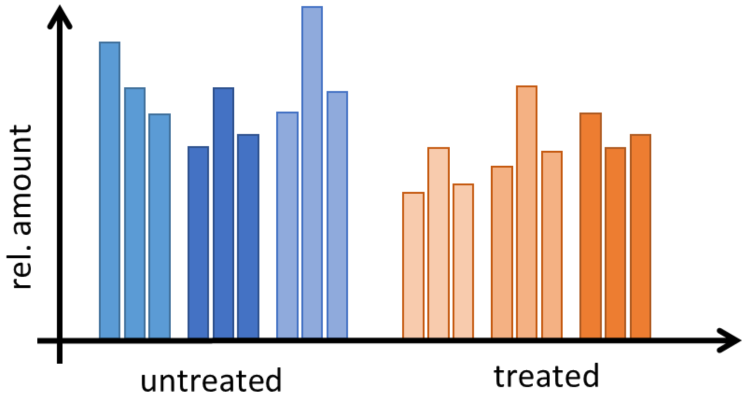
- False discovery rate < 0.05
  Expect 0.05 * 550 = 27.5 false positives

- Family wise error rate < 0.05
  The probability of at least 1 false positive ≤ 0.05

Be aware…

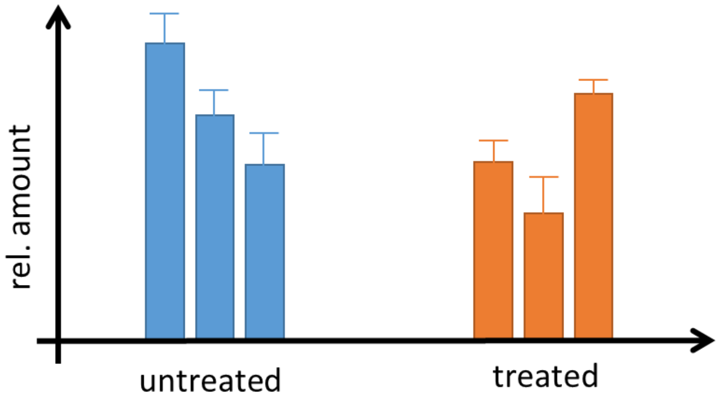
- Statistical significance can mean totally different thing depending on how it is used!

Aggregation and error propagation



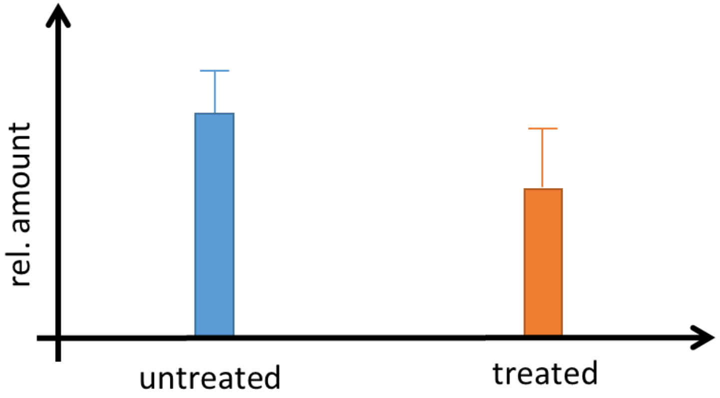
mean

stDev$_{n-1}$

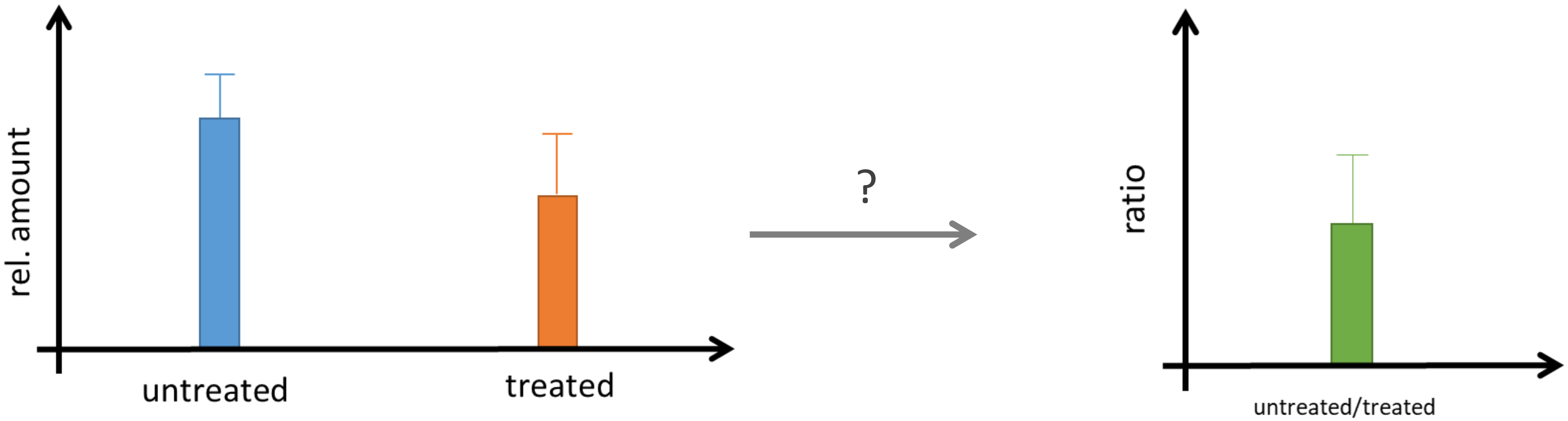


$$X_c = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

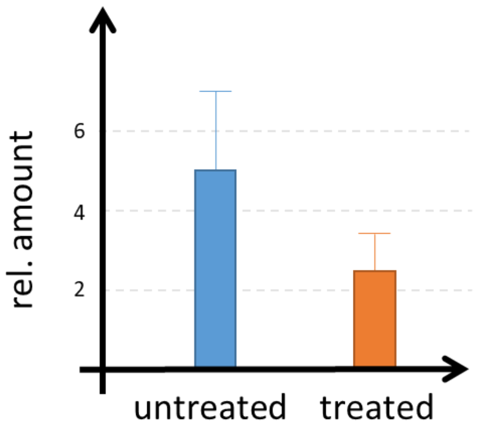
$$S_c{}^2 = \frac{n_1 \left[S_1{}^2 + \left(\overline{X}_1 - \overline{X}_c \right)^2 \right] + n_2 \left[S_2{}^2 + \left(\overline{X}_2 - \overline{X}_c \right)^2 \right]}{n_1 + n_2}$$

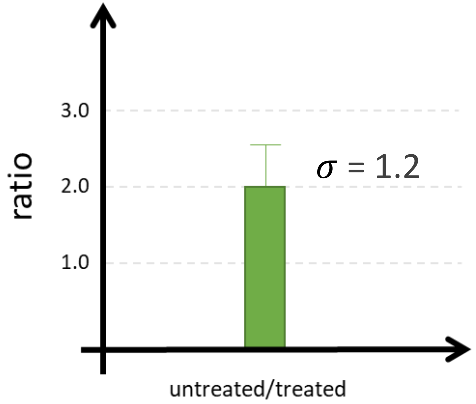


Aggregation and error propagation



Aggregation and error propagation





ratio

$$x_1 = 5.0 \qquad \delta x_1 = 2.0$$

$$x_2 = 2.5 \qquad \delta x_2 = 1.0$$

$$f_{(x1,x2)} = \frac{x_1}{x_2} = 2.0$$

$$\frac{\partial f}{\partial x_1} = \frac{1}{x_2}$$

$$\frac{\partial f}{\partial x_2} = \frac{x_1}{{x_2}^2}$$

error propagation

$$\sigma = \sqrt{\sum_{j=1}^{m} \left(\frac{\partial f}{\partial x_j}\right)^2 \cdot \sigma_{x_j}^2}$$

$$\sigma = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 \cdot \sigma_{x_1}^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \cdot \sigma_{x_2}^2}$$

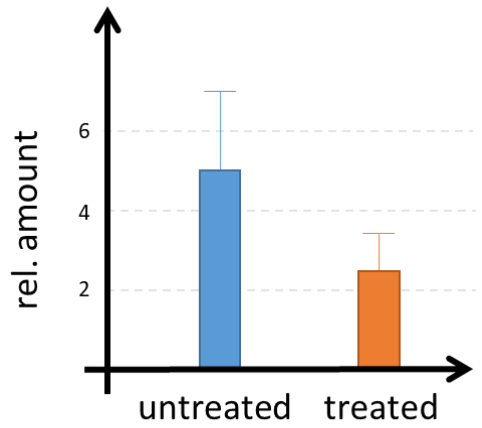
$$\sigma = \sqrt{\left(\frac{1}{x_2}\right)^2 \cdot \delta {x_1}^2 + \left(\frac{x_1}{x_2^2}\right)^2 \cdot \delta {x_2}^2}$$

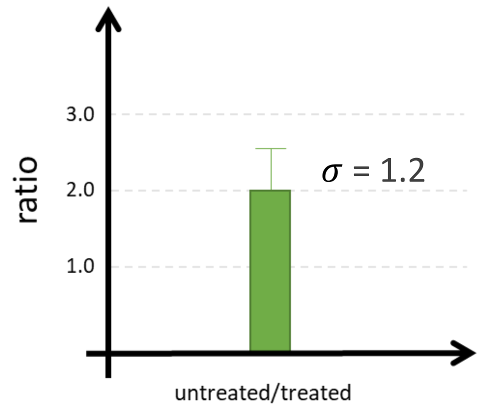
$$\sigma = \sqrt{\left(\frac{1}{2.5}\right)^2 \cdot 2^2 + \left(\frac{5}{6.25}\right)^2 \cdot 1^2}$$

$$\sigma = \sqrt{1.28} = 1.1314 = 1.2$$

Computational
Systems Biology

Aggregation and error propagation





ratio

$$x_1 = 5.0 \qquad \delta x_1 = 2.0$$

$$x_2 = 2.5 \qquad \delta x_2 = 1.0$$

$$f_{(x1,x2)} = \frac{x_1}{x_2} = 2.0$$

$$\frac{\partial f}{\partial x_1} = \frac{1}{x_2}$$

$$\frac{\partial f}{\partial x_2} = \frac{x_1}{{x_2}^2}$$

error propagation

addition or subtraction

$$Q = x_1 + x_2 + \cdots$$

$$\delta Q = \sqrt{(\delta x_1)^2 + (\delta x_2)^2 + \cdots}$$

multiplication or division

$$Q = \frac{x_1 \cdot x_3 \ldots}{x_2 \cdot x_4 \ldots}$$

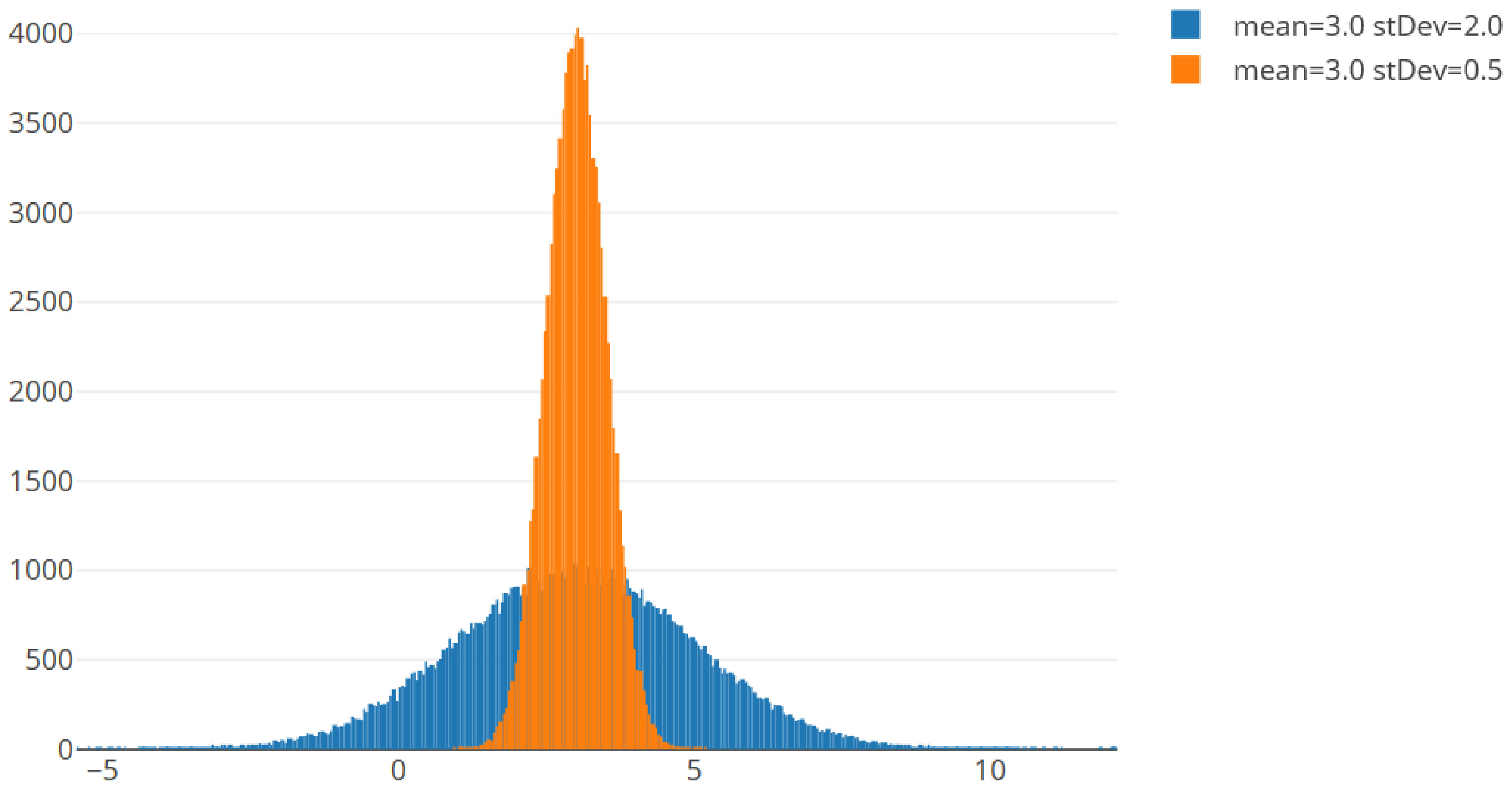
$$\frac{\delta Q}{|Q|} = \sqrt{\left(\frac{\delta x_1}{x_1}\right)^2 + \left(\frac{\delta x_2}{x_2}\right)^2 + \cdots}$$

$$\frac{\delta Q}{2} = \sqrt{\left(\frac{2}{5}\right)^2 + \left(\frac{1}{2.5}\right)^2}$$

$$\frac{\delta Q}{2} = 0.56569$$

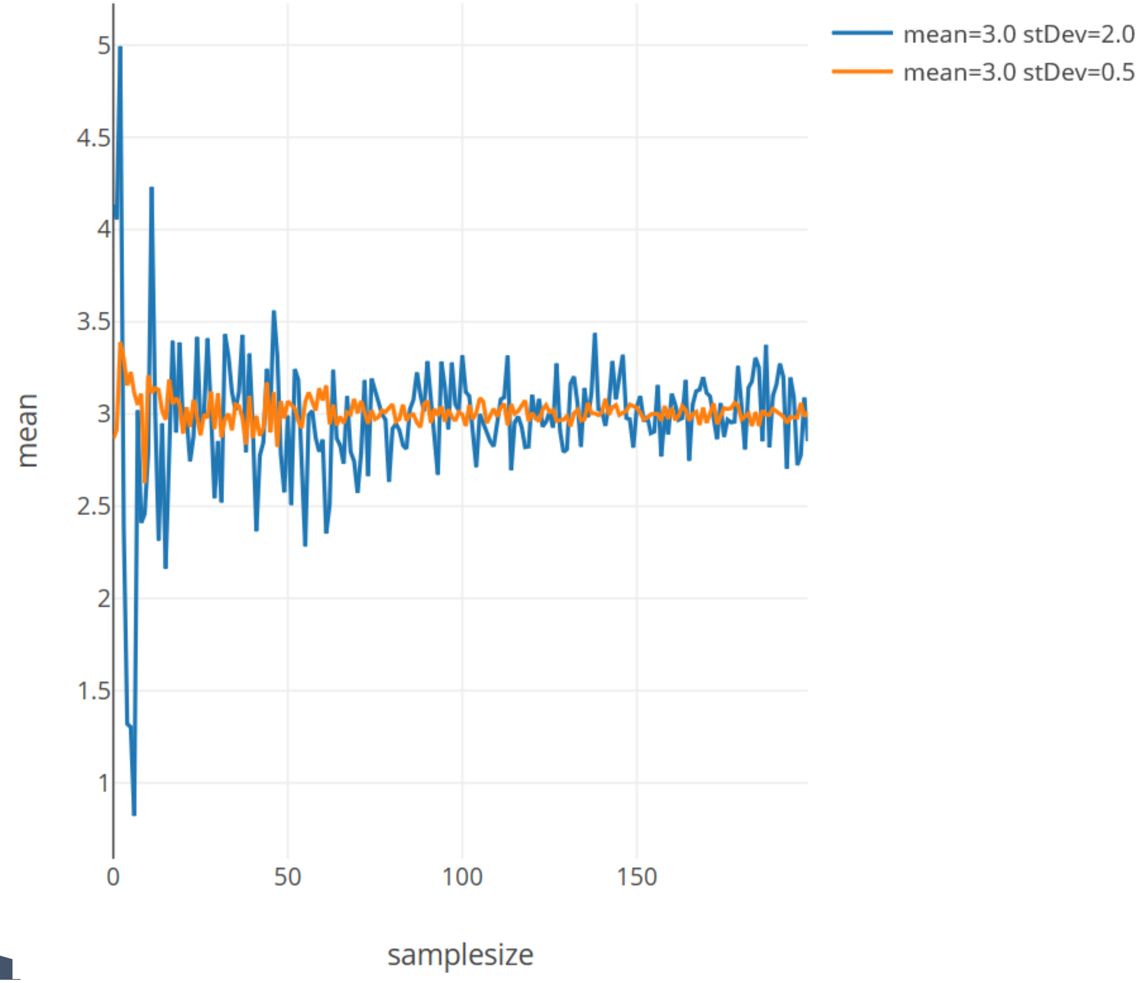
$$\delta Q = 1.1318 = 1.2$$

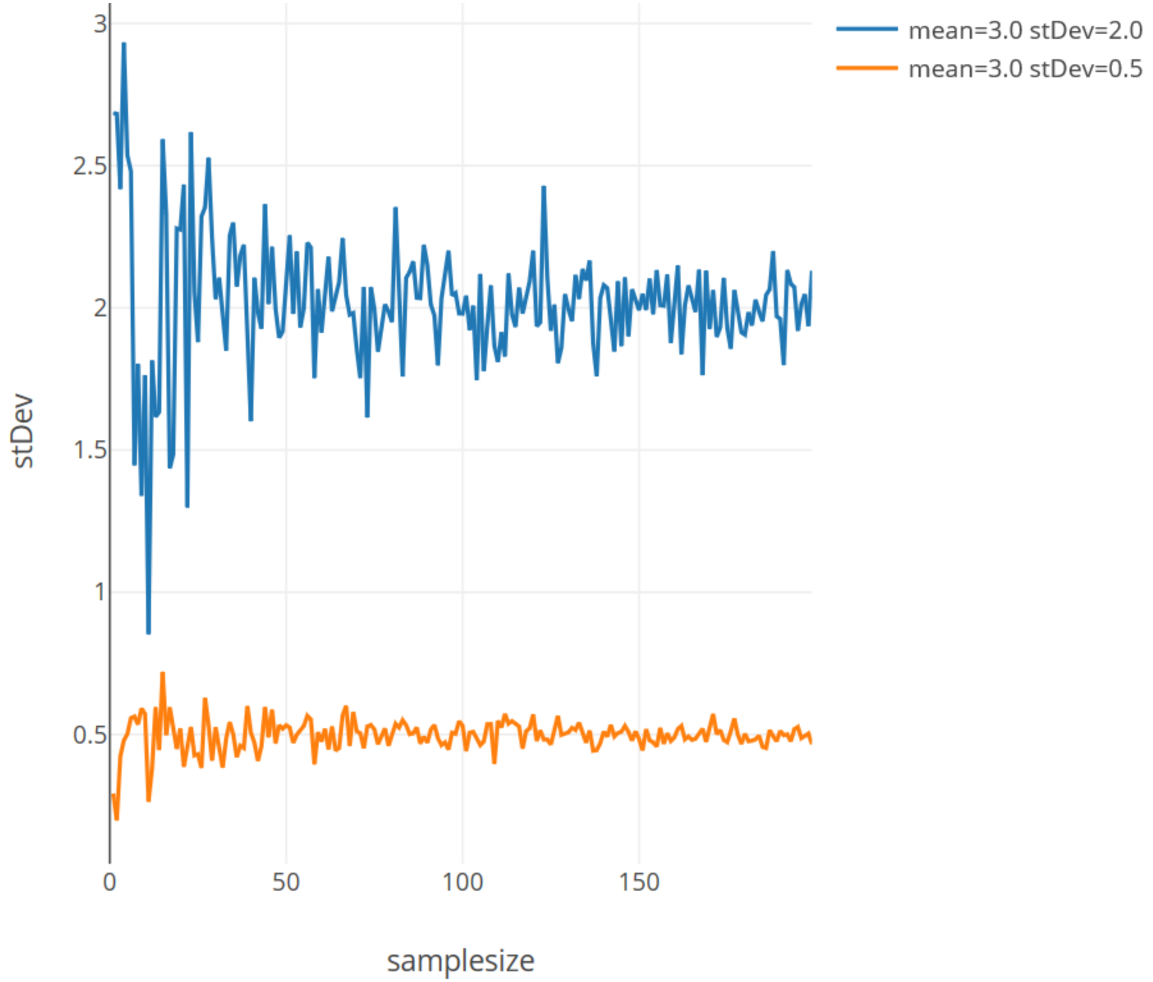
Normal distribution with different σ



N = 100,000

n vs. σ (sample size vs. stDev)

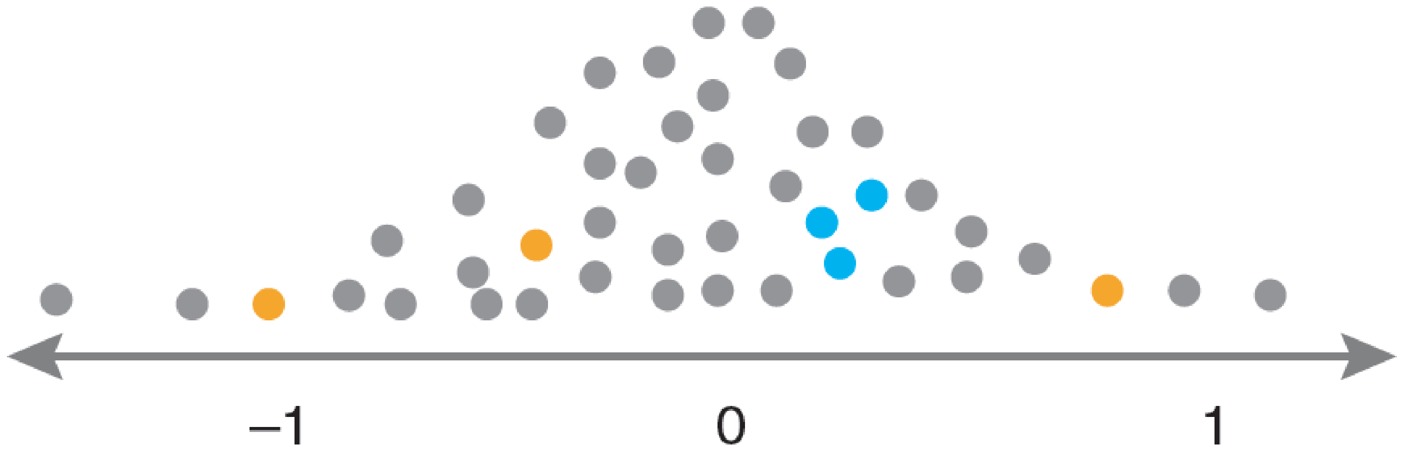




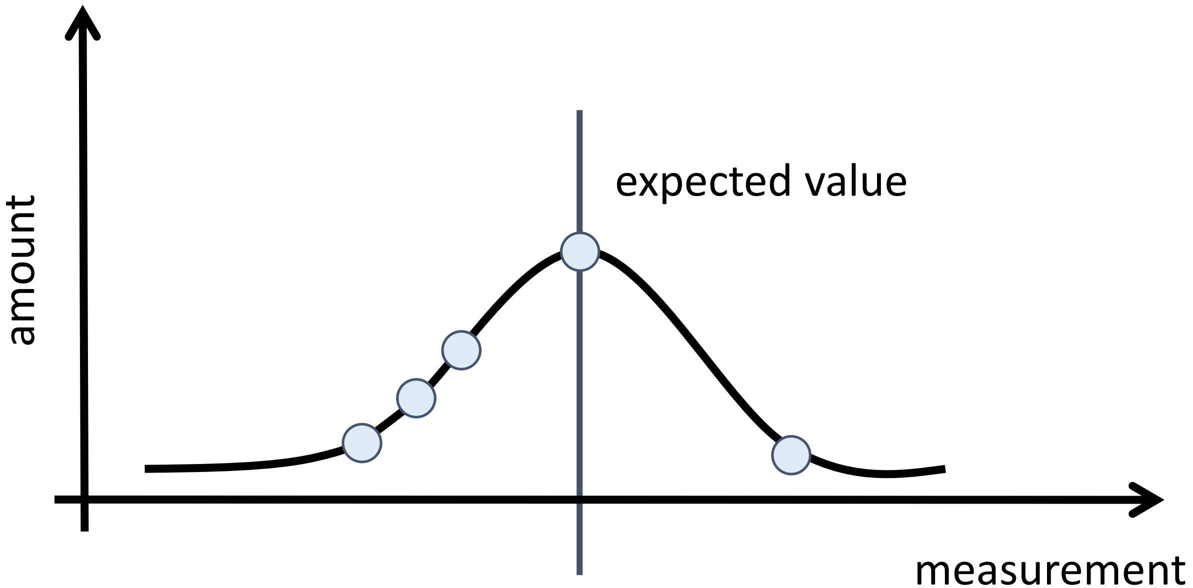
Computational
Systems Biology

Pitfall: Small sample sizes

- Small sample sizes (n < 10) can have a strong effect on the estimation of the central tendency and data dispersion of a population



Knowing the shape of a distribution?



- Assumption about the distribution shape of our measured values borrows additional information to describe the data

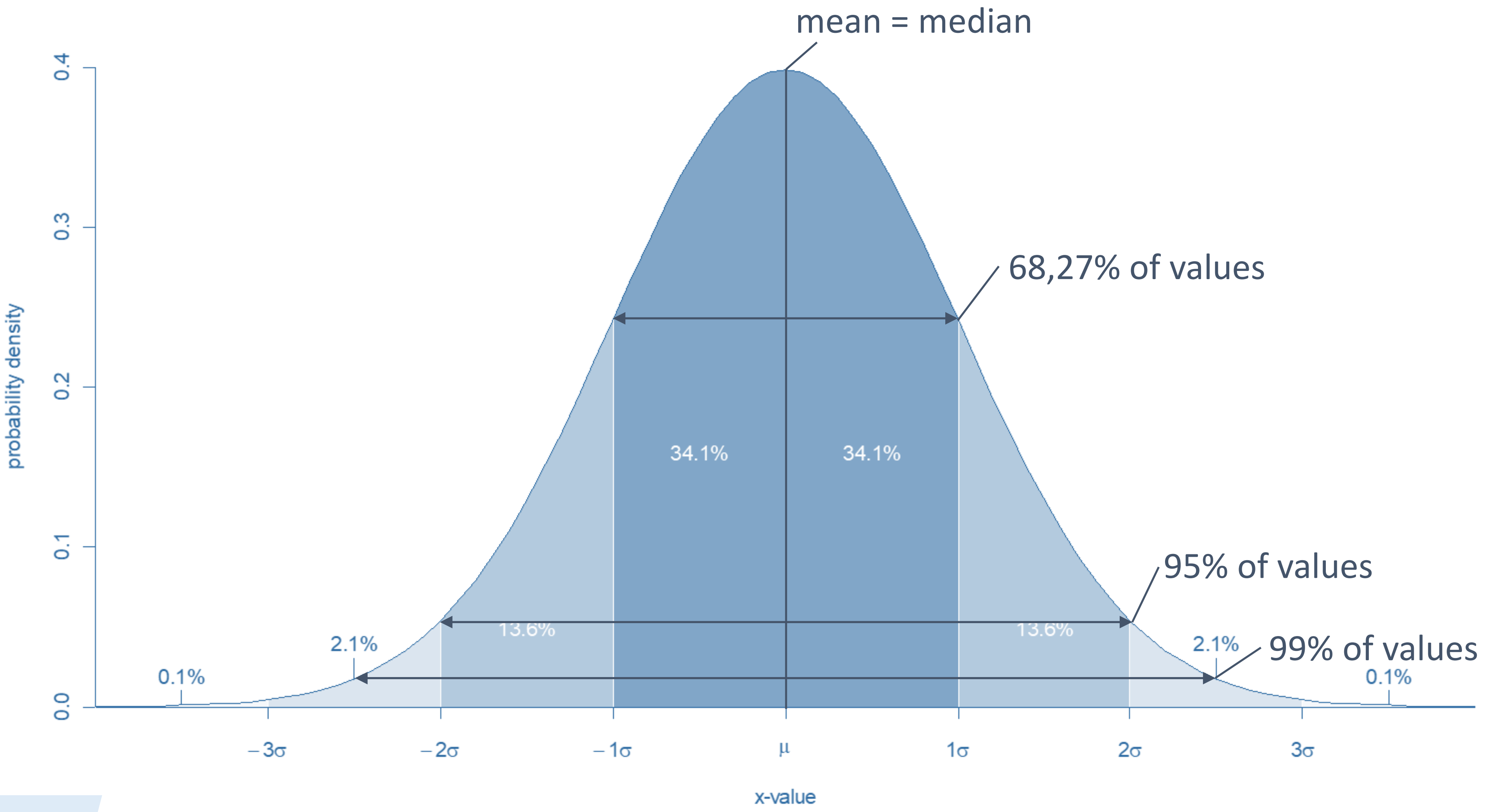
Central limit theorem

No matter how the population is distributed: the population of sample means will approximate a Gaussian distribution if the sample size is large enough

- “Large” depends on the real population distribution
  - Less normal population distribution => more sample (N >= 100)
  - More normal population distribution => N >= 10)

Computational Systems Biology

The Gaussian „Normal Distribution“



Symmetric around the mean