



Statistics: review of basic concepts



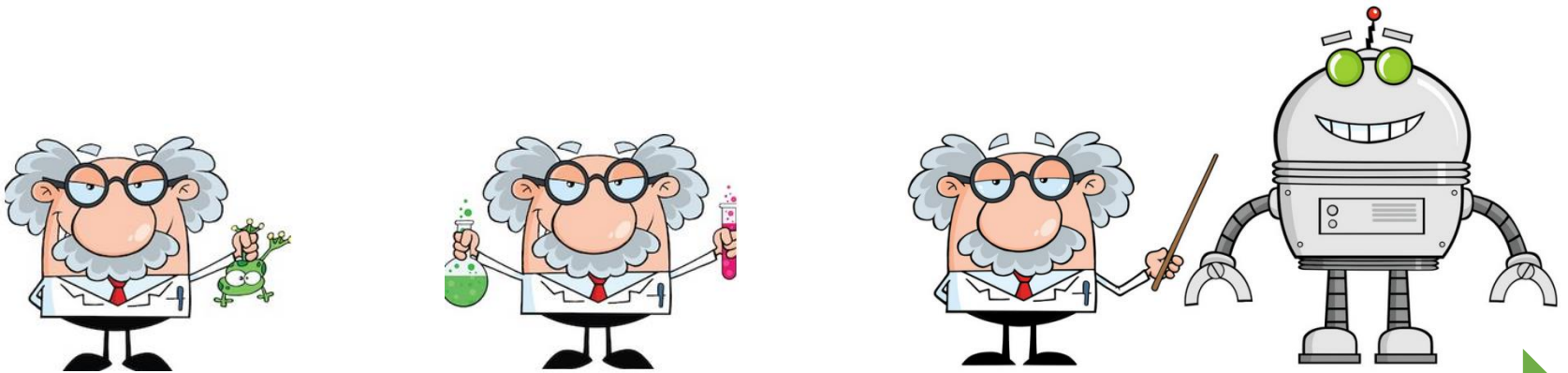
Francesca M. Buffa

Computational Biology &
Integrative Genomics

University of Oxford

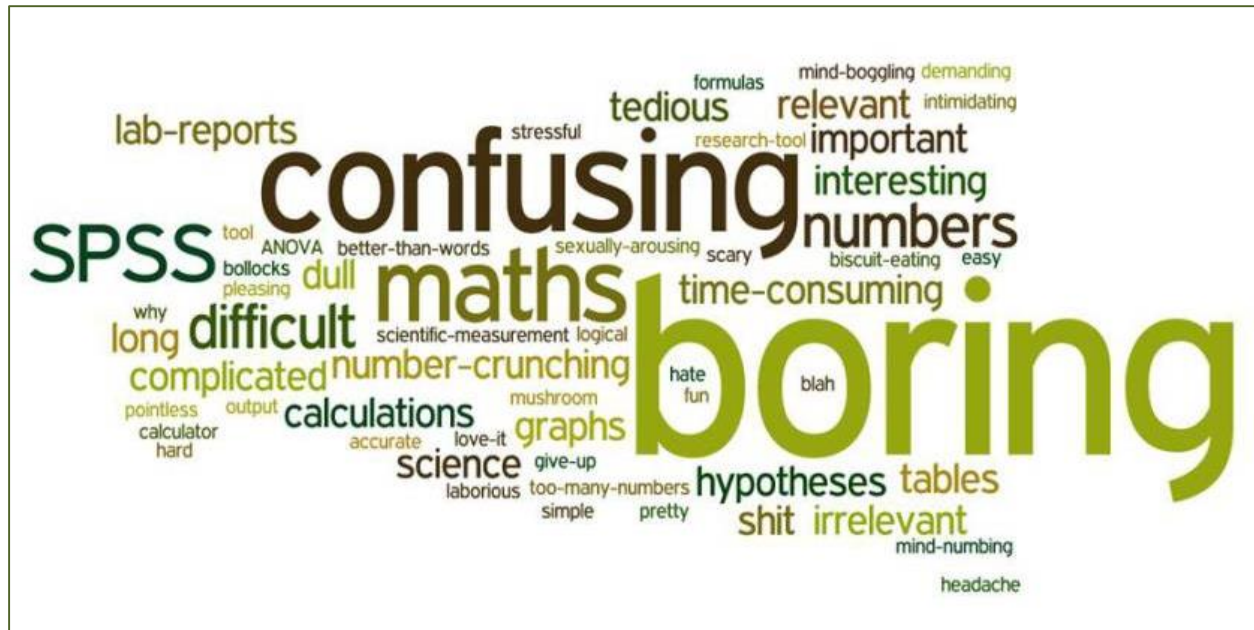
From descriptive to quantitative biology

*Biology is developing rapidly
from a mostly descriptive science
to a quantitative and predictive science,
like chemistry and physics or engineering.*



The biologist tool set is changing!

What is statistics about?



What is statistics about?

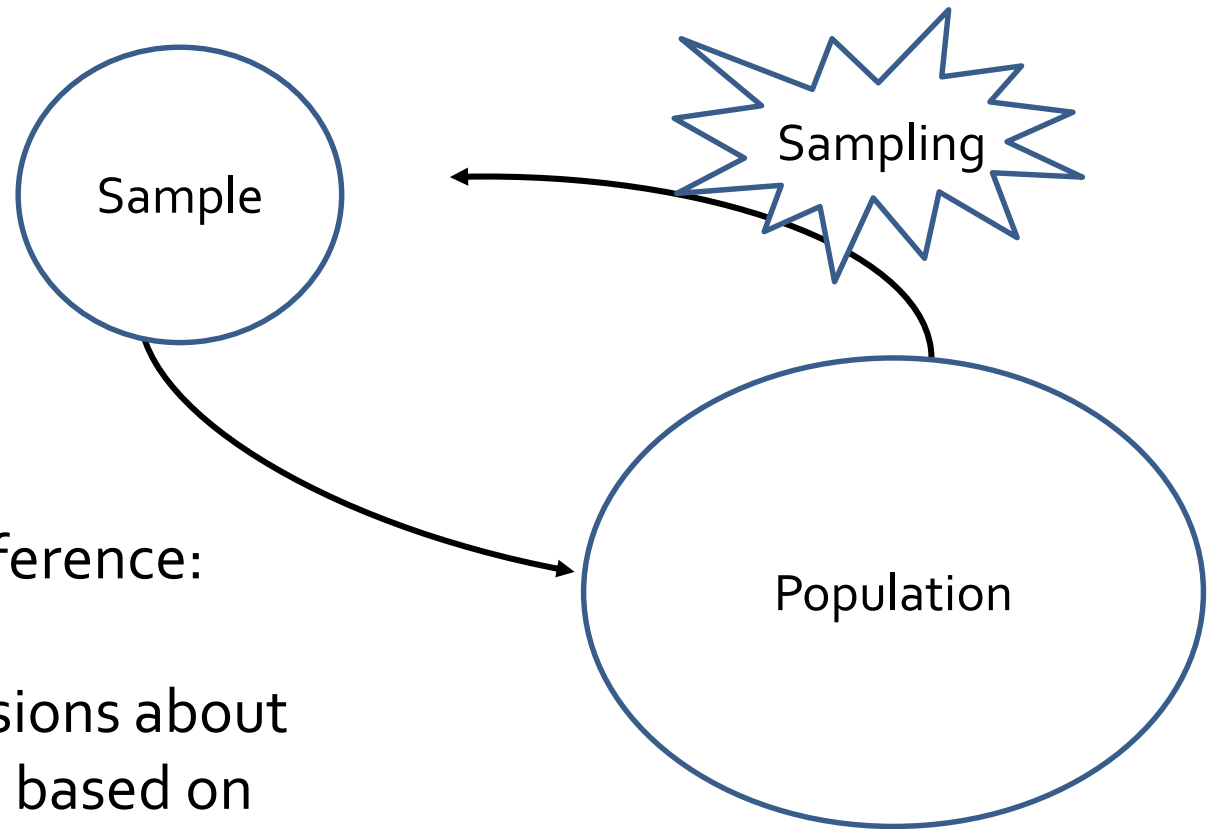
Given a population,
How do we measure and describe its attributes?
How do we test hypotheses about the population?



Spot the difference



Sampling

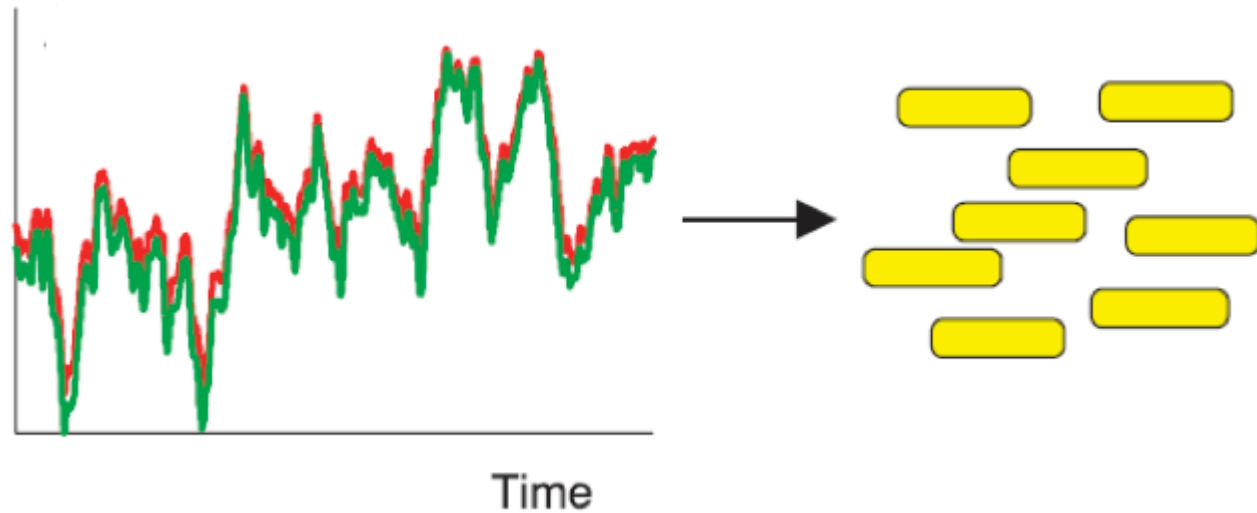


Statistical Inference:

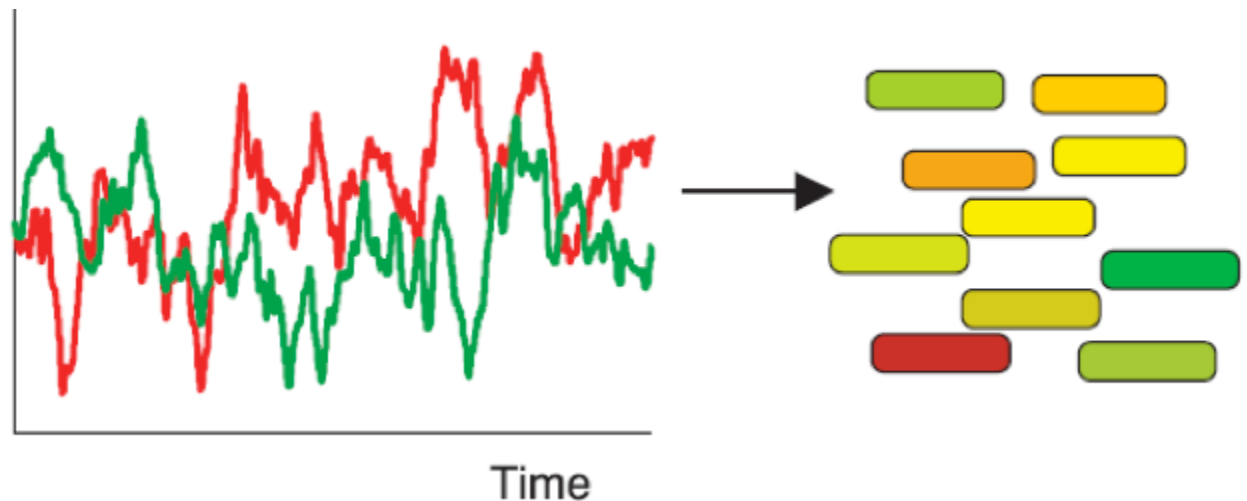
Drawing conclusions about
the population based on
the observed sample.

Transcriptional data are extremely variable

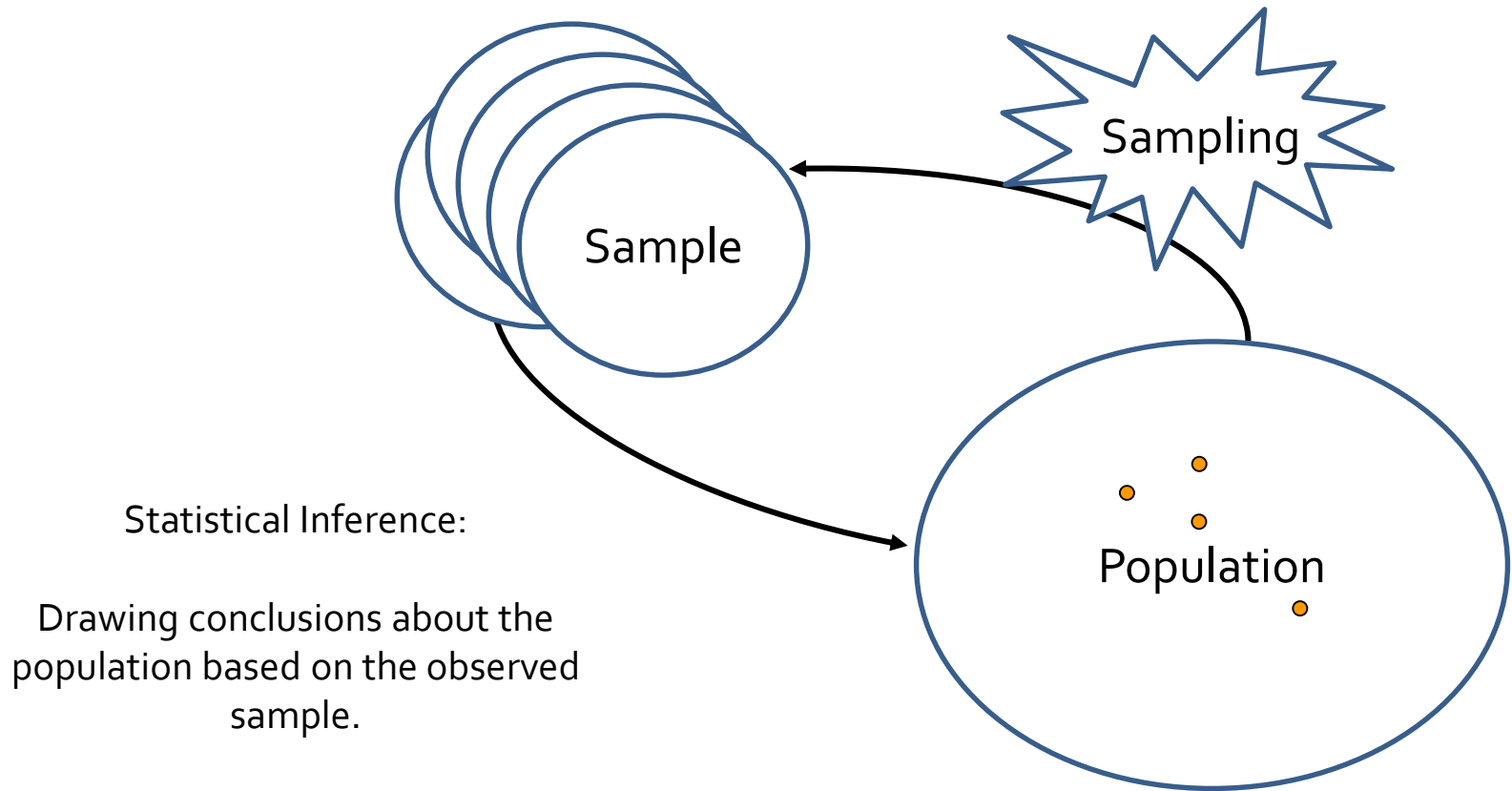
Extrinsic:
Fluctuations in states of
other cellular
components lead to
variation in expression of
a particular gene



Intrinsic: inherent
stochasticity of
biochemical processes
such as transcription (e.g.
transcriptional bursting)



Sampling and statistical inference



Resampling and error

VOLUME VI

MARCH, 1908

No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

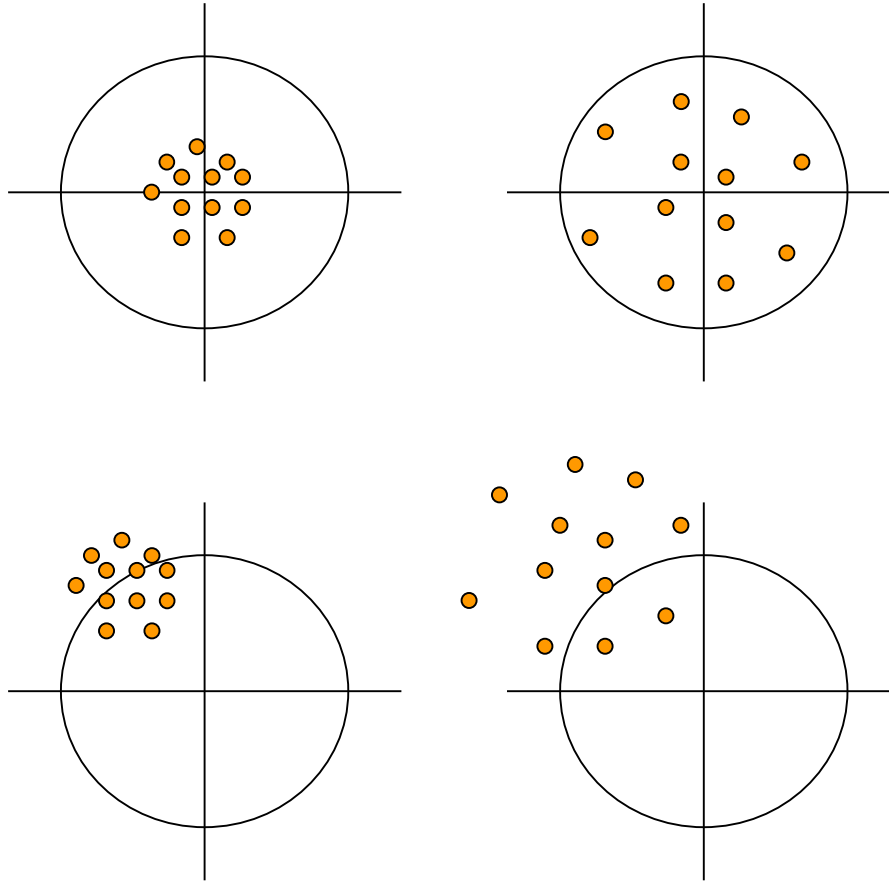
If the number of experiments be very large, we may have precise information



Random Variation

Systematic error

Biological
variability



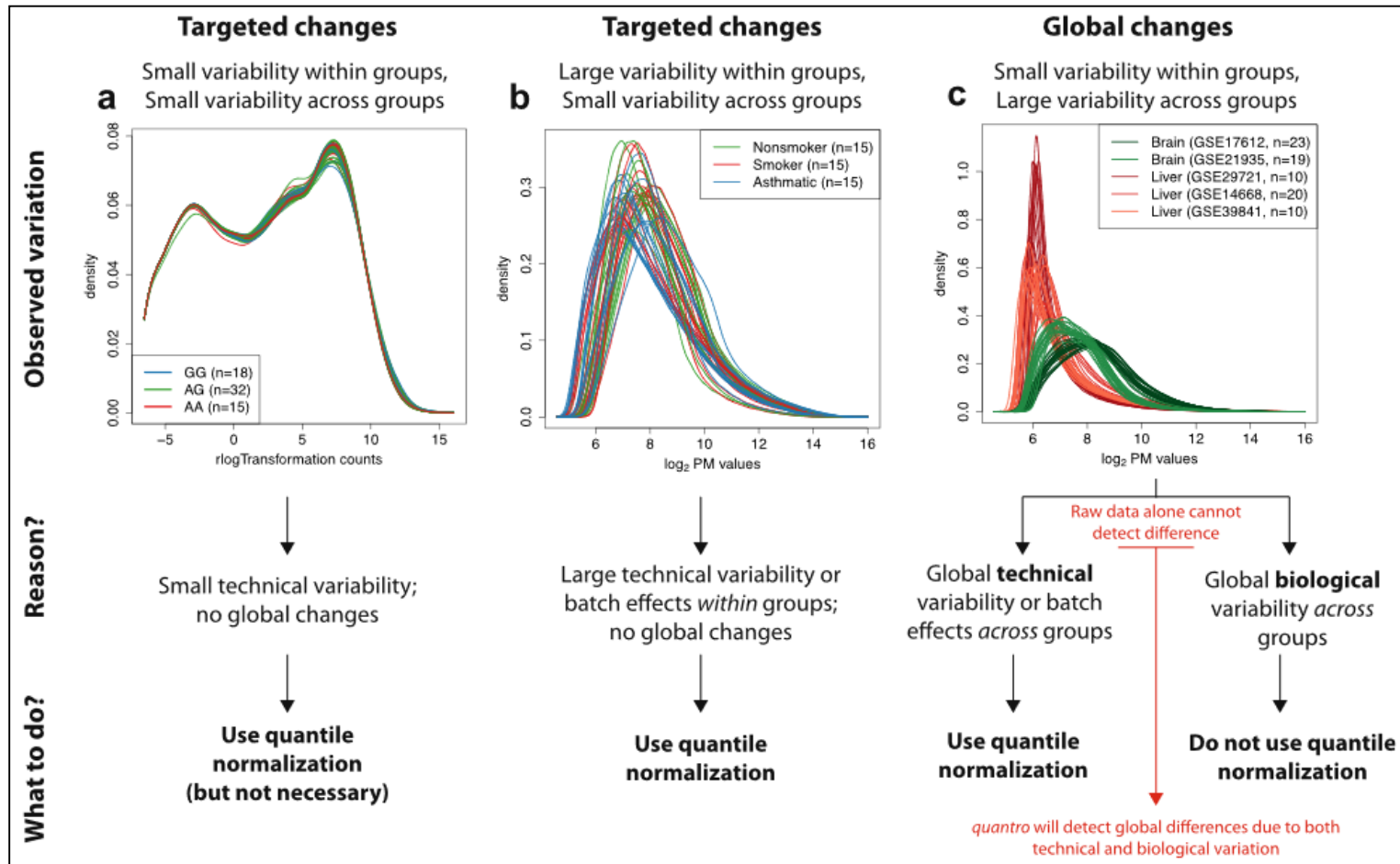
Technical bias

Data normalization

Taking care of bias (technical)

- Multi-sample global normalization methods (e.g. quantile normalization) have been successfully used to remove technical variation.
- These methods rely on the assumption that observed global changes across samples are due to unwanted technical variability.
- Applying global normalization methods has the potential to remove biologically driven variation.

How to normalize?



quantro: a data-driven approach to guide the choice of an appropriate normalization method

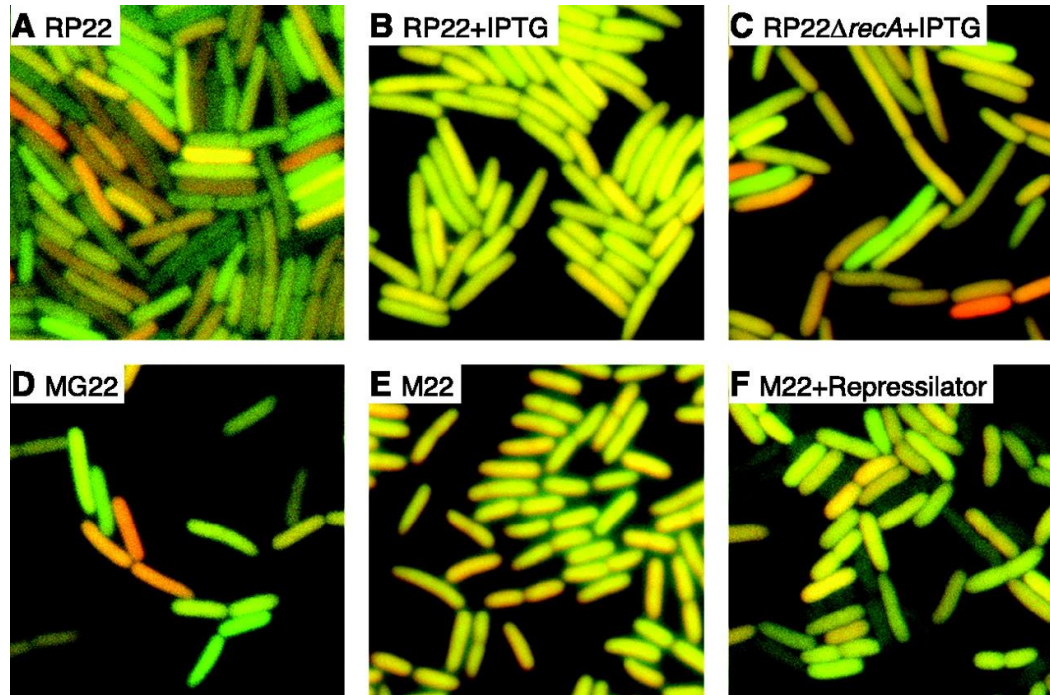


Quantifying biological variability: we need replicates!

How much?

The frequency of occurrence.

The numbers of times each possible outcome occurs in a sample.

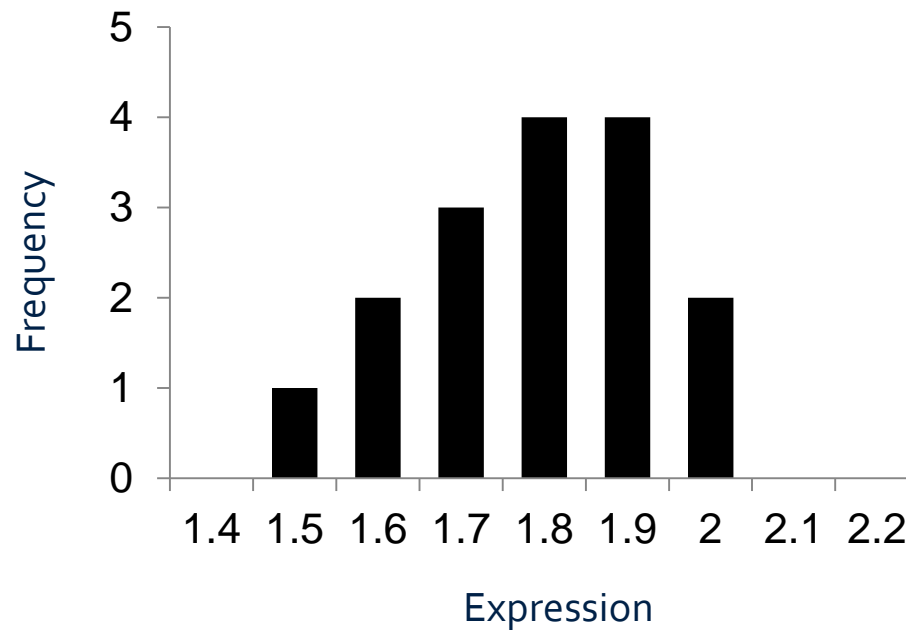


Distribution visualization

How much?

The frequency of occurrence.

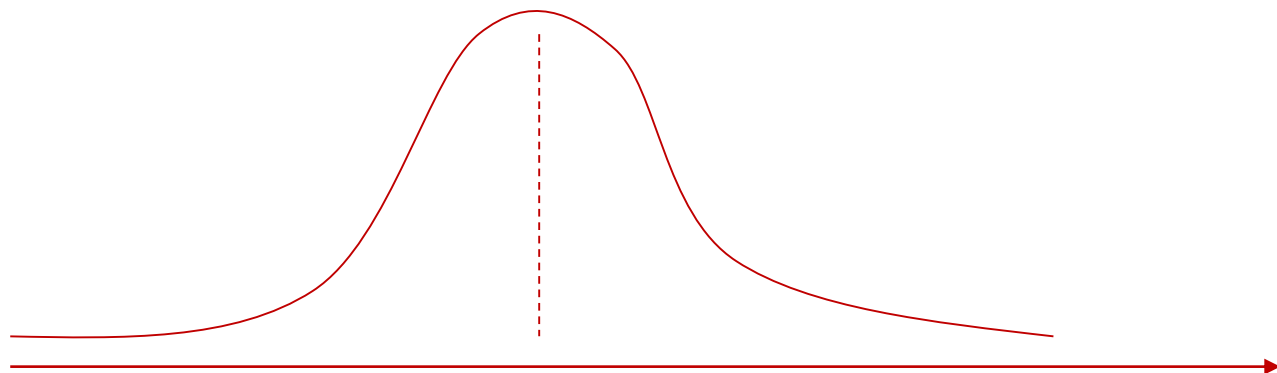
The numbers of times each possible outcome occurs in a sample.



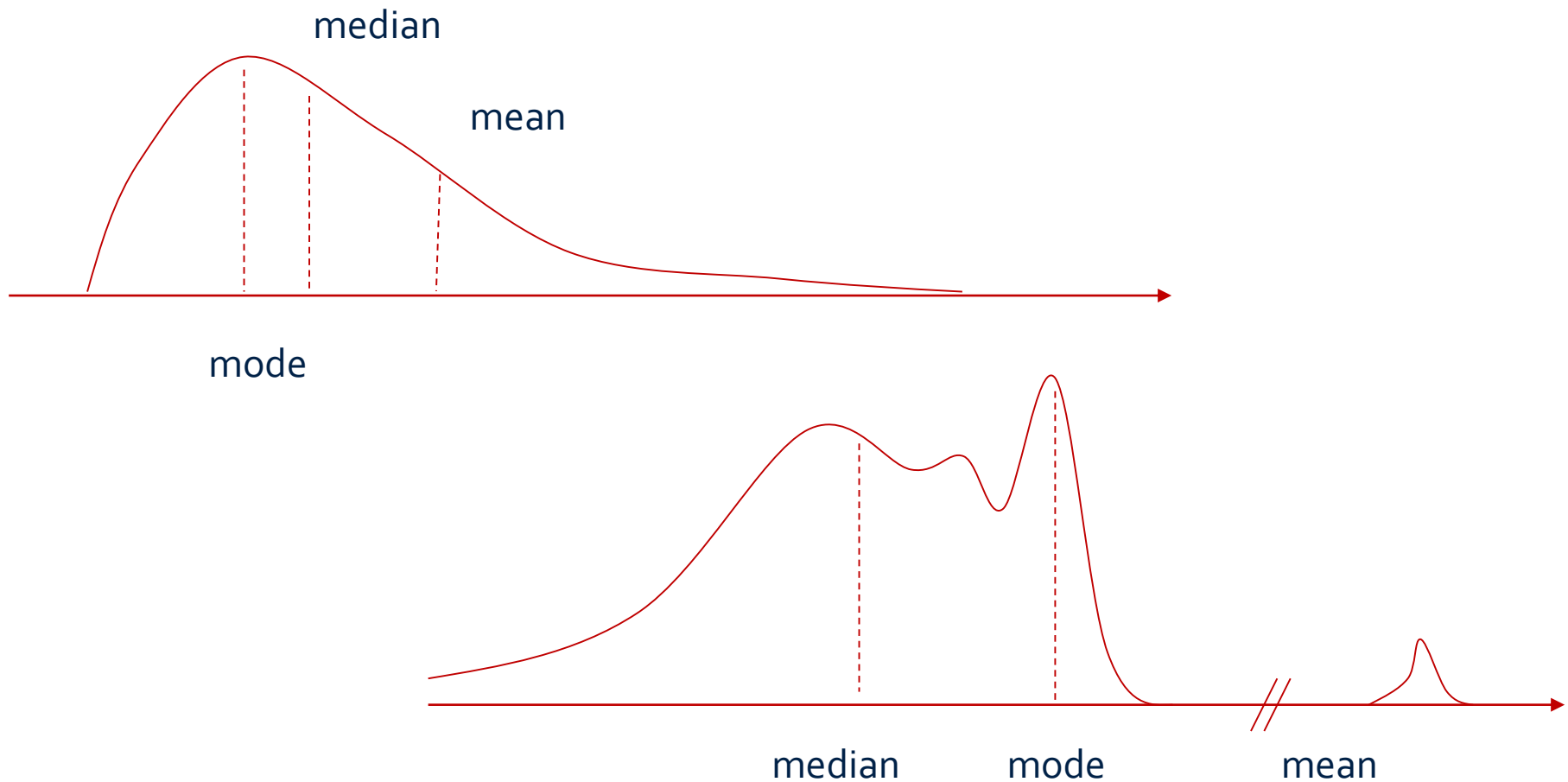
How do we describe a distribution?

LOCATION
or average of the distribution

- **Median** is the value m for which $p(x < m) = p(x > m) = 1/2$
 - **Mode** is the most common value
 - **Mean** is the expectation



Median= Mean = Mode



“Which one should I use?”

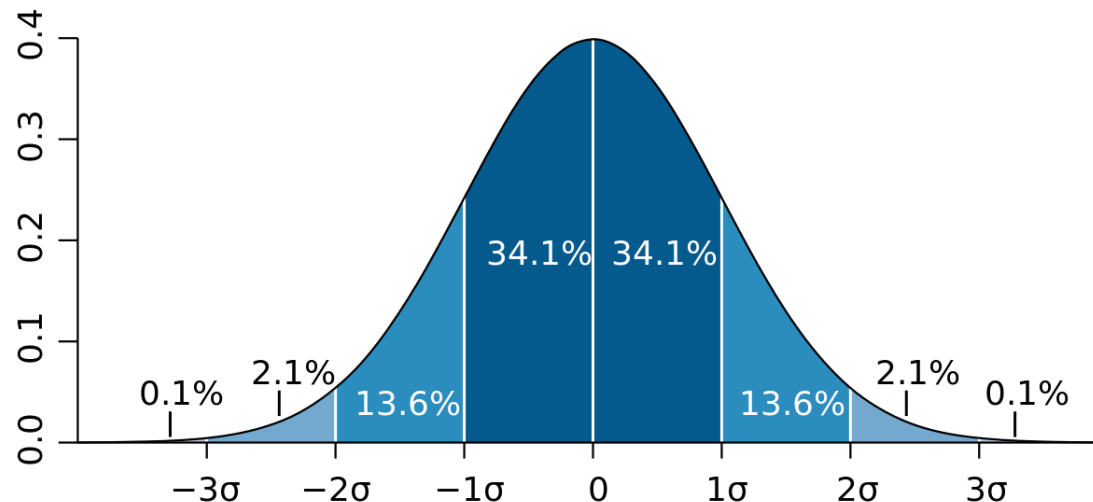
- Measures of location are used to summarise information about a distribution.
- The measure of location to use is the one that best summarises the data in the specific case.

How do we describe a distribution?

Spread
or scatter about the average of the distribution

- **Variance:** describes how far the numbers lie from the mean
- **Standard deviation** = square root of variance
- **Coefficient of variation** = standard deviation / mean = noise / signal
- **Range:** smallest interval which contains all the data
- **Interquartile range:** difference between the upper and lower quartile

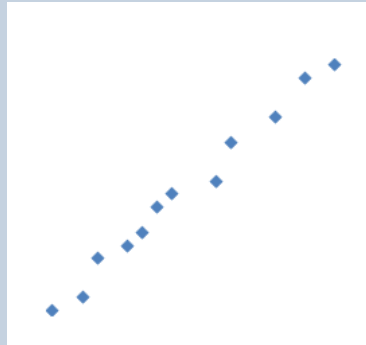
Example:
normal
distribution



Are my data normally distributed?

Empirical distribution \leftrightarrow Theoretical distribution

Expected
normal
value



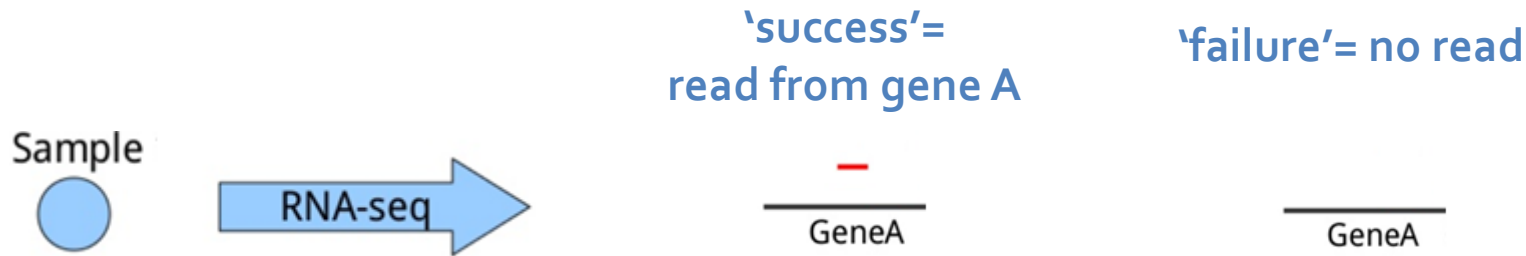
Variable

Observed values of a single numeric variable are plotted against the expected values if the sample were from a normal distribution. If the sample is from a normal distribution, points will cluster around a straight line.


- P-P plots compare the cumulative distribution functions
- Q-Q plots compare the quantiles of the two distributions

Detecting reads: Bernoulli trials

Two possible outcomes:



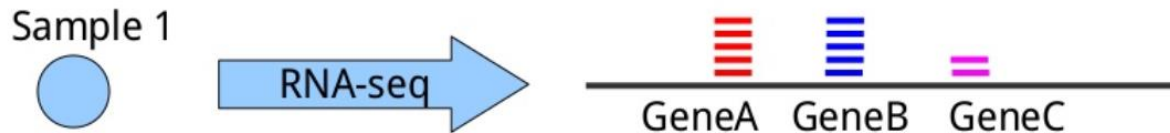
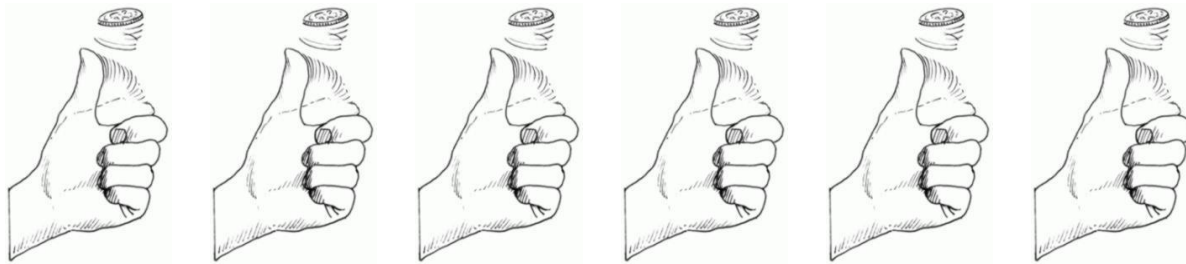
Bernoulli random variable X :
number of successes (either 0 or 1) in 1 trial where the probability of success is p .


$$p_X(x) = \begin{cases} p^x (1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

An orange arrow points from the text 'probability of success is p' to the p^x term in the equation. The term $x = 0, 1$ is circled in orange.

Counting reads

Sequence of Bernoulli trials \rightarrow Binomial distribution



If:

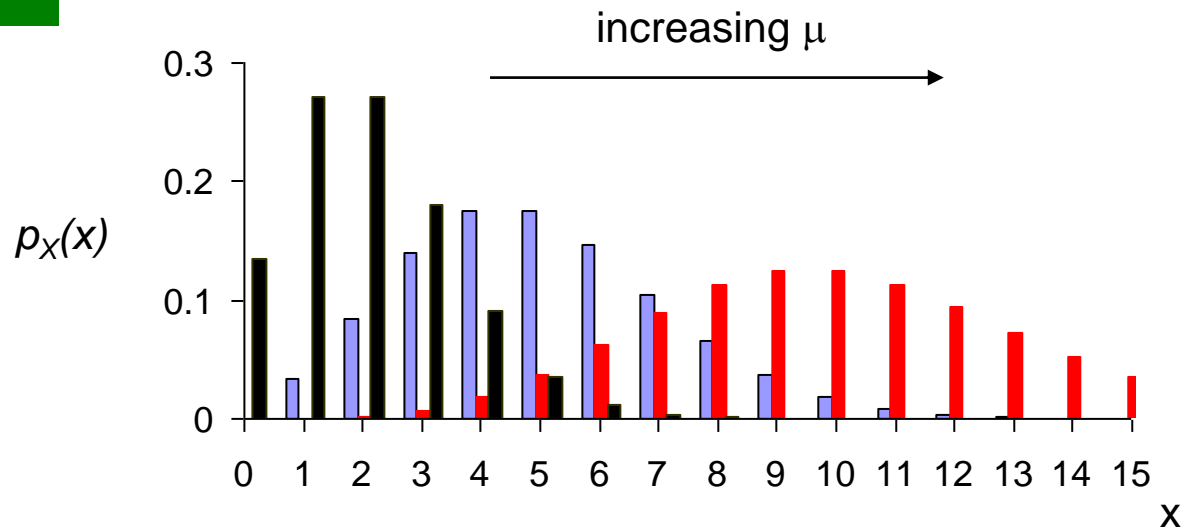
- trials are independent
- each trial results in 2 possible outcomes: success or failure
- probability of success p is constant from trial to trial



Rare events: Poisson distribution

$$p_X(x) = \begin{cases} e^{-\mu} \cdot \frac{\mu^x}{x!}, & x = 0, 1, 2, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

Low counts region

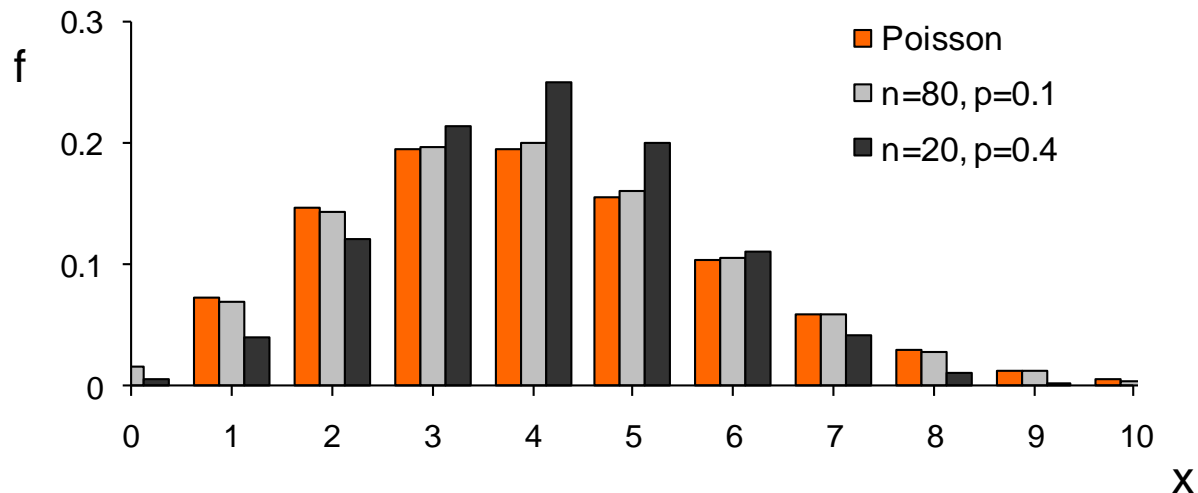


Binomial $[n, p] \rightarrow$ Poisson $[m=n \cdot p]$

when:

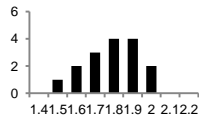
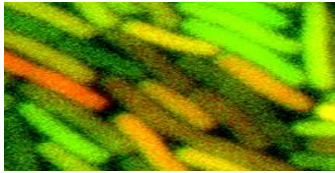
the number of trials, n , is large and
the probability of success in each trial, p , is small

Low counts region



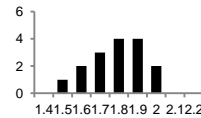
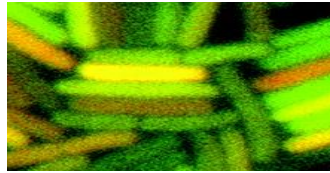
Sampling distribution

S₁



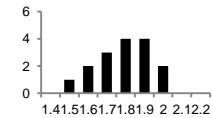
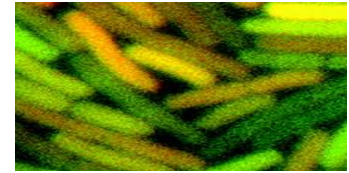
Average

S₂



Average

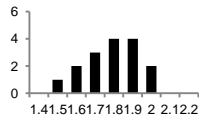
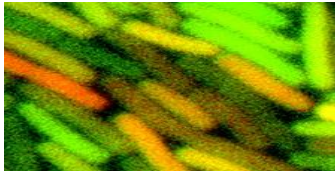
S₃



Average

Sampling distribution

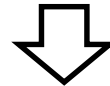
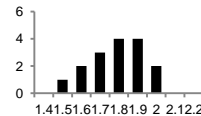
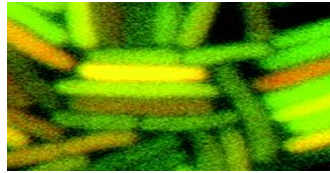
S₁



Average



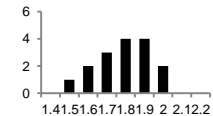
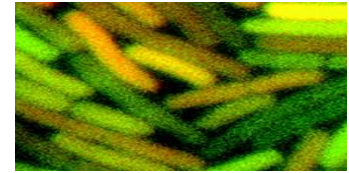
S₂



Average



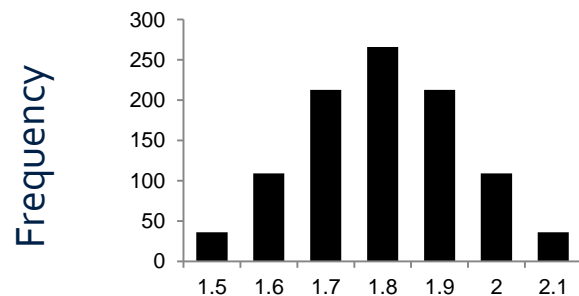
S₃



Average

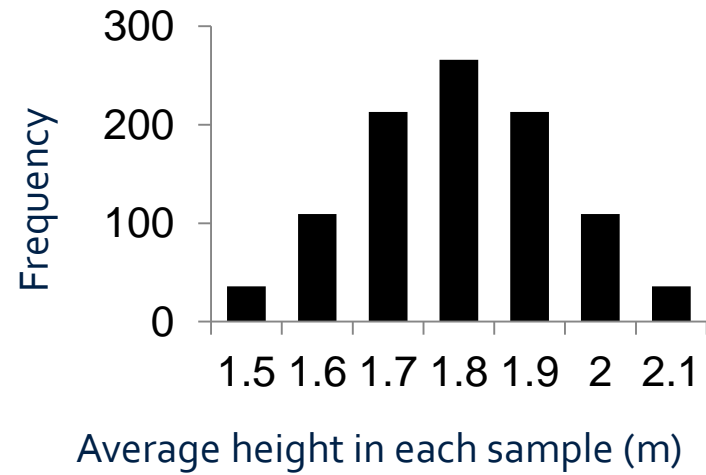
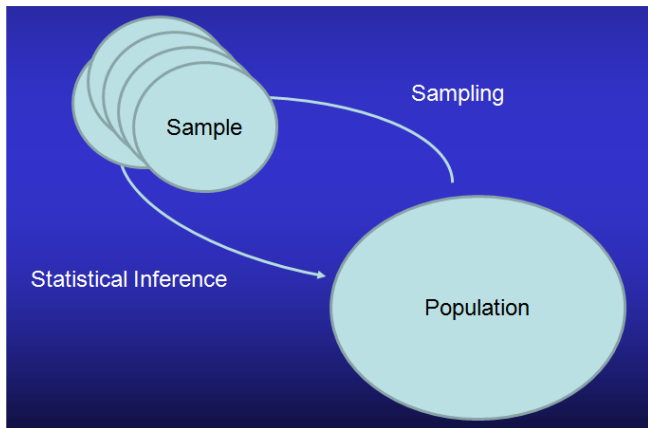


Distribution across samples



Average height in each sample (m)

What can I say about the whole population given my observation?



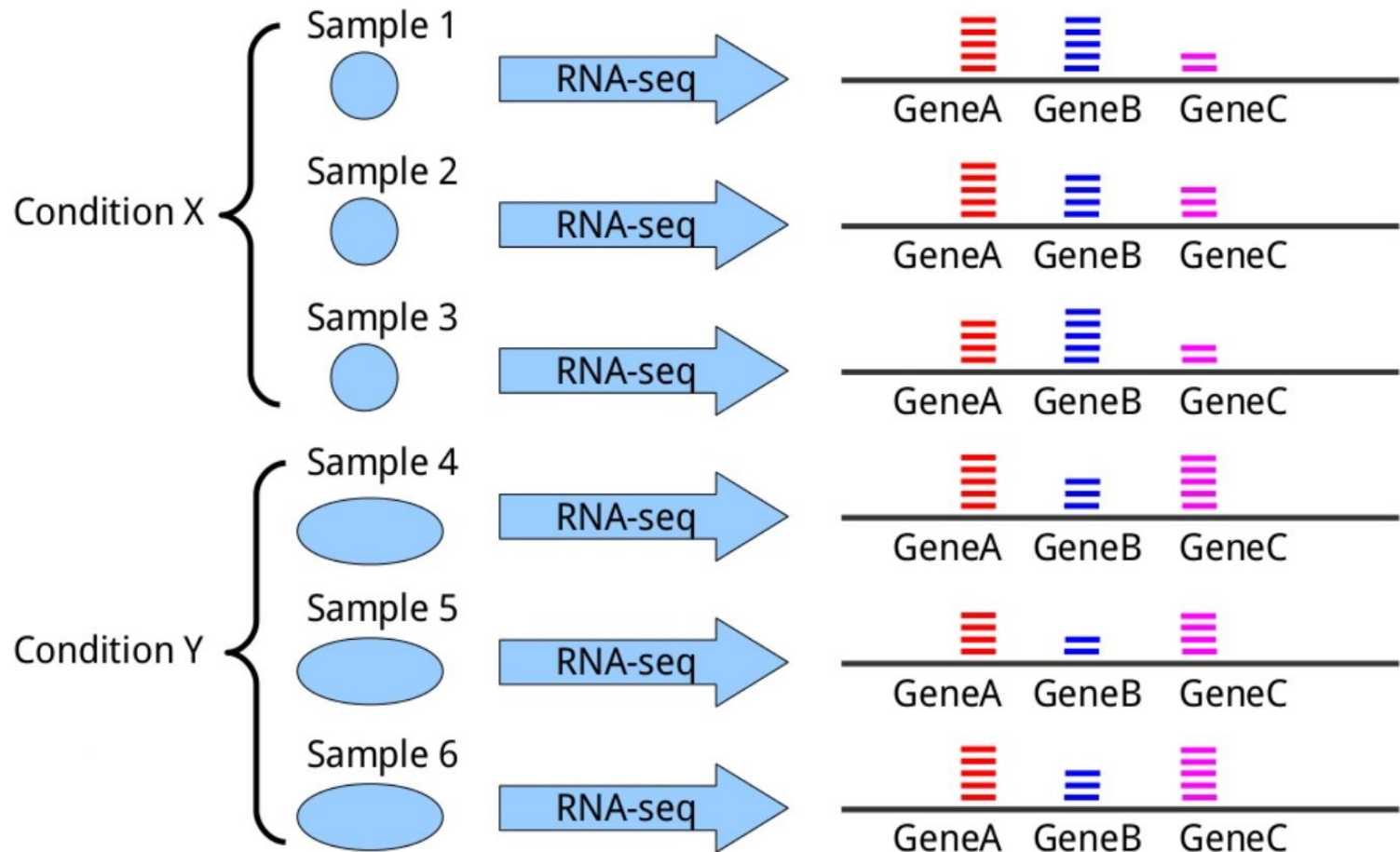
- The sampling distribution ~ normal distribution if high number of samples
- Mean of whole population ~ Mean of sampling distribution
- Standard error of population ~ deviation of samples means

(Central Limit Theorem)

Quantifying uncertainty: confidence intervals

- The range of values around the mean for which if we drew an infinite number of samples of the same size from the same population, $x\%$ of the time the true population mean would be included in the confidence interval calculated from the samples.
- It gives us the information about the **precision of a point estimate** such as the sample mean.
- CI will tell you the **most likely range of the unknown population mean**

Differential expression

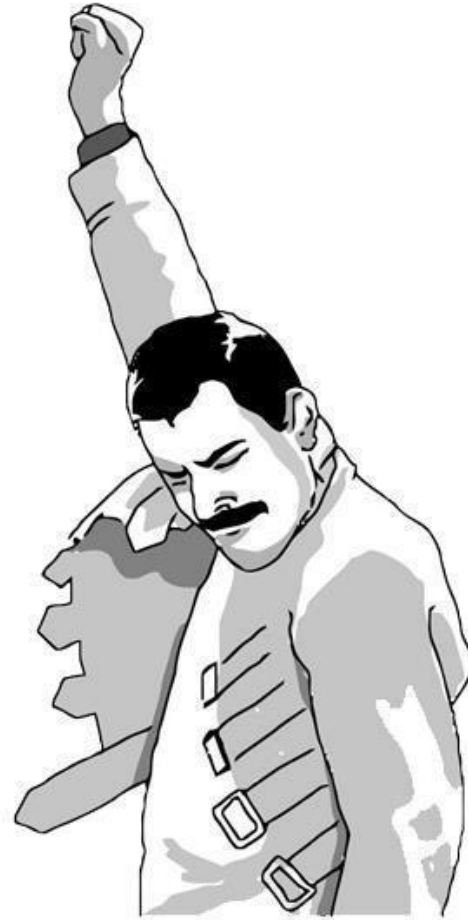


Statistical inference: basic assumption

Improbable events do not occur


- **Make a statement**
 - Gene G is over-expressed in cancer
- **Define a null hypothesis**
 - Expression of gene G is not different between normal and cancer samples
- **Find probability of same or 'more extreme' result by chance**
 - p-value
- **If the probability is small, we can reject the null hypothesis**

$p < 0.05$



WARNING!

- The opposite is not true!!!!
 - Two conditions are not proven equal by showing that the p-value is high!
- **Statistics can only reject a null hypothesis – not prove it!**

- 
- ~~▪ Expression of gene G is not different between normal and cancer samples~~
 - **Reject the null hypothesis that gene G is not different between normal and cancer samples**

Little history of the p-value

Sir Ronald A. Fisher (1890-1962):

- *"...it is certain that the interest of statistical tests for scientific workers depends entirely from their use in rejecting hypotheses which are thereby judged to be incompatible with the observations."*
- *"...if one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred. Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fails to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance..."*
- *"... no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."*



American Statistical Association Statement, 2016

- *"Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning"*
- *"Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value"*

Testing a hypothesis

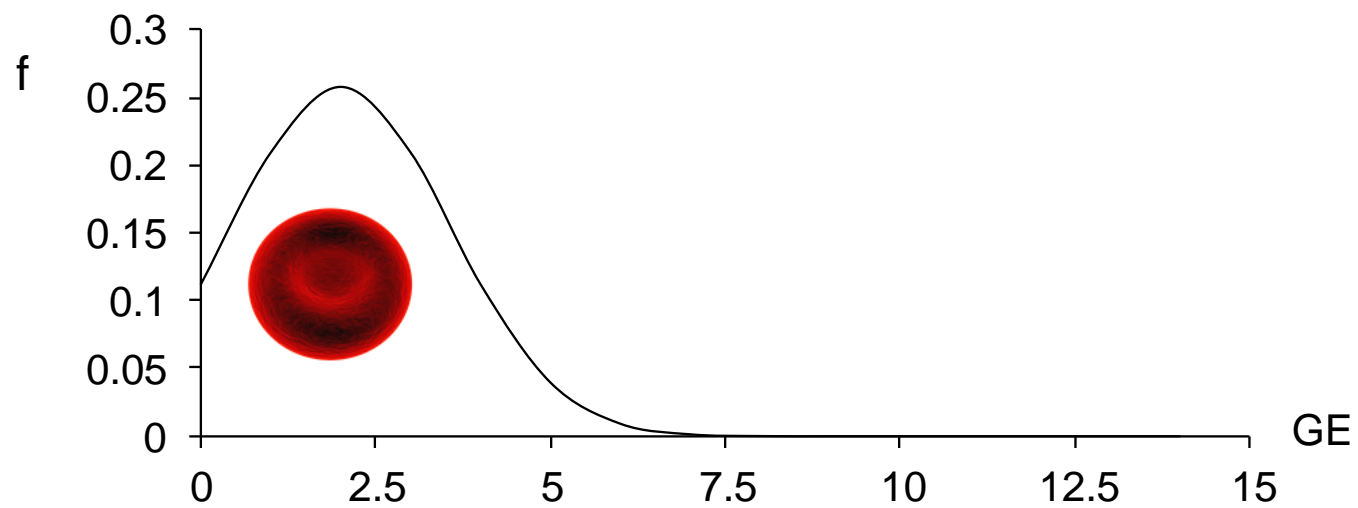
We make a claim a difference and we test it:

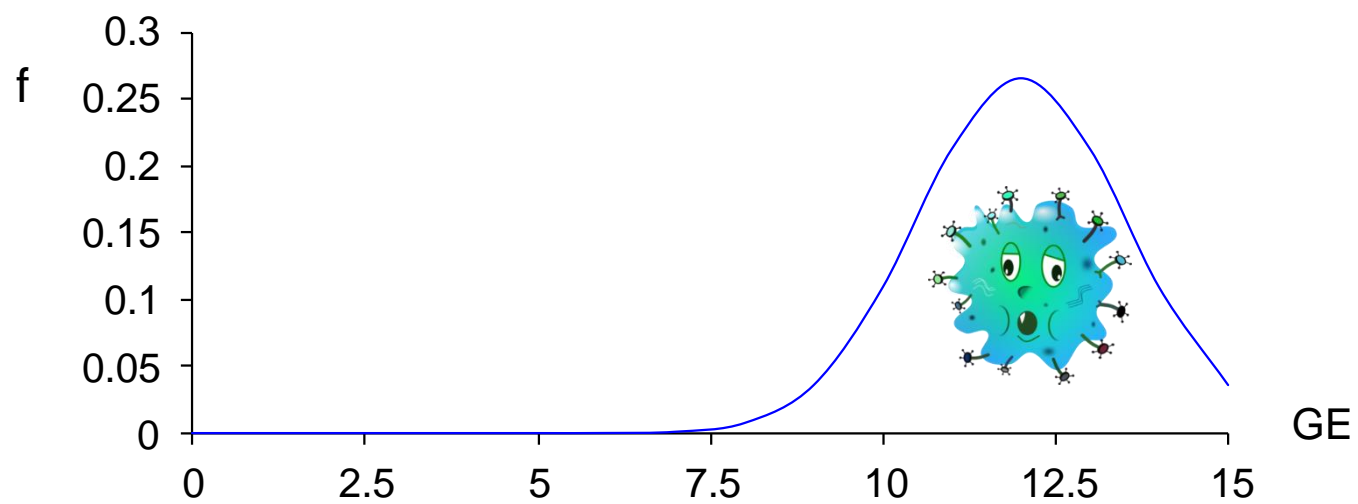
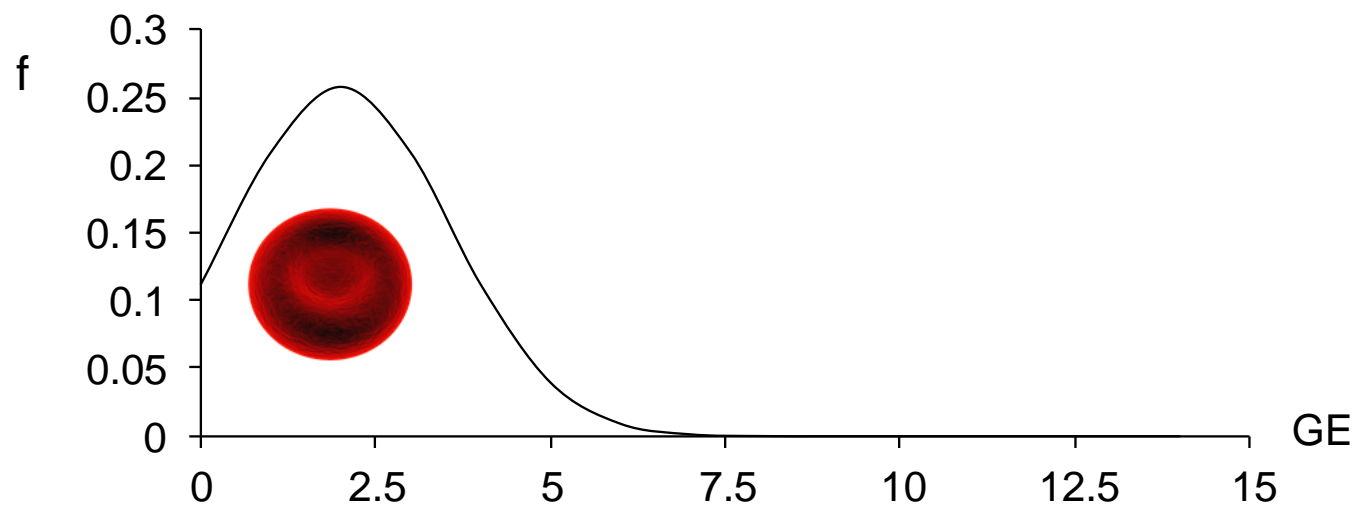
- **We** state a null hypothesis, H_0 (No disease)
 - ▶ We assume that H_0 is true
- How much evidence do **we** want before rejecting H_0 ? ▶ **Level of significance**
- We collect the sample
- How far is our sample from the null?
 - ▶ We Calculate a test statistics: a score which measure the distance of our sample from the null
- What is the probability of obtaining this score if H_0 is true?
 - ▶ We obtain a p value for the test statistics
- We make a decision to accept or reject the null hypothesis based on the p value

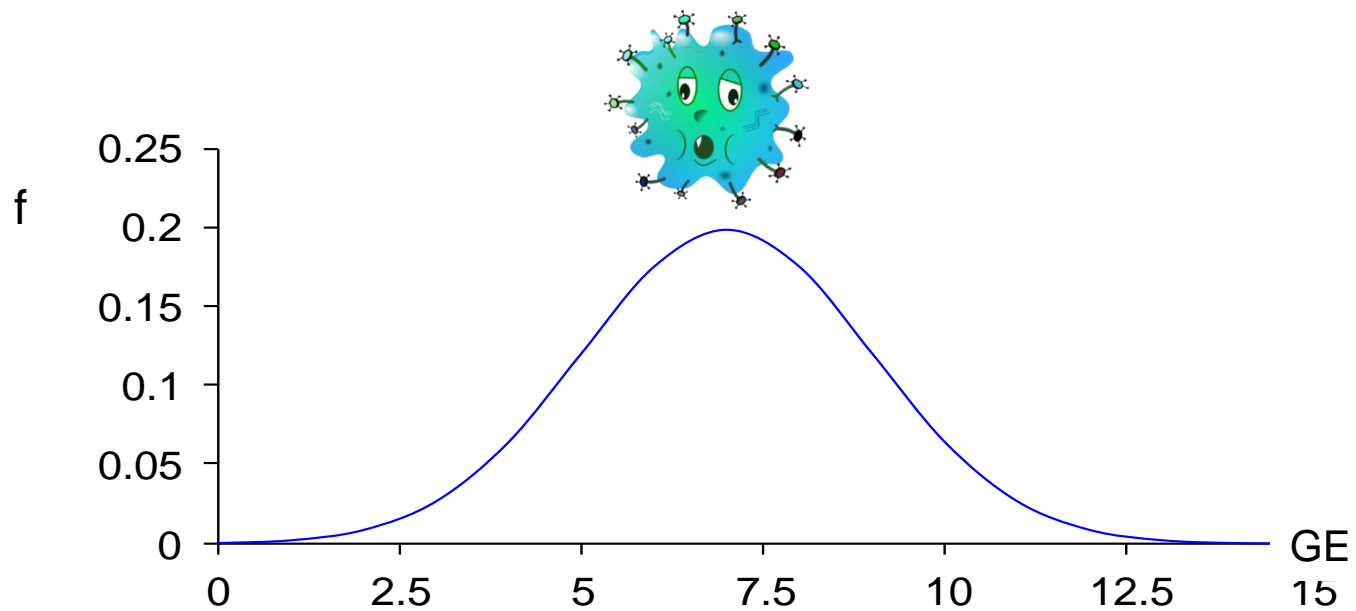
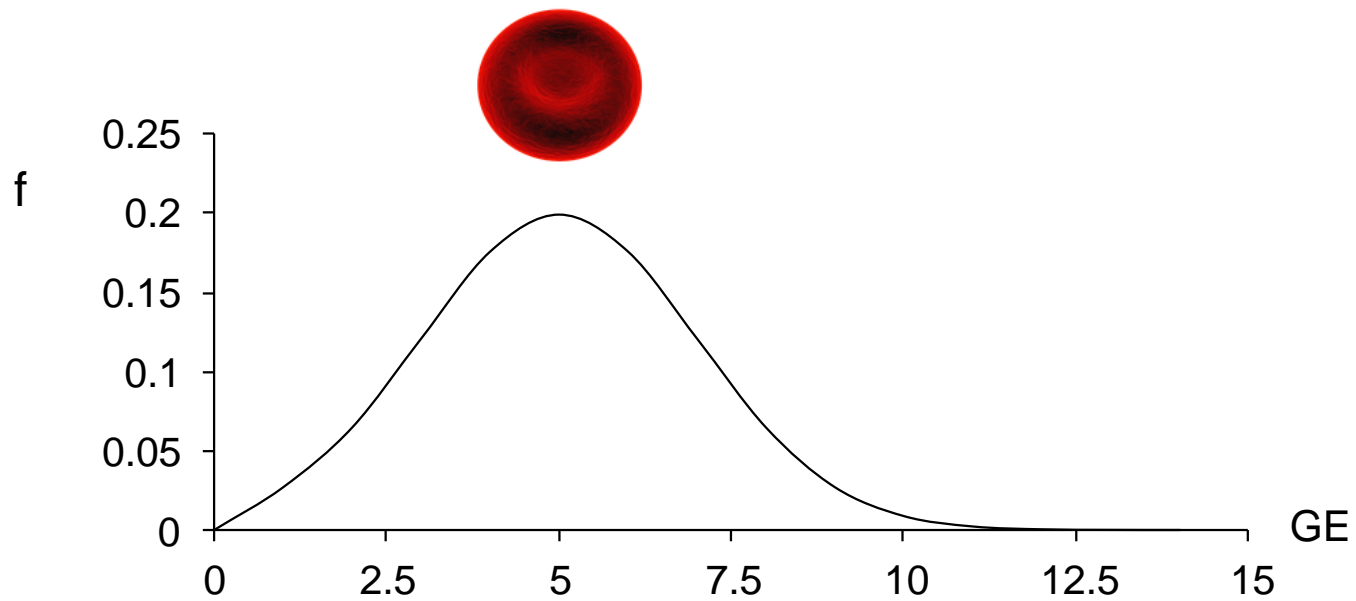
REJECT: we have evidence to reject H_0 / **ACCEPT:** we do not have evidence to reject H_0

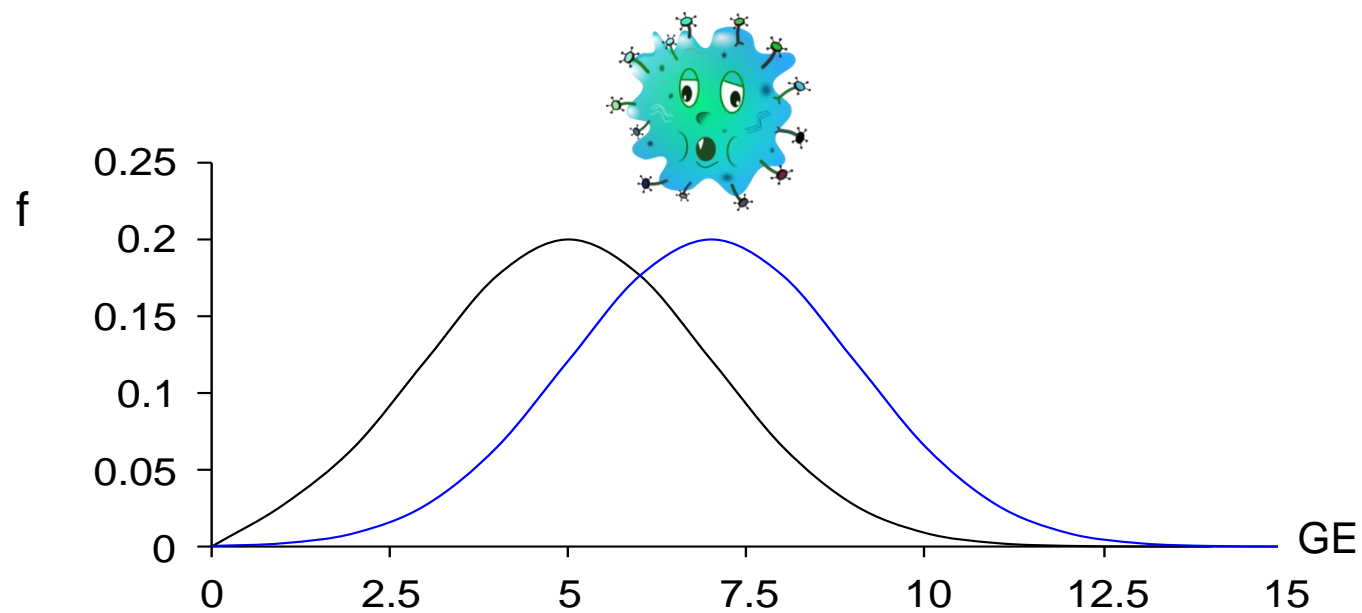
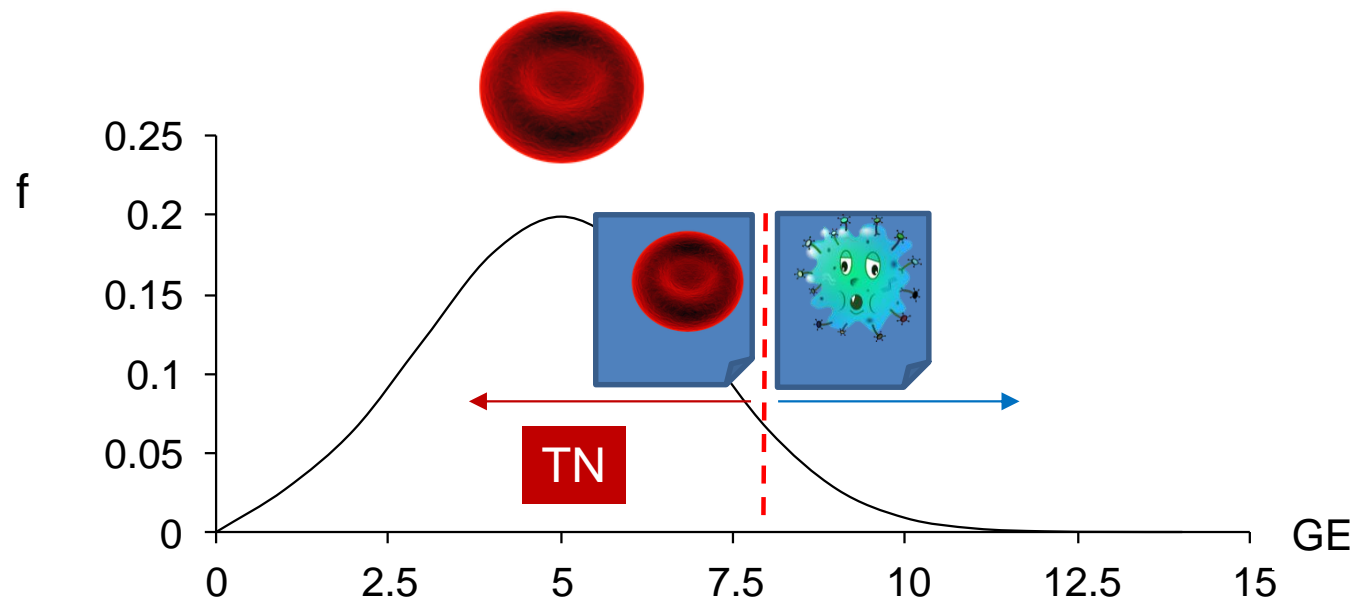
STATISTICAL ERRORS

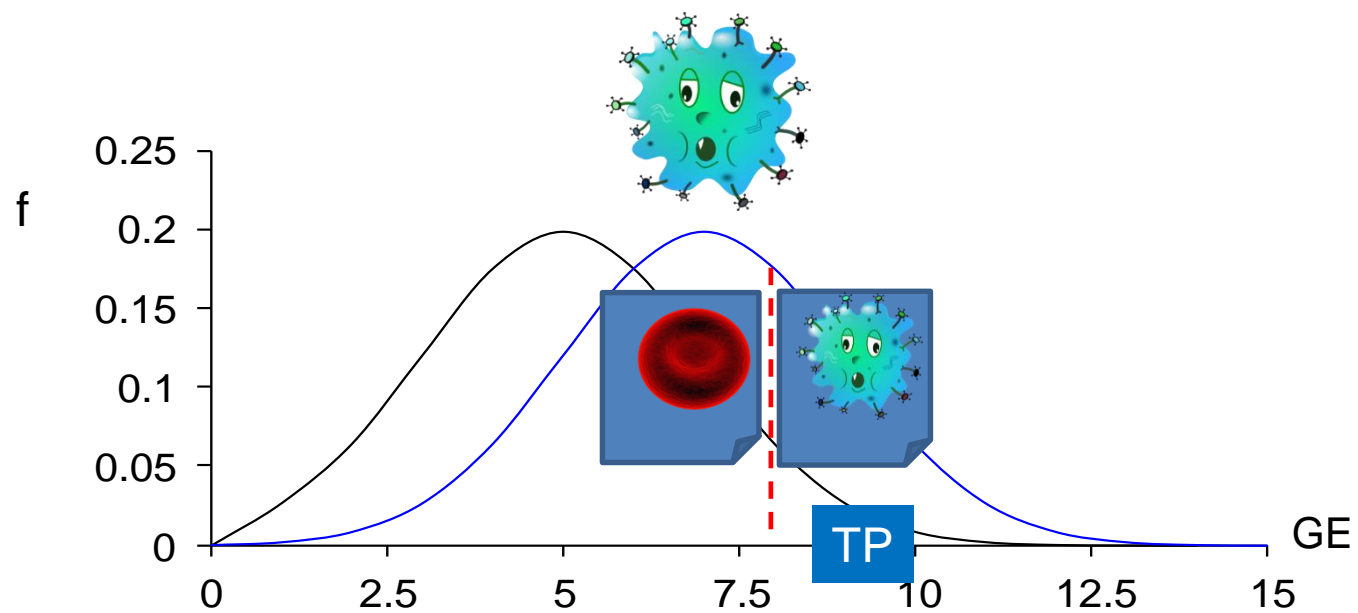
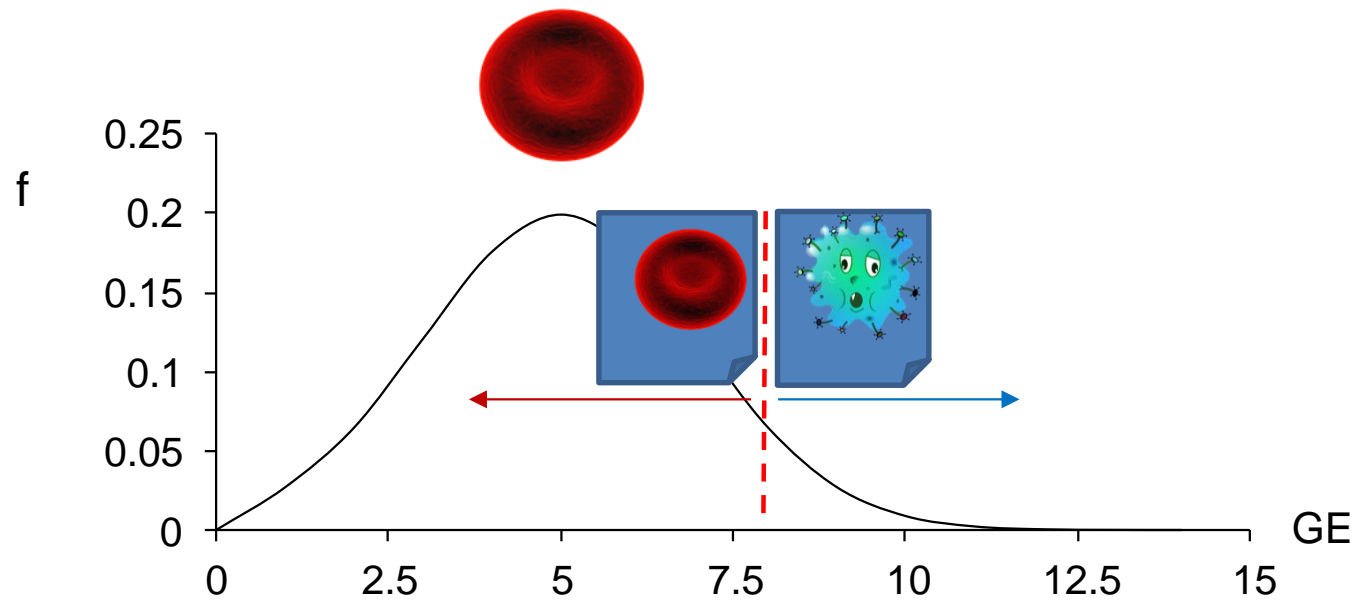
Test result Real status	Fail to reject H_0	Reject H_0
H_0 is true (No difference)	Correct decision TRUE negatives $P=1-\alpha$	Type I error FALSE positives $P=\alpha$
H_a is true (Difference)	Type II error FALSE negatives $P=\beta$	Correct decision TRUE positives $P=1-\beta$

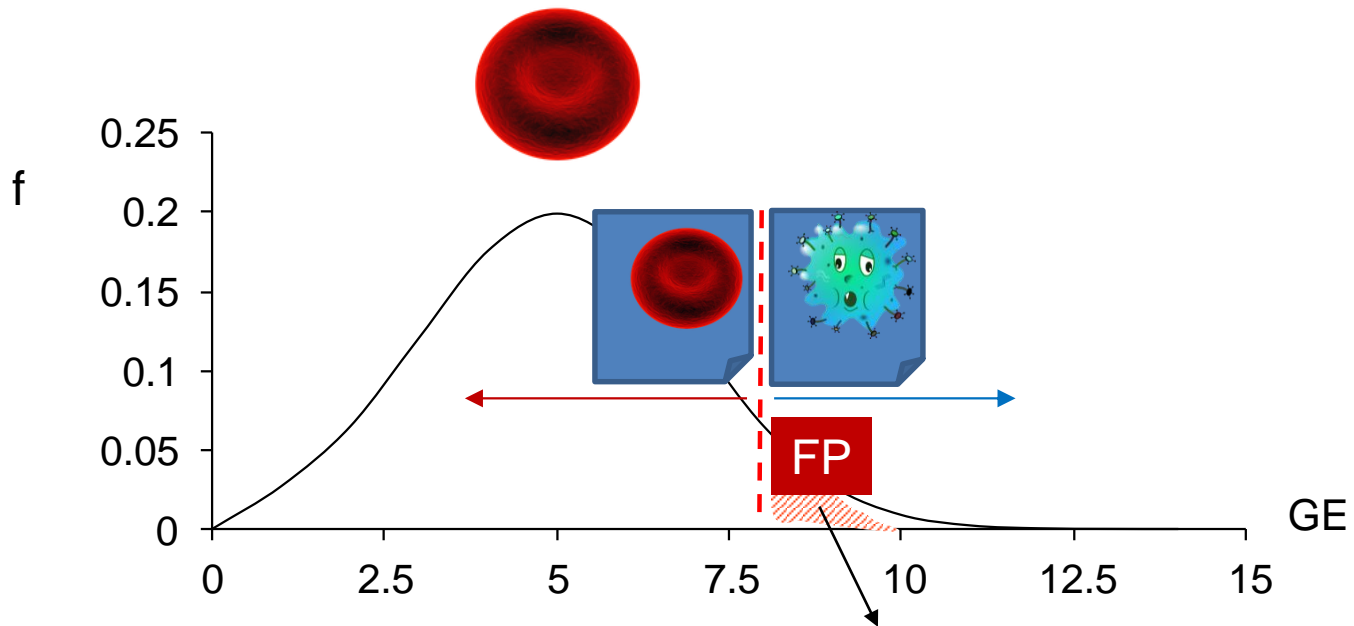




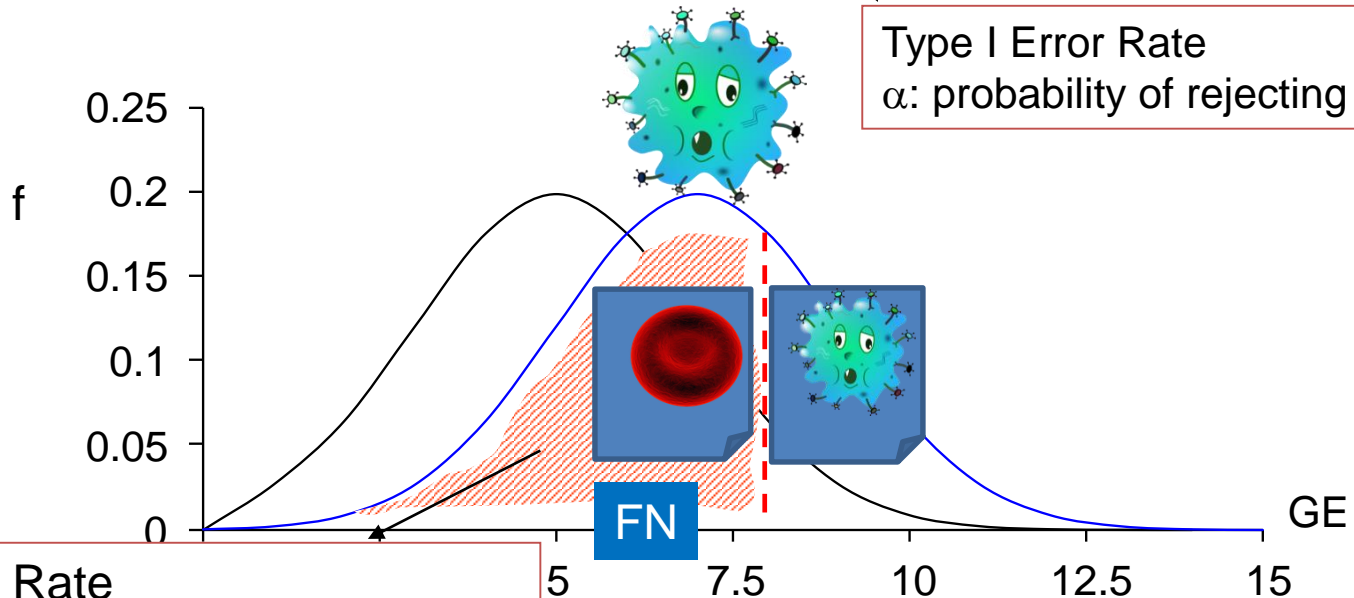






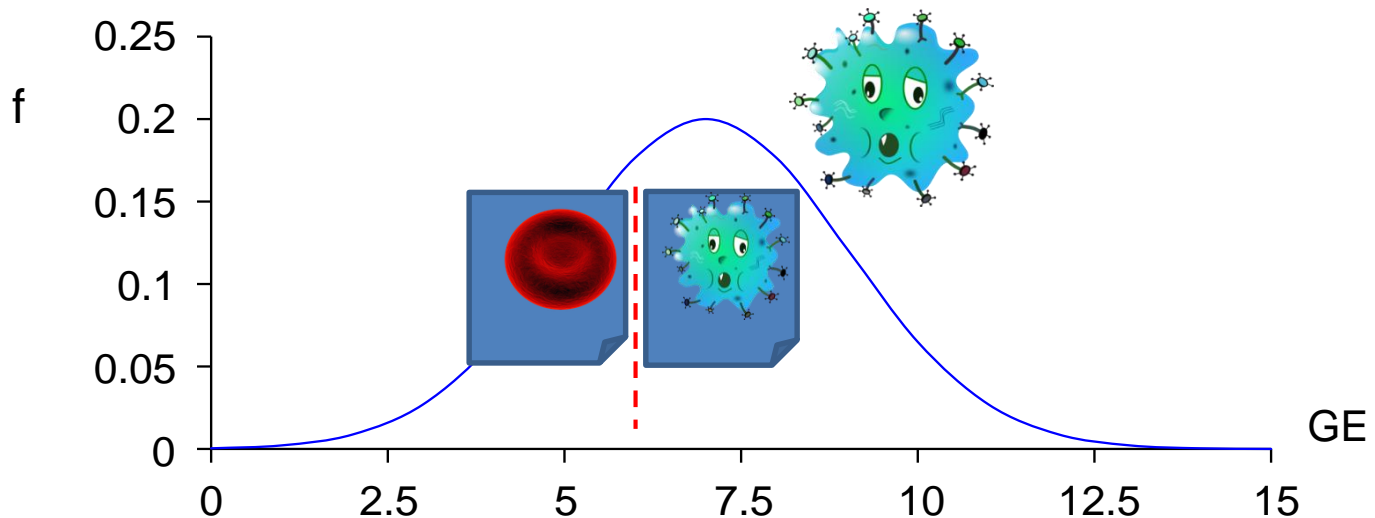
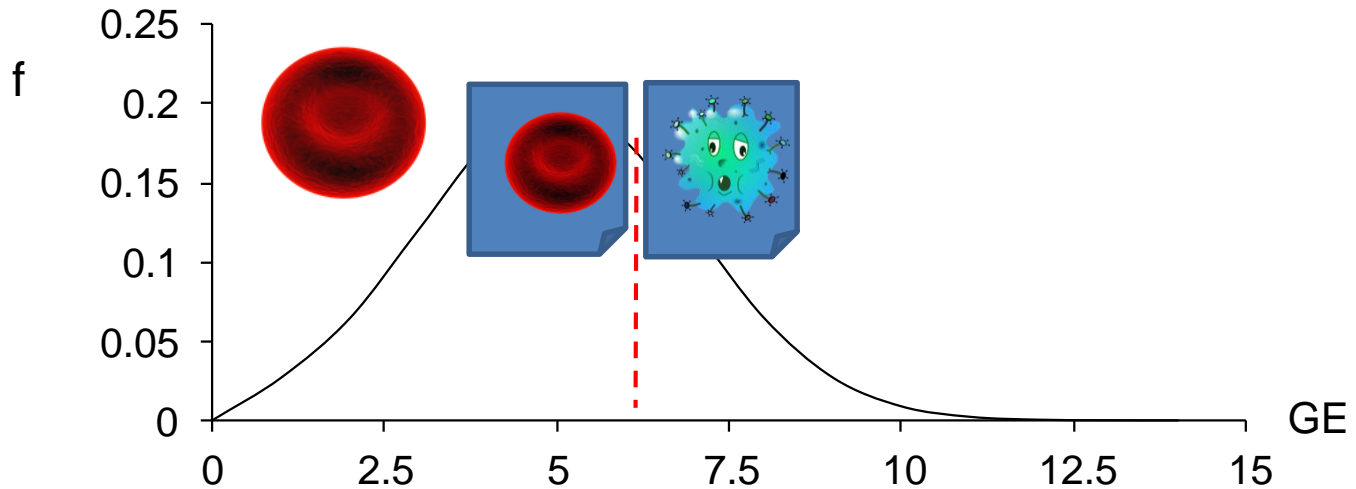


Type I Error Rate
 α : probability of rejecting a true H_0



Type II Error Rate
 β : Probability of accepting a false H_0

If we set α too high we increase the number of false positives
If we set α too low we might miss a true difference



STUDY SIZE

Power:

the probability of detecting a real difference as specified by the alternative hypothesis

Test result \ Real status	Fail to reject H_0	Reject H_0
H_0 is true (No difference)	Correct decision TRUE negatives $P=1-\alpha$	Type I error FALSE positives $P=\alpha$
H_a is true (Difference)	Type II error FALSE negatives $P=\beta$	Correct decision TRUE positives $P=1-\beta$

To minimize the false positives AND maximize the power we need to increase sample size.

Hypothesis testing and sample size

$p=0.45$

	Apoptosis	No apoptosis	Proportion
Control	5	15	25%
siRNA	10	15	40%

If $p > 0.05$: no effect or my sample is too small?

$p=0.05$

	Apoptosis	No apoptosis	Proportion
Control	20	60	25%
siRNA	40	60	40%

The p value depends on effect size but also on sample size!

Which test?



Parametric or non parametric?

	Positive	Negative
Parametric	Provides description of distribution of data	Relies on assumption about distribution
Non-parametric	No assumption on distribution	No information on distribution

	Output
Parametric	P-value + distribution of data
Non-parametric	P-value

Parametric statistics

- Assumes a parametric distribution of data
- Tests are based on this distribution
- Examples:
 - Paired and unpaired T-test
 - Binomial tests
 - Regression models

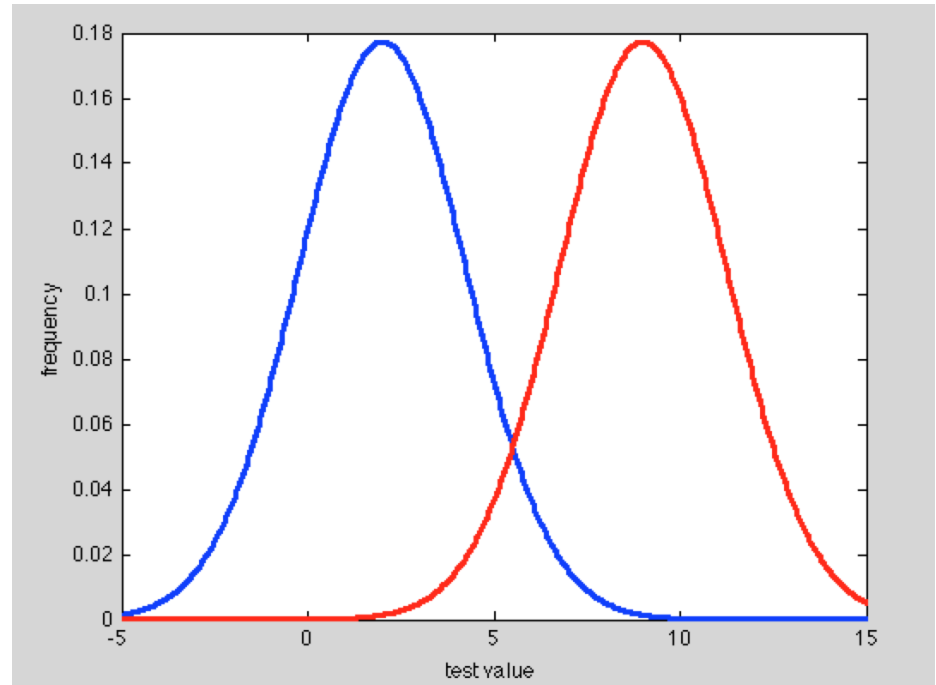
Compare two normally distributed series with equal variance: t-test (parametric)

What is under the hood of your statistical software?

Null hypothesis: the two samples come from the same normal distribution

Test quantity. Note common variance

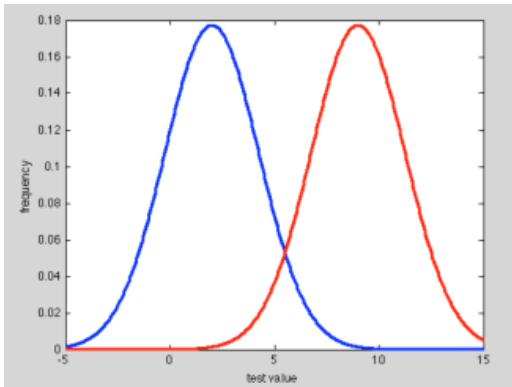
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$



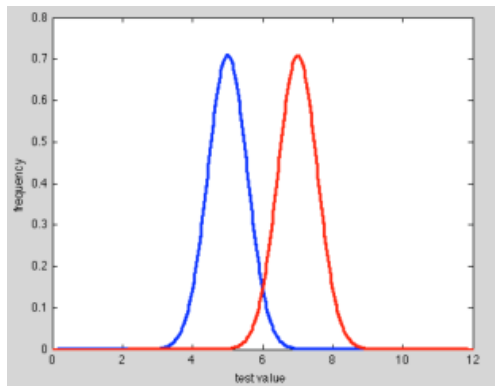
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

High power

Large effect size

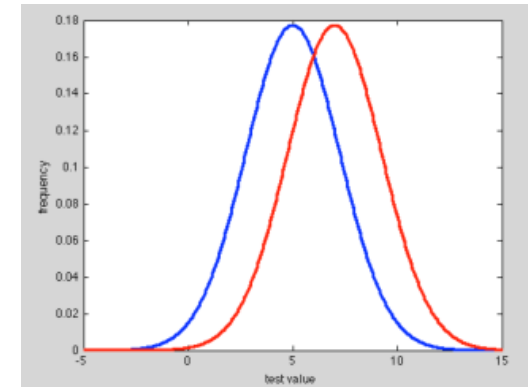


Small variance

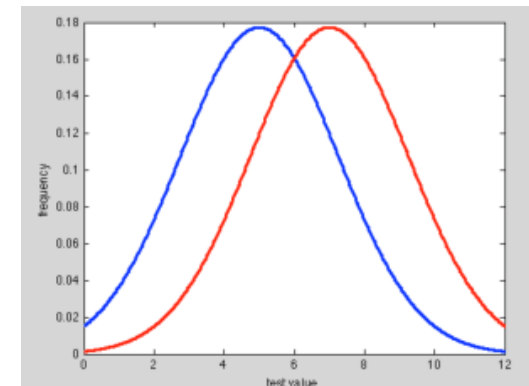


Low power

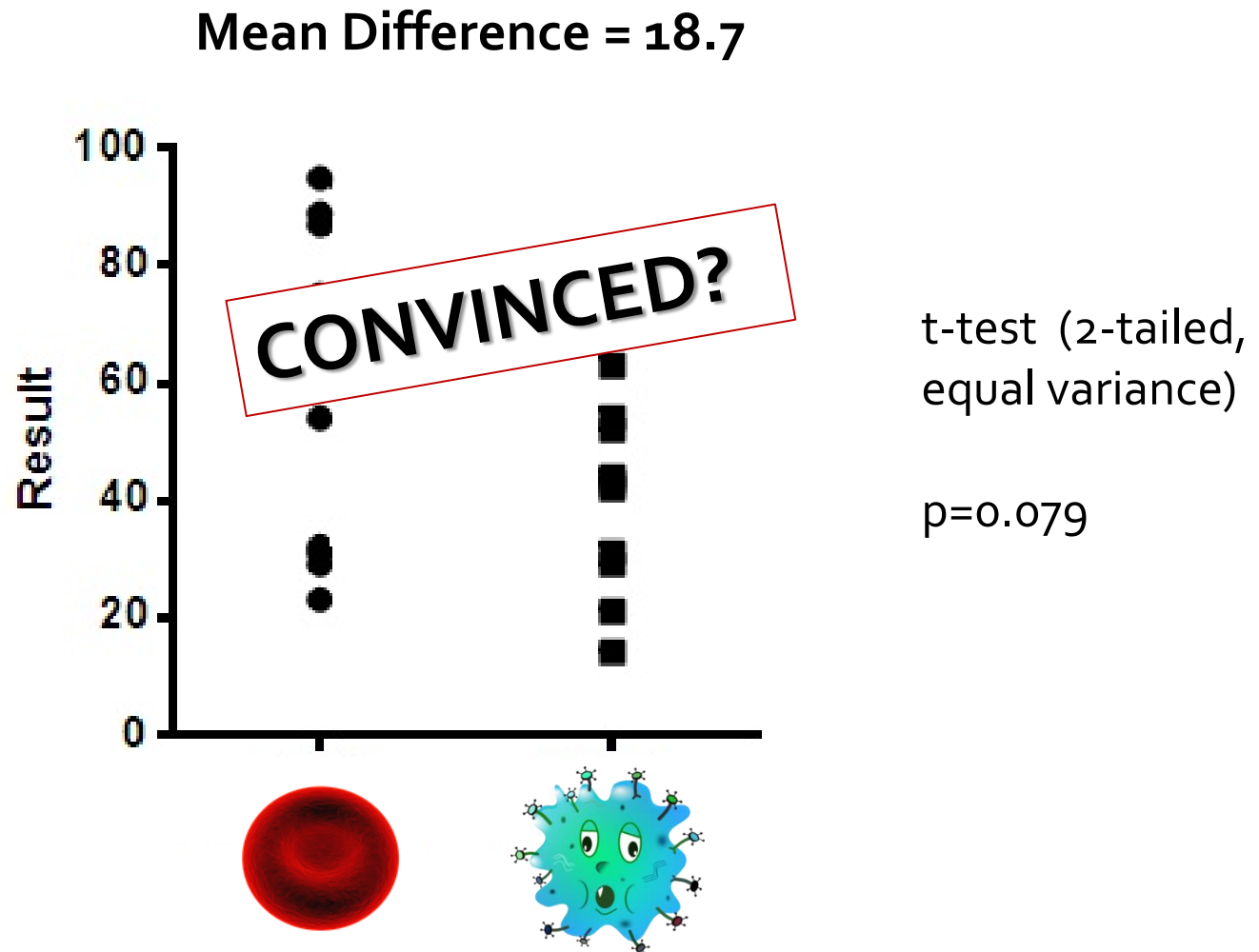
Small effect size



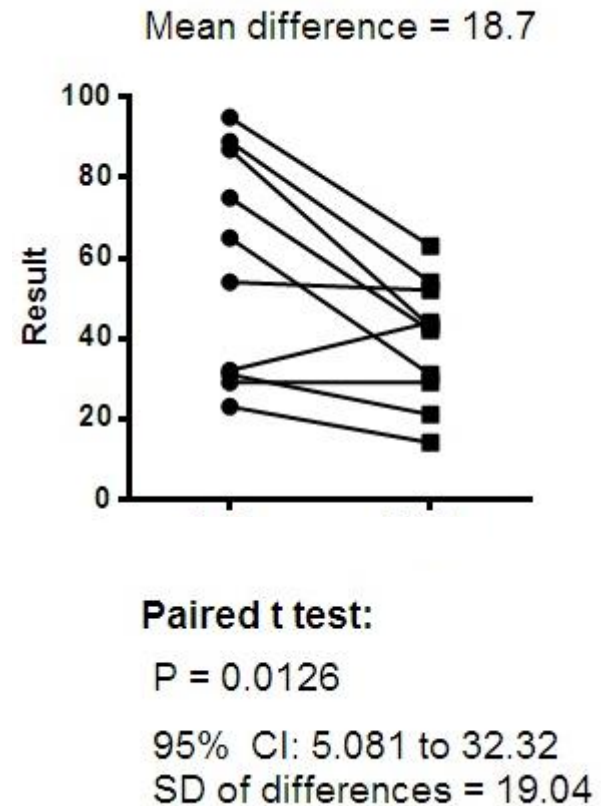
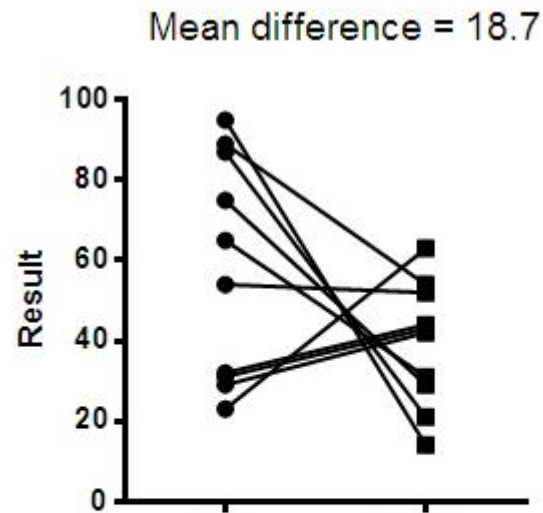
Large variance



Example

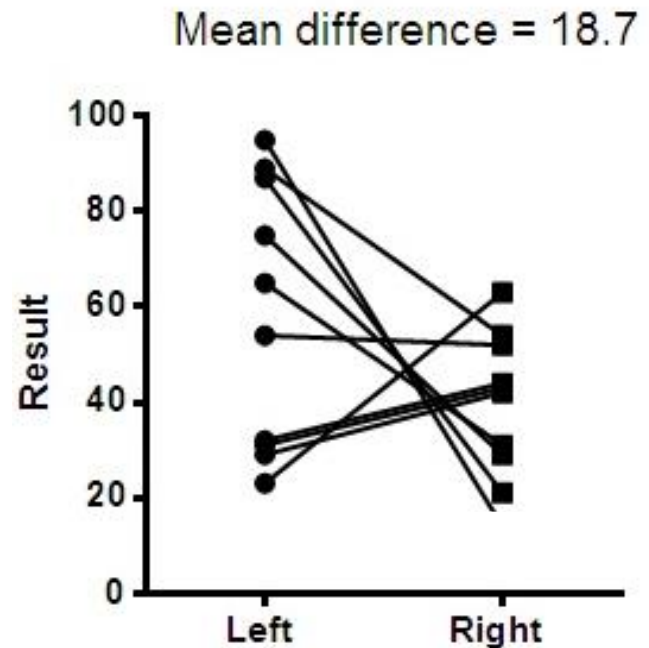
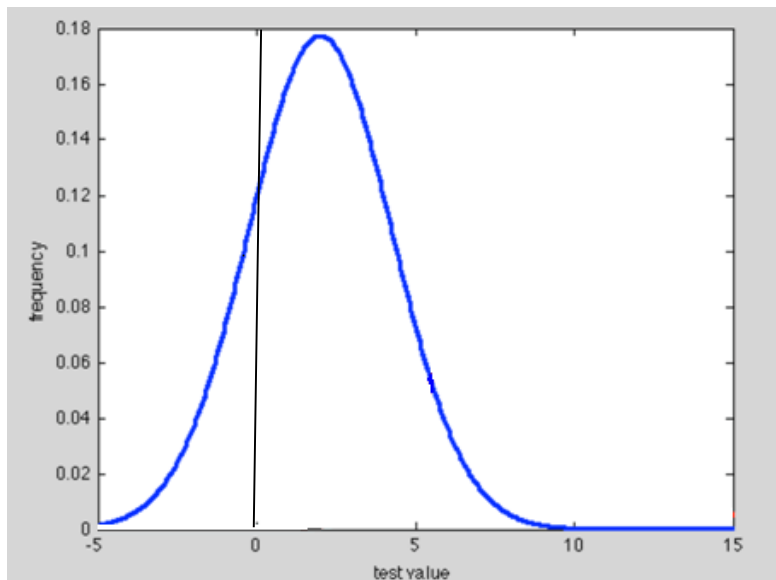


Data structure



Paired t-test

$$t = \frac{\sum (x_1 - x_2)}{\sqrt{\frac{s^2_{(x_1 - x_2)}}{n}}}$$



Paired t test:

P = 0.1677

95% CI: -9.492 to 46.89

SD of differences = 39.41

Non-parametric statistics

- No assumption about distribution
 - Example: rank tests
 - Rank all values
 - Under null hypothesis: group 1 beats group 2 as often as vice versa.
 - Compare with coin flip
 - Less than 5% probability of result?
 - p-Value is probability of that or 'more extreme' result

Examples

- Wilcoxon rank sum (signed or unsigned)
- Fishers exact test

Two-sample comparison: Mann-Whitney (non-parametric)

Null hypothesis: the distributions of the two groups are identical

- Rank cases based on measure value (order from smaller to greater value)

- What is the chance that random sampling would result in the mean ranks of the two series being as far apart (or more so) as observed in this experiment?

Raw Measures			Ranks		
S1	S2		S1	S2	
4.6	5.2		1	5.5	A & B Combined
4.7	5.3		2	7	
4.9	5.4		3	8	
5.1	5.6		4	10	
5.2	6.2		5.5	13	
5.5	6.3		9	14	
5.8	6.8		11	17	
6.1	7.7		12	19	
6.5	8.0		15.5	20	
6.5	8.1		15.5	21	
7.2			18		
average of ranks			8.8	13.5	11

A couple of good books

