Statistics of RNA-seq analysis

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
gene1	32.80405	0.359444	0.598072	0.601004	0.5478372	0.923764
gene2	4.01072	3.407763	1.649827	2.065527	0.0388732	0.641407
gene3	7.01837	0.743337	0.994100	0.747749	0.4546118	0.923764
gene4	1.51006	2.814822	2.464686	1.142061	0.2534287	0.923764
gene5	11.23166	0.480522	0.894709	0.537071	0.5912189	0.923764
...	...	...	...	...	...	...
gene96	16.21864	0.684962	0.809892	0.845745	0.3976952	0.923764
gene97	2.91349	1.784327	1.790046	0.996805	0.3188590	0.923764
gene98	13.29915	-0.634070	0.768728	-0.824830	0.4094680	0.923764
gene99	82.45653	-0.963147	0.505109	-1.906810	0.0565452	0.799710
gene100	6.25763	1.673078	1.252839	1.335429	0.1817359	0.923764

OUTLINE

- **Experimental Design**
- **General Statistical Concepts**
- **Statistical aspects specific to bulk RNA-seq analysis**

OUTLINE

- **Experimental Design**
- **Statistical Concepts - Bite size statistics**
- **Statistical aspects of bulk RNA-seq analysis**

Let's see if the subject responds to magnetic stimuli... ADMINISTER THE MAGNET!

Interesting...there seems to be a significant decrease in heart rate. The fish must sense the magnetic field.

CONSEQUENCES OF POOR EXPERIMENTAL DESIGN

Inability to answer the questions we would like to answer

- Cost of experimentation.
- Limited & Precious material, esp. clinical samples.
- Immortalization of data sets in public databases and methods in the literature. Our bad science begets more bad science.
- Ethical concerns of experimentation: animals and clinical samples.

A WELL-DESIGNED EXPERIMENT

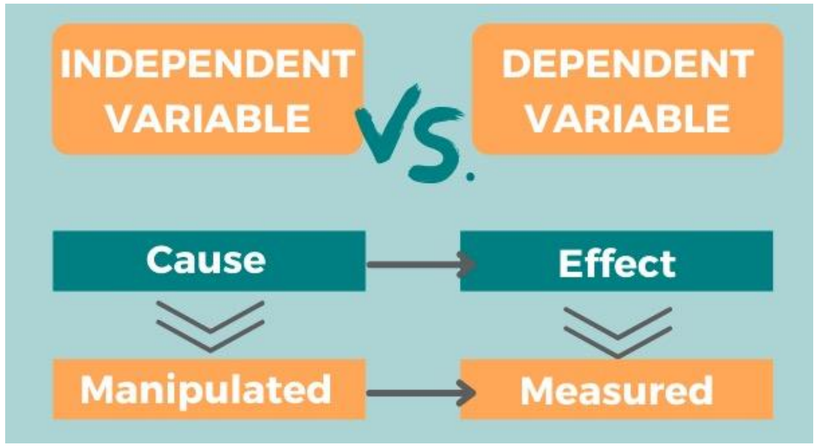
Should have

- Clear objectives
- Focus and simplicity
- Sufficient power
- **Randomised** comparisons

And be

- Precise
- Unbiased
- Amenable to statistical analysis
- Reproducible

VARIABLES IN THE EXPERIMENT

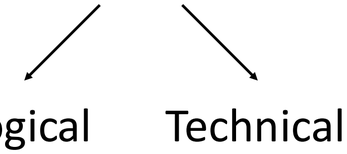


- Independent variable also called as Input or Predictor or Explanatory variable
- Dependent variables also called as output or Response variable

- Based on the type of measurements both Independent and Dependent variables further classified …
  - Continuous: Height, weight, Microarray intensities
  - Discontinuous: RNAseq counts
  - Categorical : Sex, color (Categorical independent variables also called as factors)

SOURCES OF VARIATION

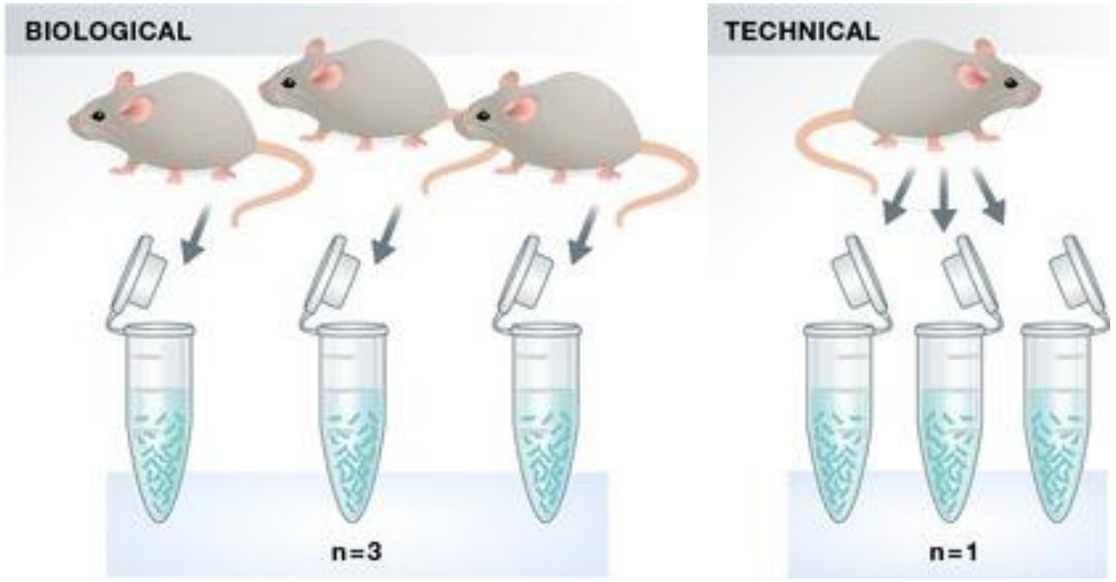
dependent variable = f (independent variable) + noise



- Biological “noise”
  - Biological processes are inherently stochastic
  - Single cells, cell populations, individuals, organs, species….
  - Timepoints, cell cycle, synchronized vs. unsynchronized
- Technical noise
  - Reagents, antibodies, temperatures, pollution
  - Platforms, runs, operators
- Replication is required to capture variance

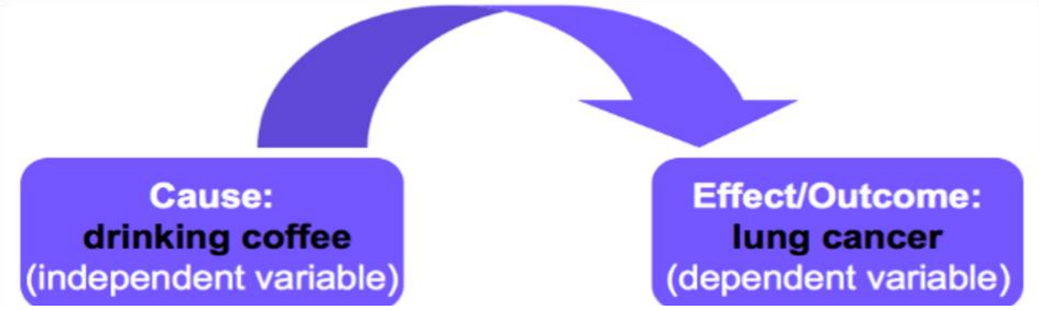
TYPES OF REPLICATION

- Biological replication:
  - In vivo:
    - Patients
    - Mice
  - In vitro:
    - Different cell lines
    - Re-growing cells (passages)
- Technical replication:
  - Experimental protocol
  - Measurement platform (i.e. sequencer)



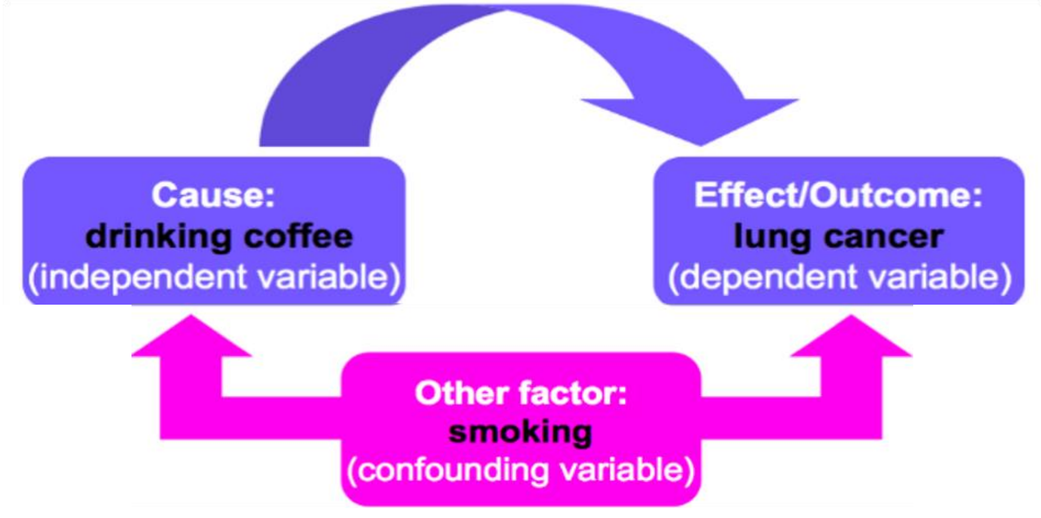
CONFOUNDING FACTORS

- Also known as extraneous, hidden, lurking or masking factors, or the third variable or mediator variable.
- May mask an actual association or falsely demonstrate an apparent association between the independent & dependent variables.
- Hypothetical Example would be a study of coffee drinking and lung cancer.



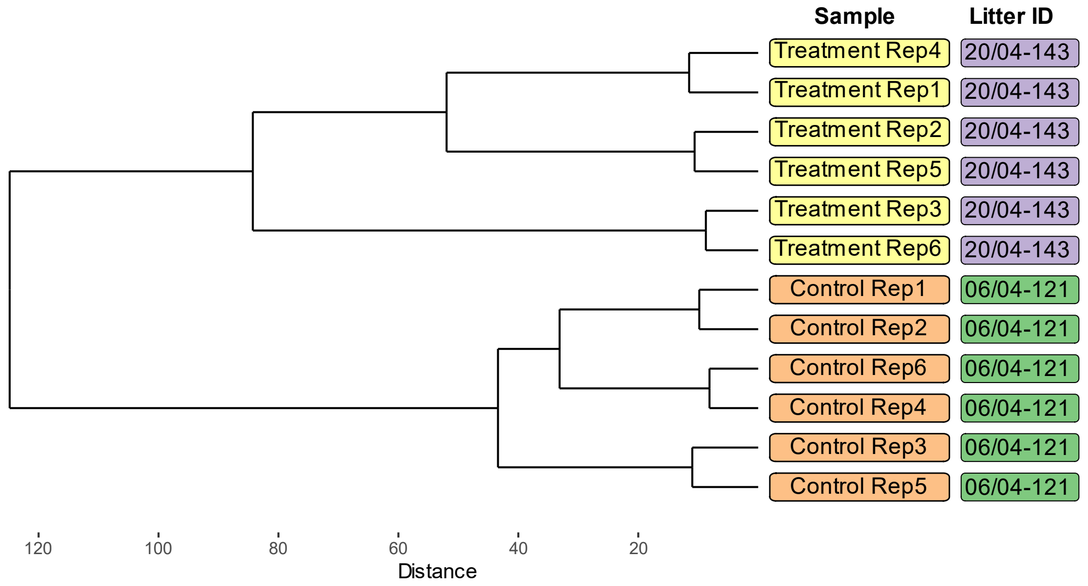
CONFOUNDING FACTORS

- Also known as extraneous, hidden, lurking or masking factors, or the third variable or mediator variable.
- May mask an actual association or falsely demonstrate an apparent association between the independent & dependent variables.
- Hypothetical Example would be a study of coffee drinking and lung cancer.



BATCH EFFECTS

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.

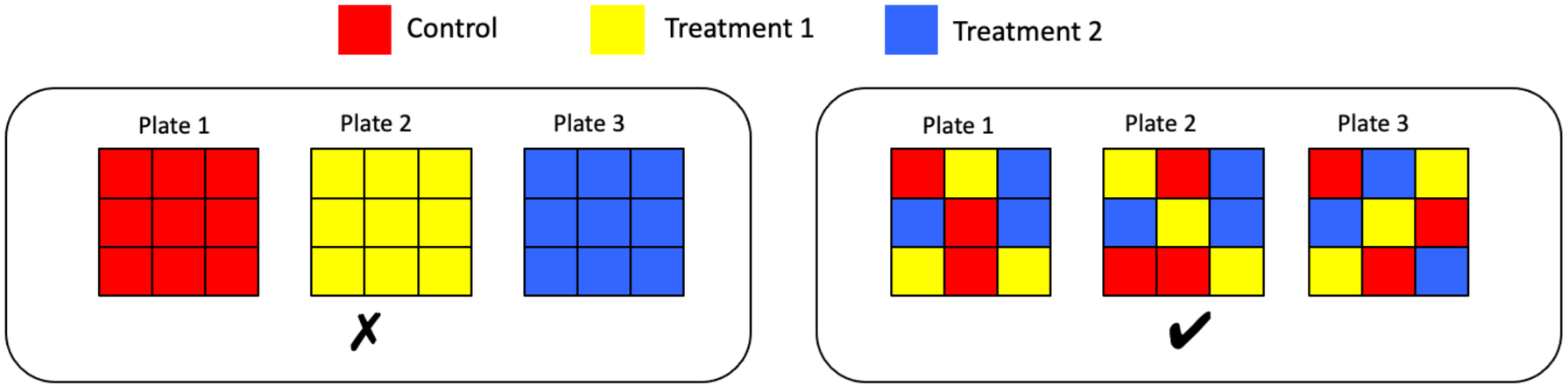


SOLUTIONS

- Write it all down!!!!!!!!
- Controlling technical effects:
  - Randomisation
    - Batch effects that are randomly distributed across experimental variables can be controlled for
    - Statistical analyses assume randomised comparisons
    - May not see issues caused by non-randomised comparisons
    - Make every decision random not arbitrary
    - Caveat: over-randomization can increase error
  - Blinding
    - Especially important where subjective measurements are taken
    - Potentially multiple degrees of blinding (eg. double-blinding)

RANDOMISED BLOCK DESIGN

- Blocking is the arranging of experimental units in groups (blocks) that are similar to one another.

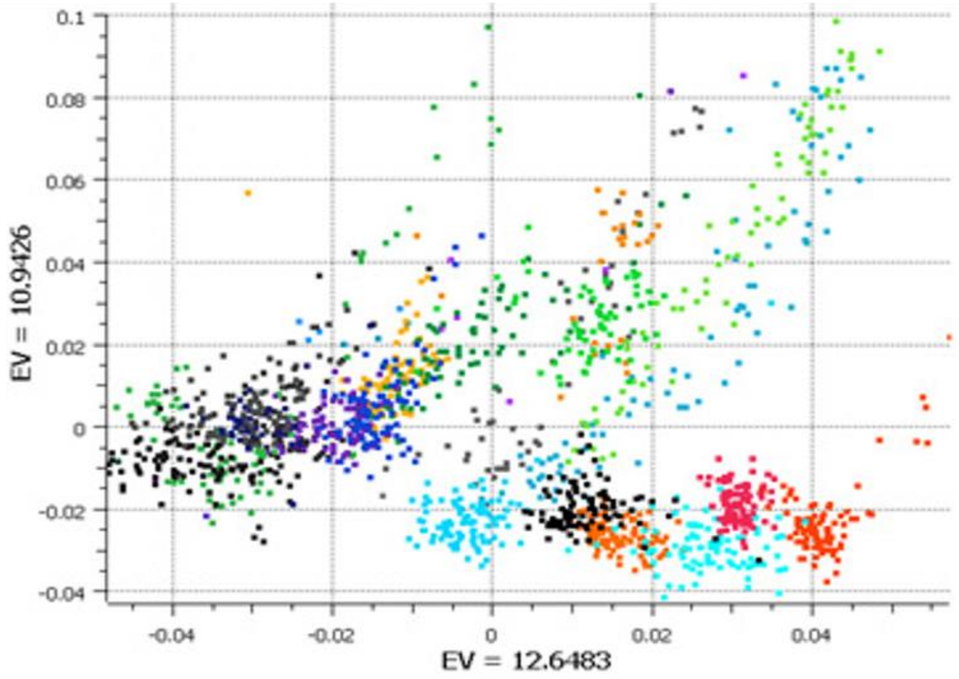


- Each plate contains spatially randomised equal proportions of:
  - Control
  - Treatment 1
  - Treatment 2
- controlling plate effects.

RANDOMISED BLOCK DESIGN

- Good design example: Alzheimer's study from GlaxoSmithKline

Plate effects by plate
Left PCA plot show large plate effects.
Each colour corresponds to a different plate



http://blog.goldenhelix.com/?p=322
c

RANDOMISED BLOCK DESIGN

- Good design example: Alzheimer's study from GlaxoSmithKline

Plate effects by plate
Left PCA plot show large plate effects.
Each colour corresponds to a different plate

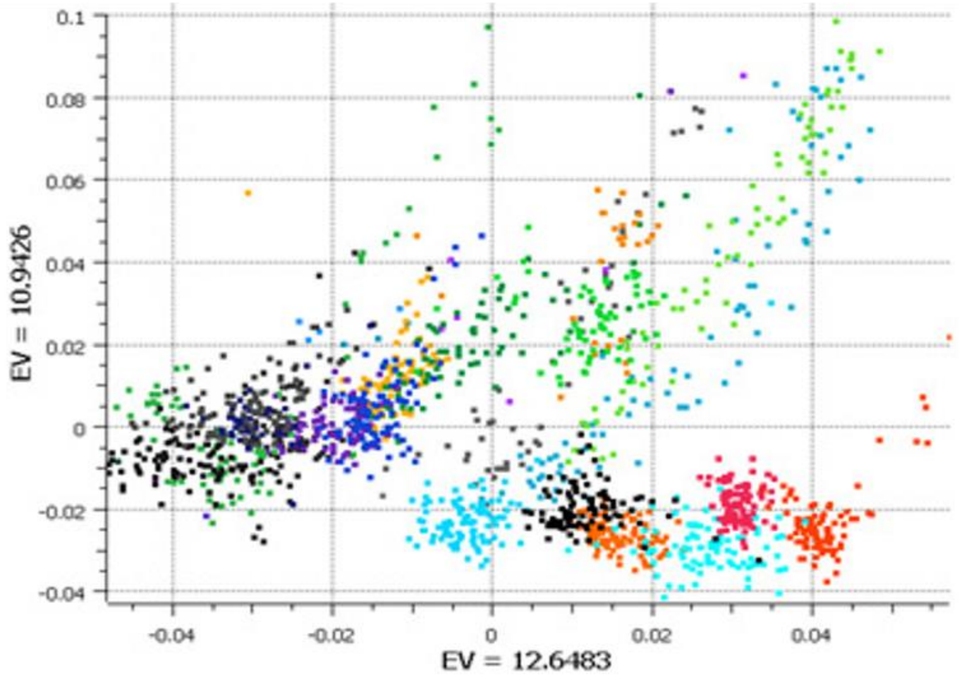
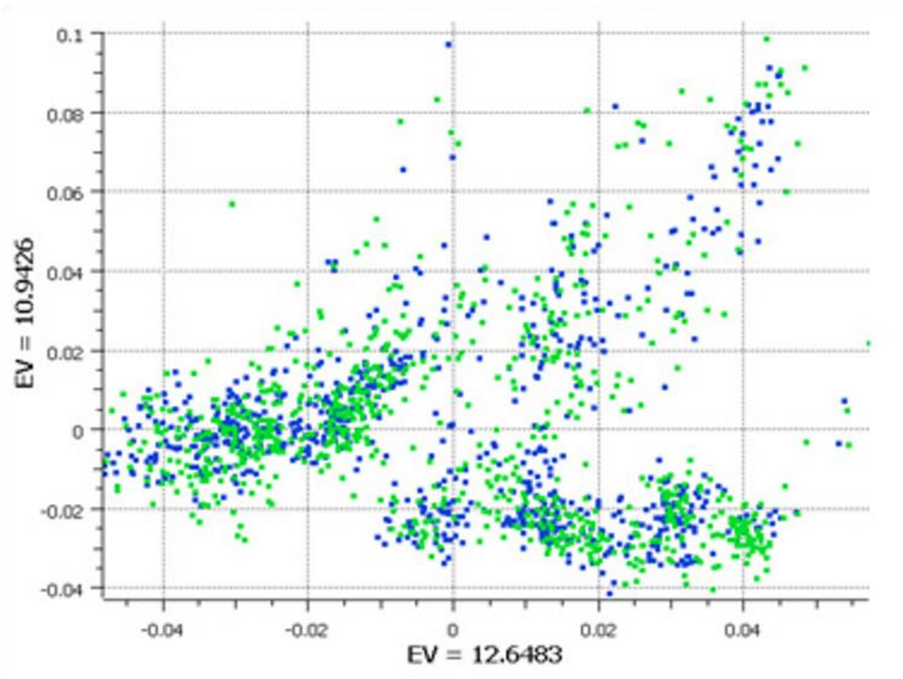


Plate effects by case/control
Right PCA plot shows each plate cluster contains equal proportions of cases (blue) and controls (green).



http://blog.goldenhelix.com/?p=322
c

EXPERIMENTAL CONTROLS

- Ideal : Everything is identical across conditions except the variable you are testing
- Controlling errors
  - Type I: False Positives
    - Negative controls: should have minimal or no effect
  - Type II: False Negatives
    - Positive controls: known effect
- Technical controls
  - Detect/correct technical biases
  - Normalise measurements (quantification)

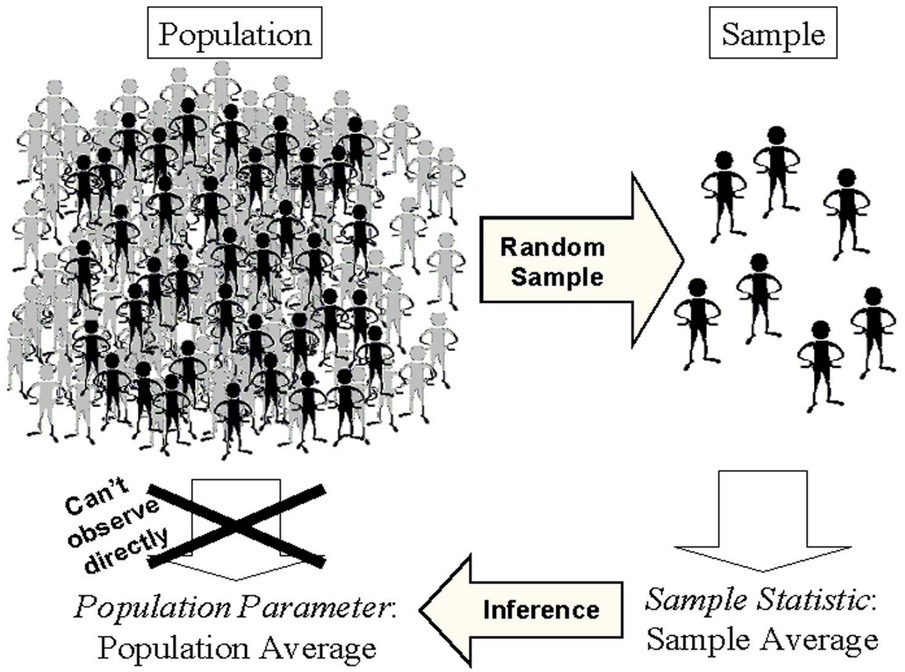
EXAMPLES OF EXPERIMENTAL CONTROLS

- Wild-type organism (knockouts)
- Inactive siRNA (silencing)
- Vehicle (treatments)
- Spike-ins (quantification/normalisation)
- “Gold standard” data points
- Multi-level controls
- e.g. contrast Vehicle/Input vs. Treatment/Input

OUTLINE

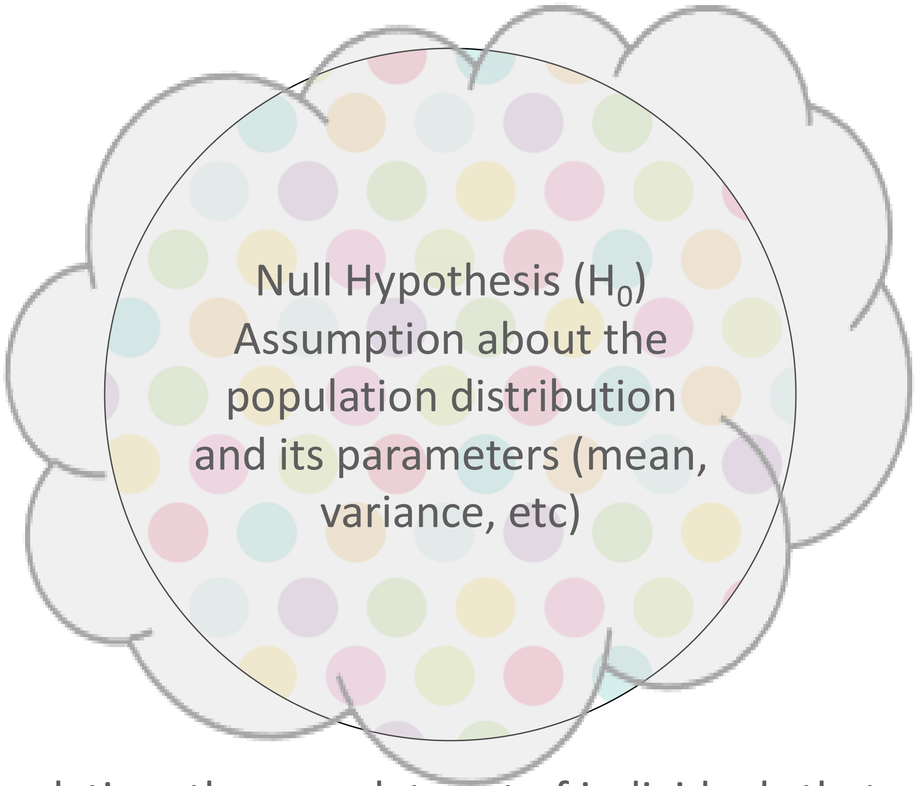
- Experimental Design
- **Statistical Concepts**
- Statistical aspects of bulk RNA-seq analysis

BASICS ON INFERENTIAL STATISTICS AND HYPOTHESIS TESTING



- Two important parameters
  - Mean
  - Variance
- Population mean and variance unknow and are constants
- Estimated using sample
- Estimated mean and variance used for inferring population parameters

BASICS ON INFERENTIAL STATISTICS AND HYPOTHESIS TESTING



Inference means two things:

1. Estimating population parameters
2. Testing hypothesis regarding the population distribution

inference

Population: the complete set of individuals that we are interested in

Sample: smaller set of individuals that is representative of the population

Variable: what we are interested in measuring

Slide content adapted from Matt Castle

BASICS ON INFERENTIAL STATISTICS AND HYPOTHESIS TESTING

Null Hypothesis (H_0)
Assumption about the population distribution and its parameters (mean, variance, etc)

Choose test

Based on data and H_0

Calculate test statistic, interpret results

Slide content adapted from Matt Castle

A SIMPLE EXAMPLE

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: Drug has no effect on response time

H_1: Drug has an effect on response time

Example adapted from Sal Khan

A SIMPLE EXAMPLE

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Example adapted from Sal Khan

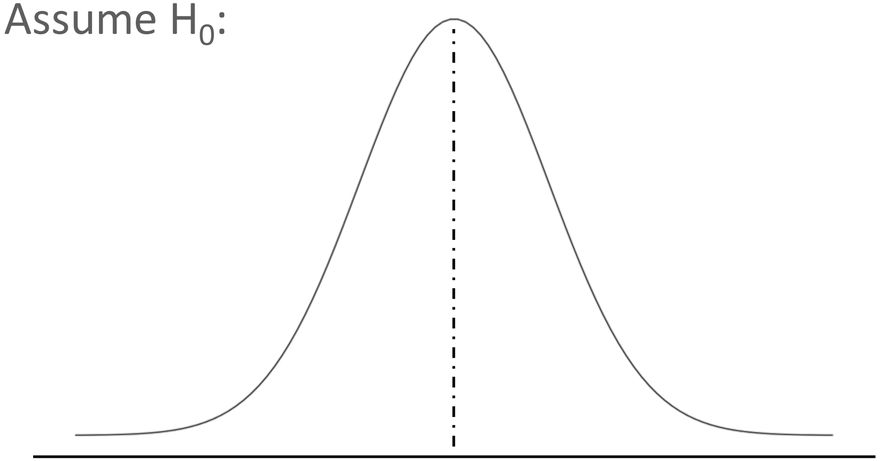
A SIMPLE EXAMPLE

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Assume H_0:



$\mu = 1.2$ s

Example adapted from Sal Khan

A SIMPLE EXAMPLE

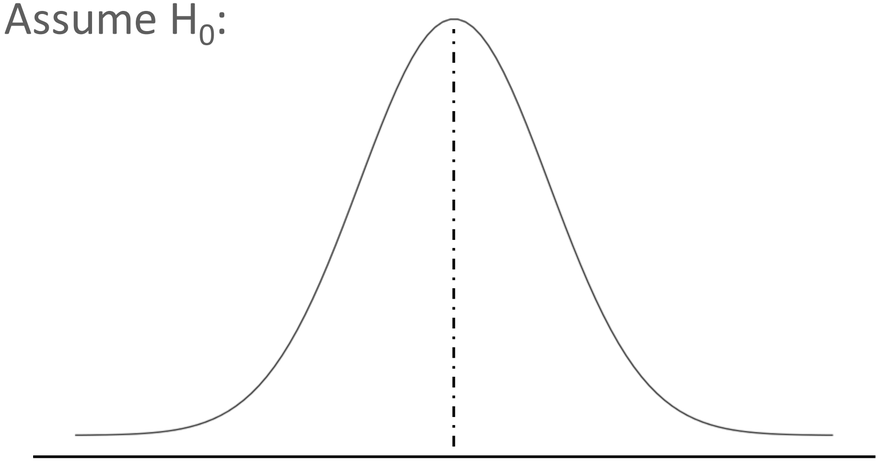
A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}}$

Assume H_0:



$\mu = 1.2$ s

Example adapted from Sal Khan

A SIMPLE EXAMPLE

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

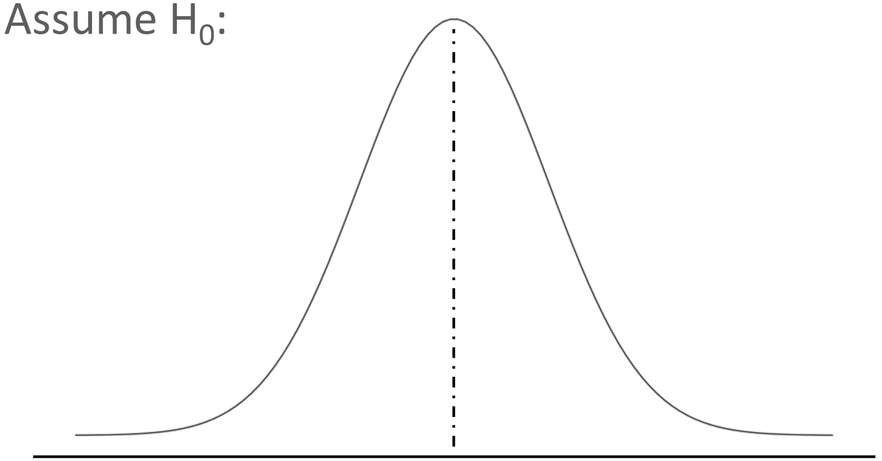
H_1: $\mu \neq 1.2$ s

Calculate test statistic

1.05 → m, 1.2 → μ, 0.5 → s, 100 → n

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

Assume H_0:



$\mu = 1.2$ s

Example adapted from Sal Khan

A SIMPLE EXAMPLE

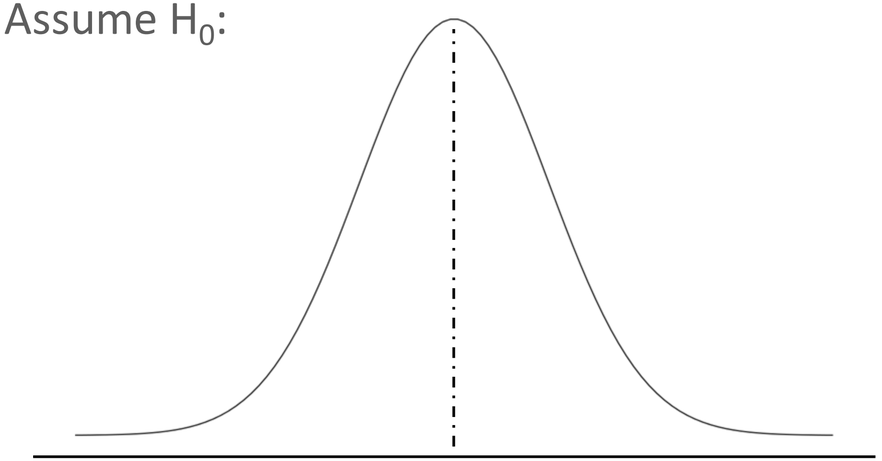
A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}}$ = - 3

Assume H_0:



$\mu = 1.2$ s

Example adapted from Sal Khan

A SIMPLE EXAMPLE

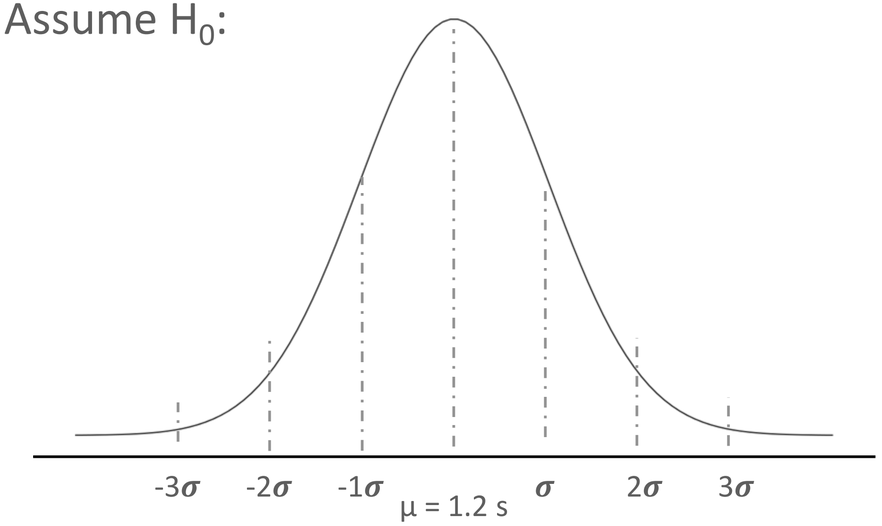
A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}}$ = - 3

Assume H_0:



This means that the sample mean (1.05) is 3 standard deviations away from the mean

Example adapted from Sal Khan

A SIMPLE EXAMPLE

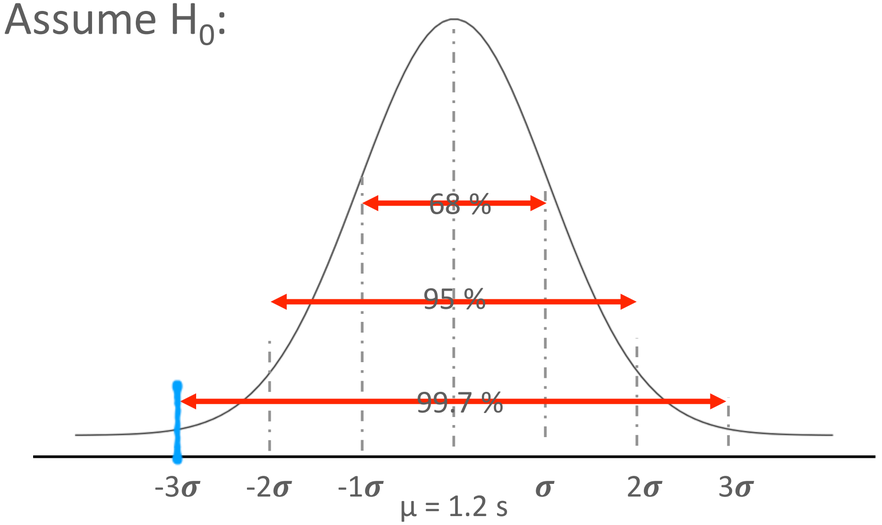
A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}} = -3$

Assume H_0:



This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

Example adapted from Sal Khan

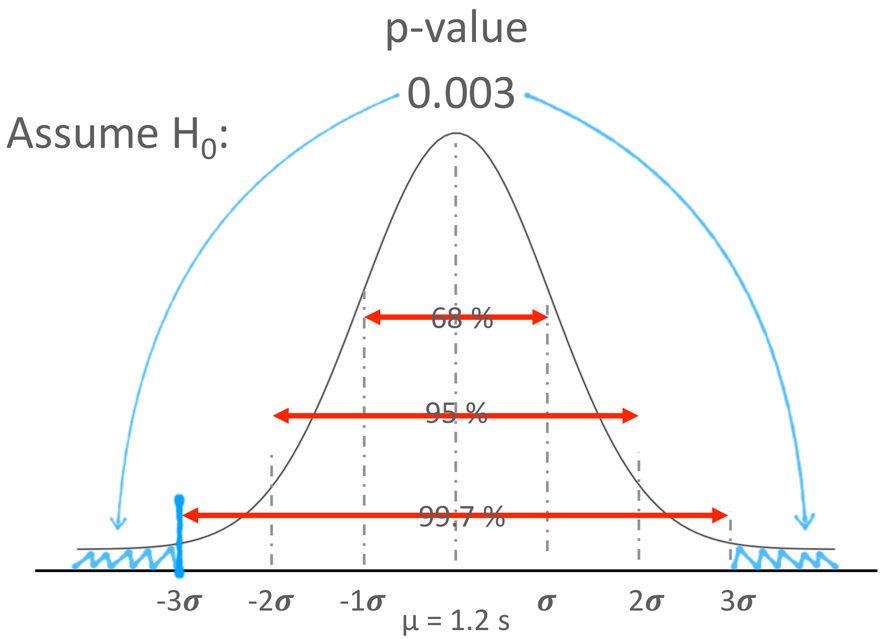
A SIMPLE EXAMPLE

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}} = -3$



This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

Example adapted from Sal Khan

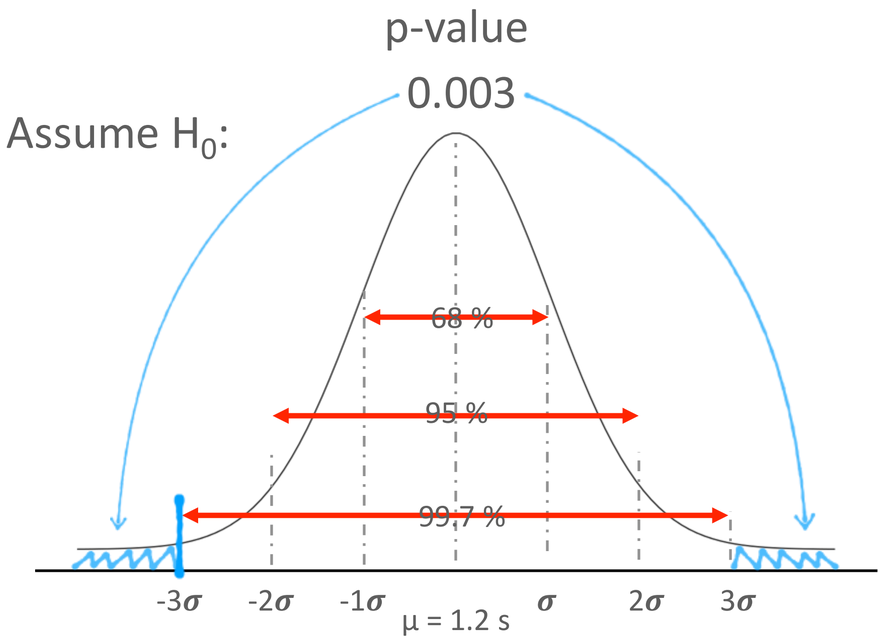
A SIMPLE EXAMPLE

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}} = -3$



This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

We reject the null hypothesis!

Example adapted from Sal Khan

A SIMPLE EXAMPLE

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

Constructed the null and alternative hypothesis about the population

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

Calculate test statistic

$$t = \frac{m - \mu}{s/\sqrt{n}} = -3$$

Calculated test statistic

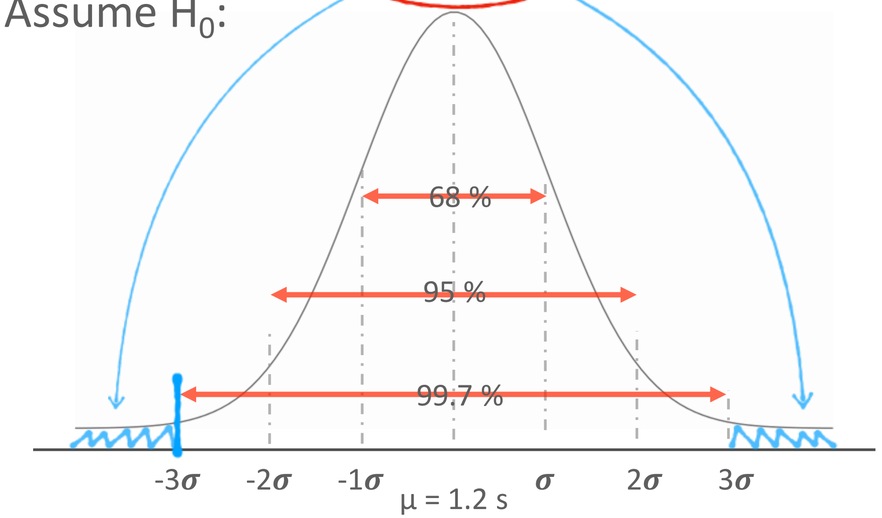
This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

Obtained p-value

p-value
0.003

Assume H_0:



Reached a conclusion

We reject the null hypothesis!

Example adapted from Sal Khan

KEY CONCEPTS - HYPOTHESIS TESTING

- **All statistical tests are based on assumptions!**
- **All statistics can be wrong**
- **Statistical tests are probabilistic in nature**
- **There is always a chance that the result is wrong (even when all assumptions met perfectly):**
  - Either significant result when no difference (Type I),
  - Or insignificant results when there is an actual difference (Type II)

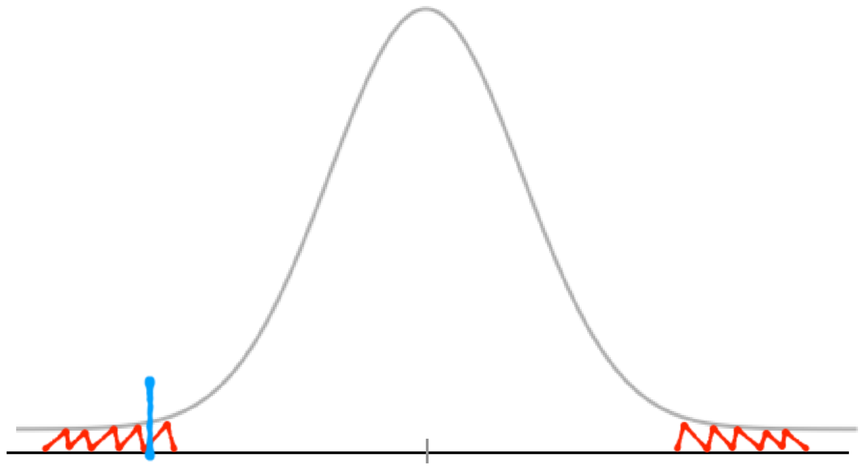
Slide content adapted from Matt Castle

TYPE I AND TYPE II ERRORS

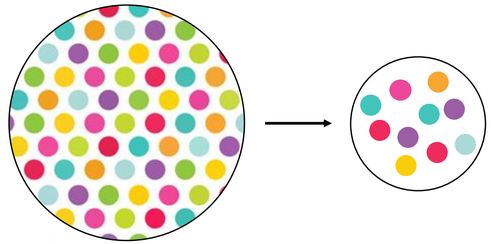
H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

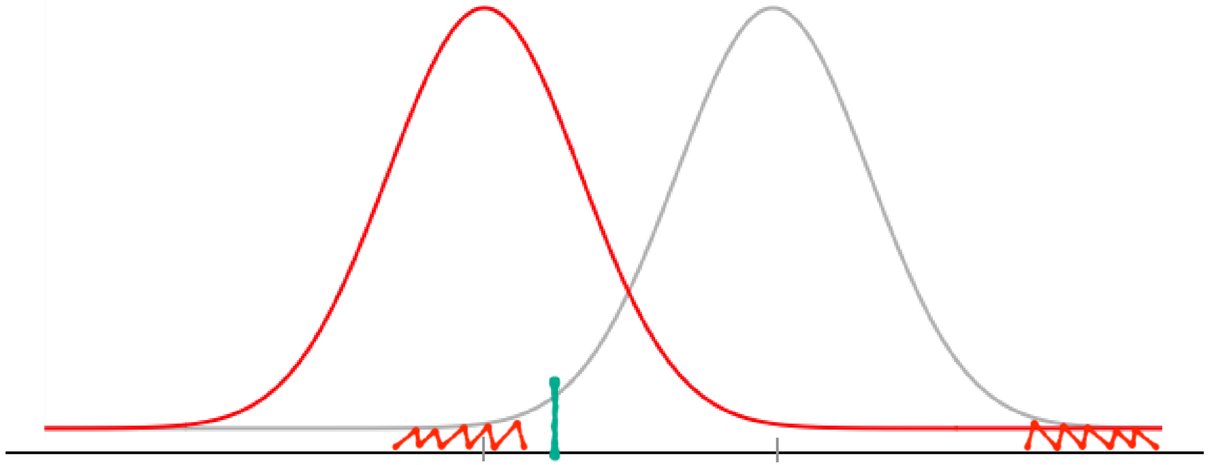
α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



Suppose H_1 true:



Depending on your sampling, you might fail to reject H_0

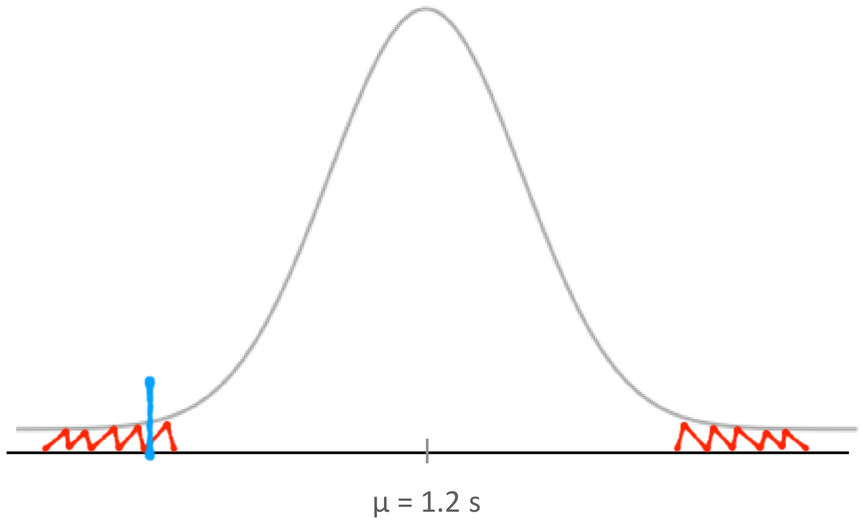


TYPE I AND TYPE II ERRORS

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

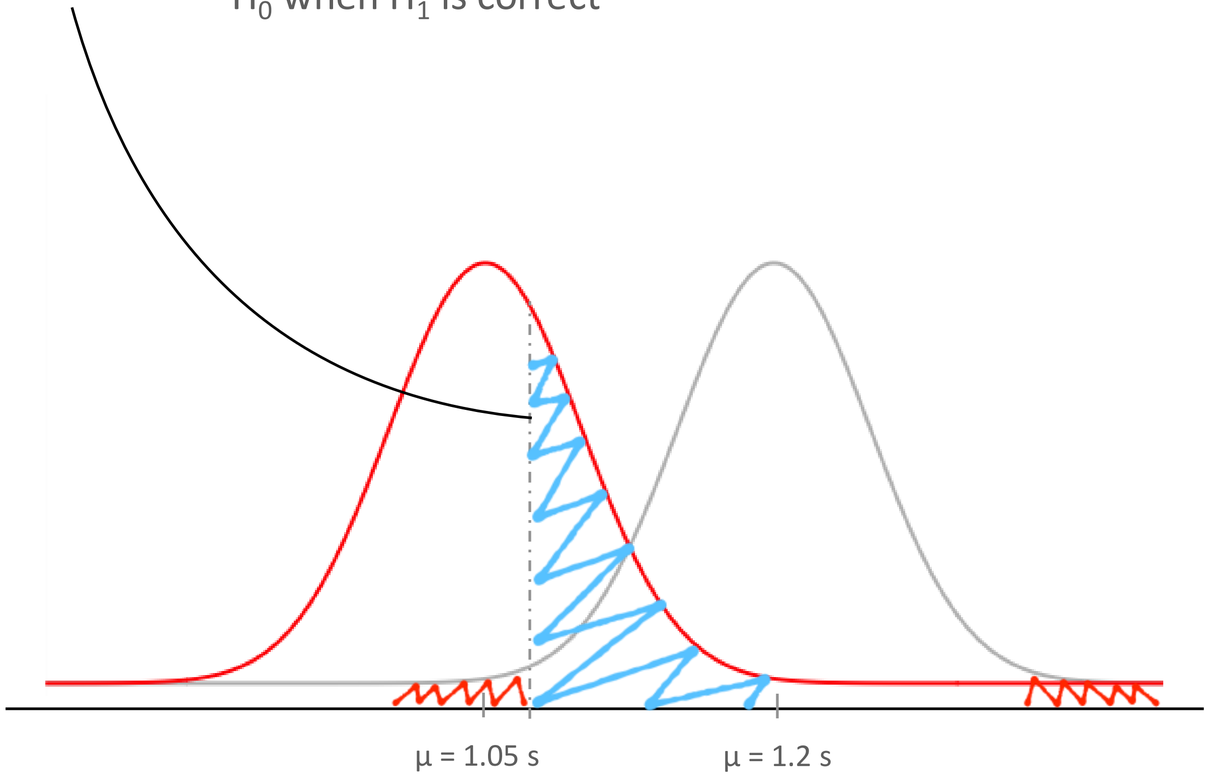
if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



Suppose H_1 true:

β → the type II error, the probability of not rejecting H_0 when H_1 is correct

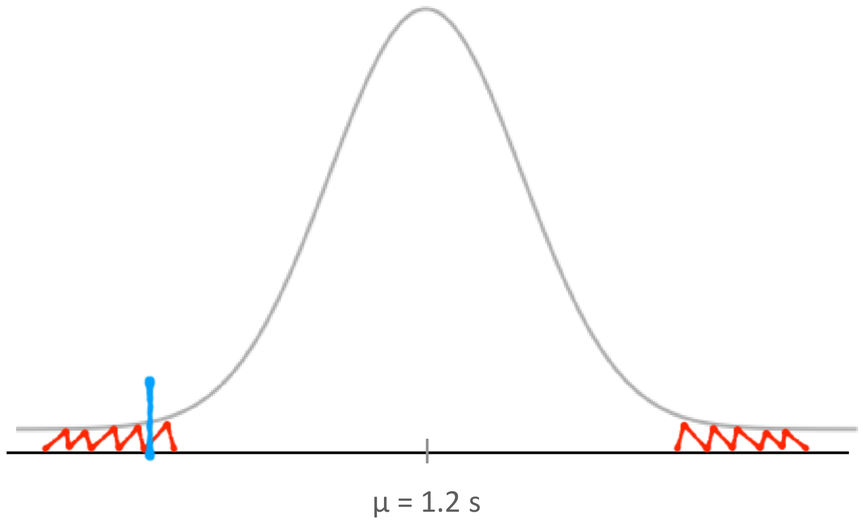


TYPE I AND TYPE II ERRORS

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

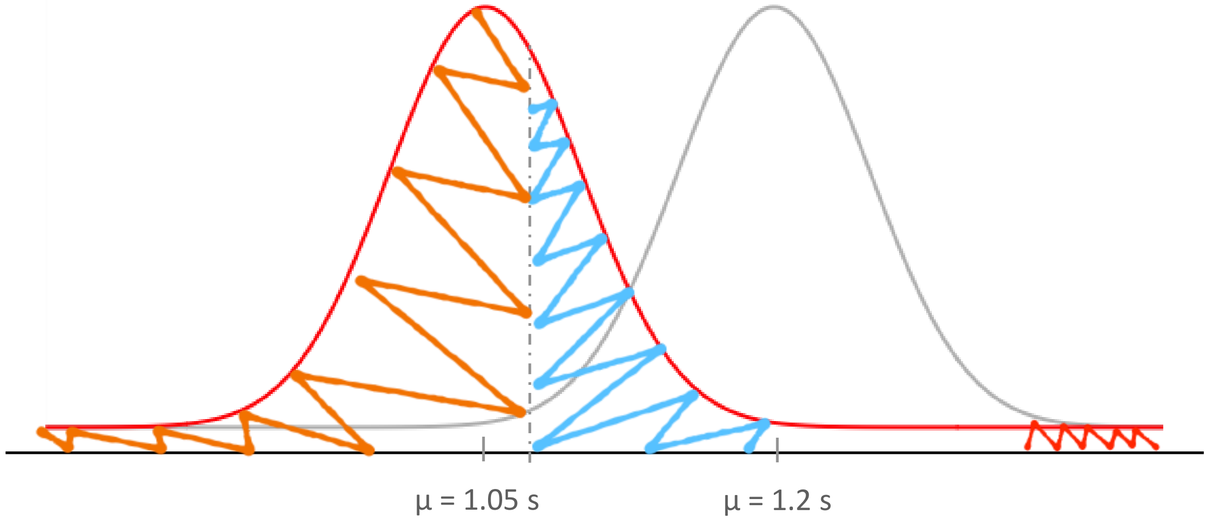
α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



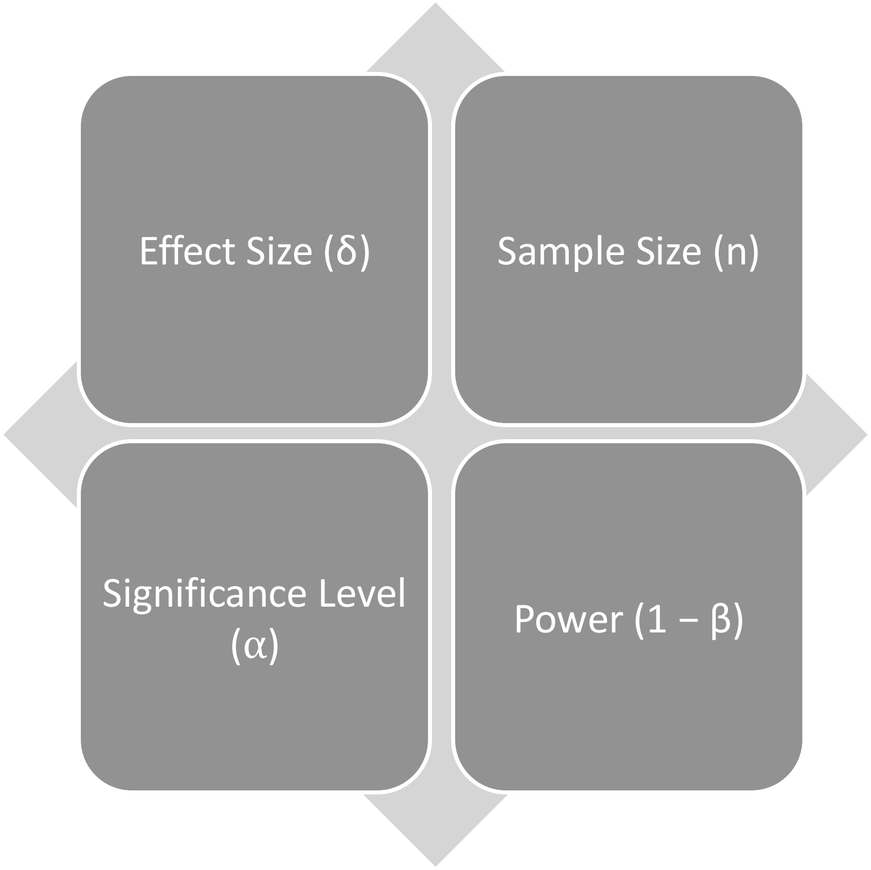
Suppose H_1 true:

β → the type II error, the probability of not rejecting H_0 when H_1 is correct

1- β → Power is the probability that we actually detect an effect that exists

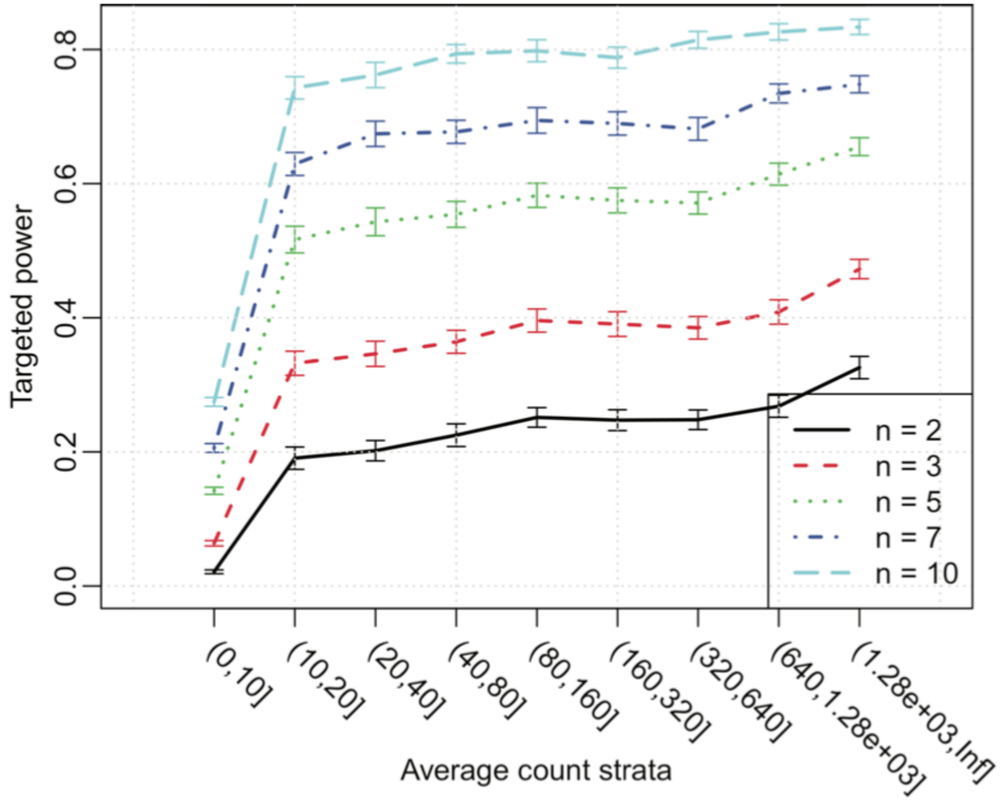


POWER ANALYSIS



- The four concepts are linked
- If we know three, we can work out the forth
- Power calculation: Aim is to define the probability (1- β) to detect an effect size of interest (δ) at the α level with a sample size of *n* biological replicates
- Sample size calculation: Aim is to define the sample size (*n*) allowing to detect an effect size of interest (δ) at the α level with a given probability (1 – β).

POWER ANALYSIS IN DIFFERENTIAL EXPRESSION ANALYSIS

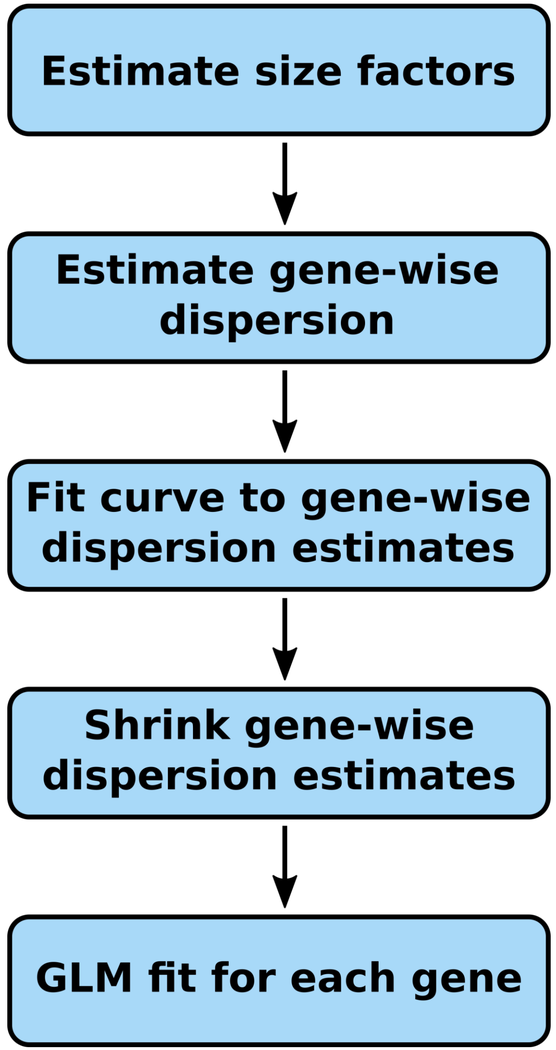


(Wu, Wang and Wu (2015))

OUTLINE

- Experimental Design
- Statistical Concepts - Bite size statistics
- **Statistical aspects of bulk RNA-seq analysis**

DESEQ2 WORKFLOW



NORMALISATION

- Counting estimates the **relative** counts for each gene
- Does this ***accurately*** represent the original population of RNAs?
- The relationship between counts and RNA expression is not the same for all genes across all samples

- Library Size – different sequence depth between samples
- Gene properties - Length, GC content, sequence
- Library composition - Quantification is relative, changes in relative abundance for one gene will affect the relative abundances of other genes. “Composition Bias”

GENERAL PRINCIPLES BEHIND NORMALISATION

Normalization has two steps

- Scaling
  - First get size factors or normalization factors
  - Usually one size factor per sample
  - Scale the counts by divide the raw counts of a sample with sample specific size factor
- Transformation: Transform the data after scaling
  - Per million
  - log2
  - square root transformation
  - Pearson residuals (eg. sctransform)

Normalization removes technical variance but not biological variance

Normalization helps in making two samples comparable

Raw data

Gene Name	Sample1	Sample2
Gene 1	1000	600
Gene 2	8	2
Gene 3	555	470
Size Factors	1563	1072

Scaling

$$\frac{\text{Raw count}}{\text{Cell size factor}}$$

Scaled data

Gene Name	Sample1	Sample2
Gene 1	0.639	0.559
Gene 2	0.005	0.001
Gene 3	0.355	0.438

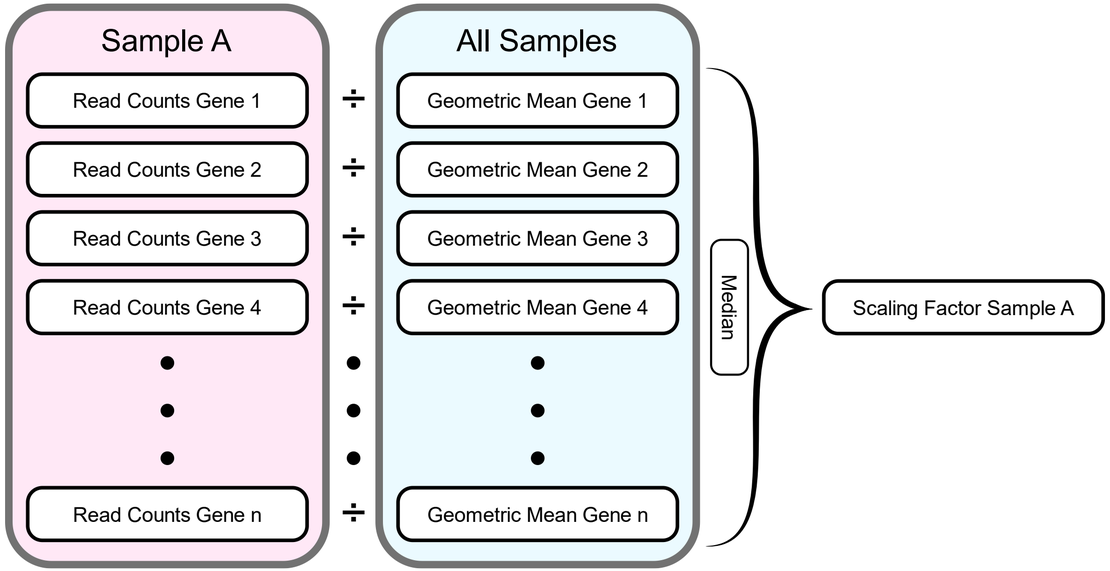
Transformation Scaled value x 1000,000

Gene Name	Sample1	Sample2
Gene 1	639000	559000
Gene 2	5000	1000
Gene 3	355000	438000
Total counts	999,000	998,000

Normalized data

DESEQ2 NORMALISATION

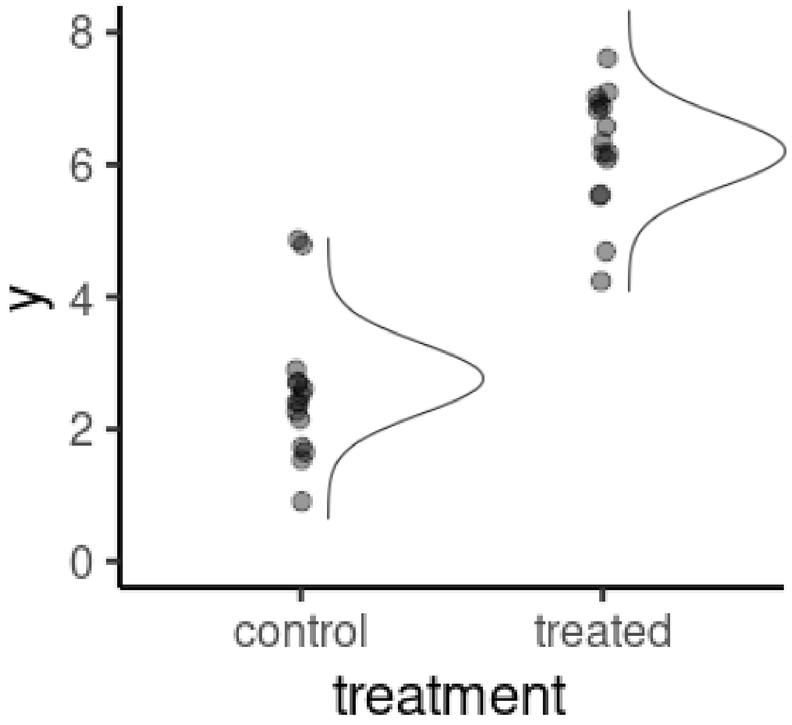
1. Geometric mean is calculated for each gene across all samples.
2. The counts for a gene in each sample is then divided by this mean.
3. The median of these ratios in a sample is the size factor (normalization factor) for that sample.
4. DESEq2 normalization corrects for library size and RNA composition bias
5. Composition bias: Arises for examples when only a small number of genes are very highly expressed in one sample but not in the other.



STATISTICAL ASPECTS OF DIFFERENTIAL EXPRESSION ANALYSIS

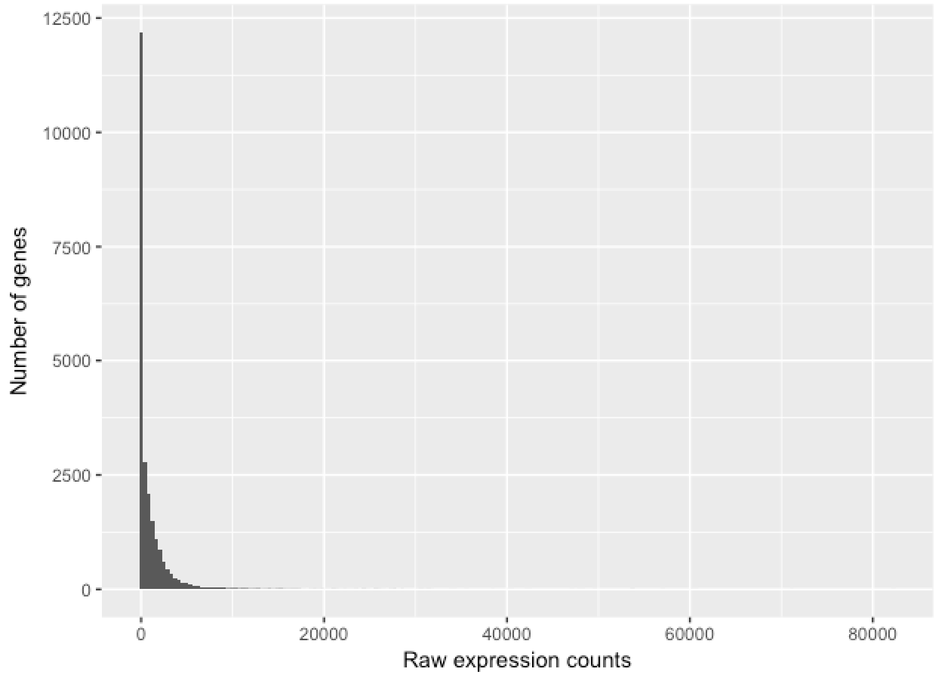
Linear Modeling

Model the expression of each gene as linear combination of explanatory factors (eg. treatment, age, sex, etc.)



STATISTICAL ASPECTS OF DIFFERENTIAL EXPRESSION ANALYSIS

Characteristics of RNA-seq data



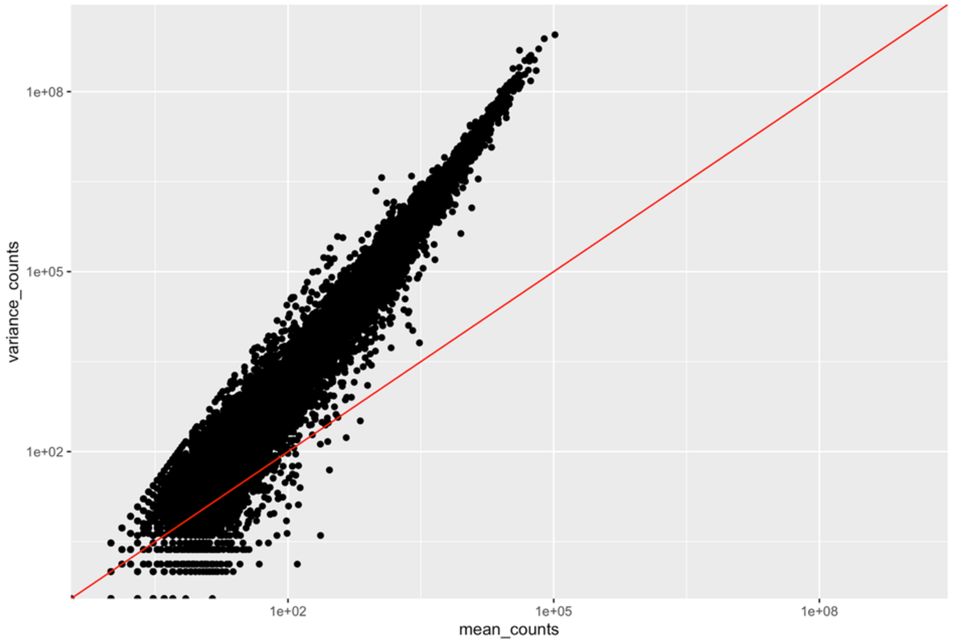
This plot illustrates some common features of RNA-seq count data:

- a low number of counts associated with a large proportion of genes
- a long right tail due to the lack of any upper limit for expression
- large dynamic range

Looking at the shape of the histogram, we see that it is not normally distributed.

STATISTICAL ASPECTS OF DIFFERENTIAL EXPRESSION ANALYSIS

Characteristics of RNA-seq data



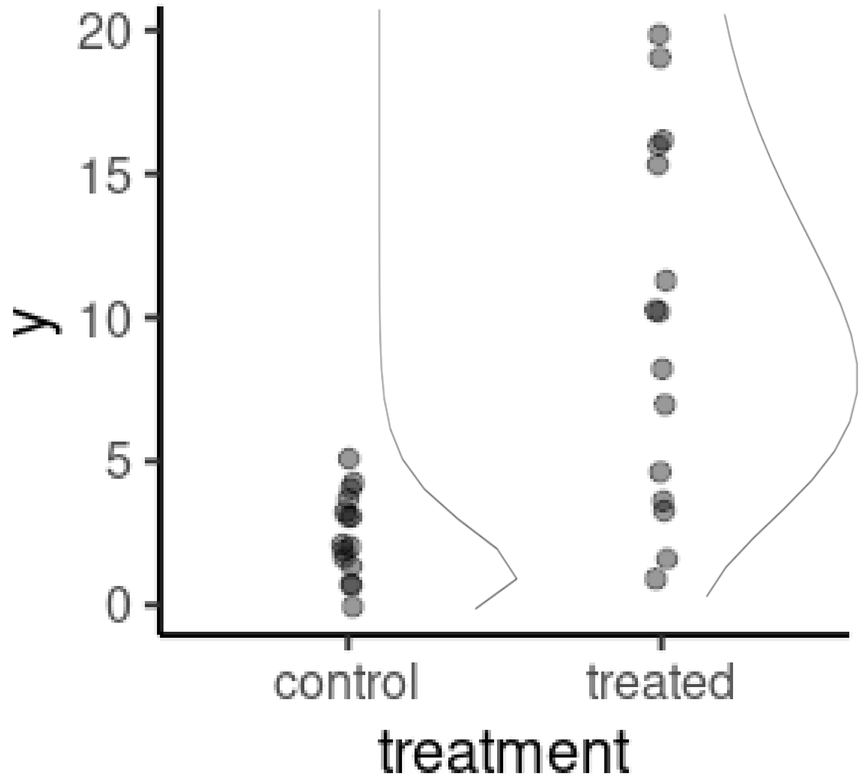
To assess the properties of the data we are working with, we can look at the mean-variance relationship.

For the genes with high mean expression, the variance across replicates tends to be greater than the mean (scatter is above the red line).

Essentially, the Negative Binomial is a good approximation for data where the mean < variance, as is the case with RNA-Seq count data.

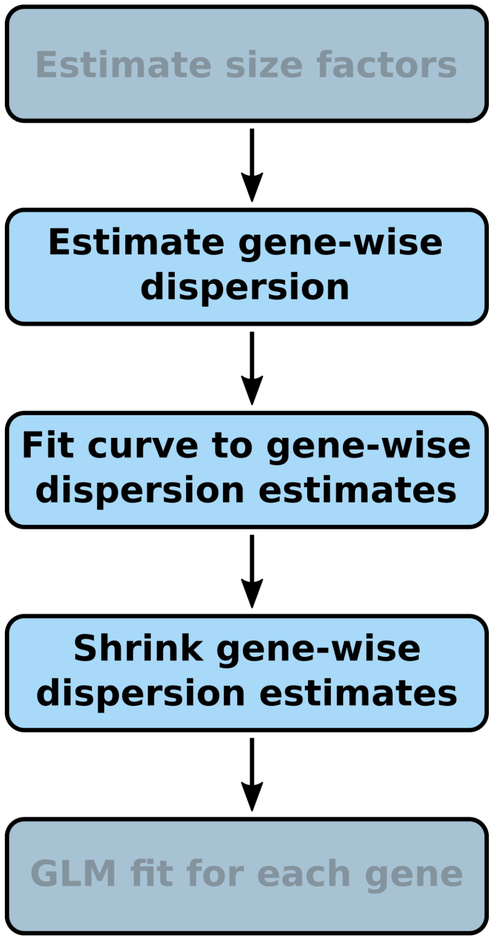
STATISTICAL ASPECTS OF DIFFERENTIAL EXPRESSION ANALYSIS

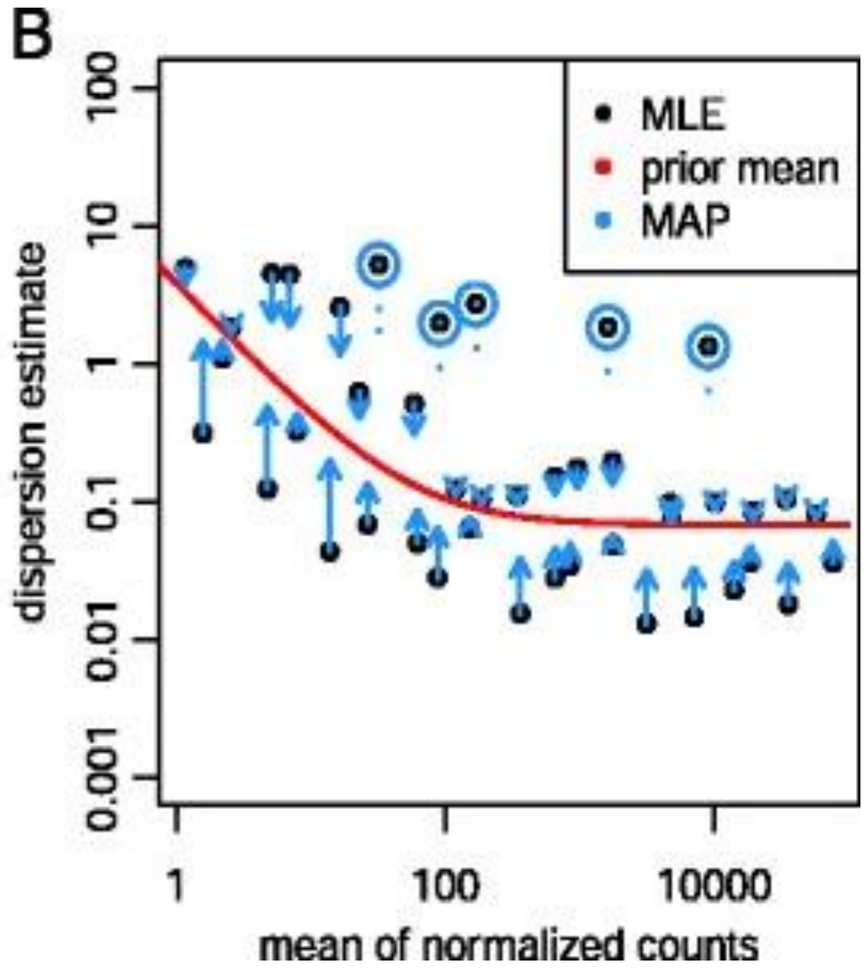
Model the expression of each gene as linear combination of explanatory factors (eg. treatment, age, sex, etc.)



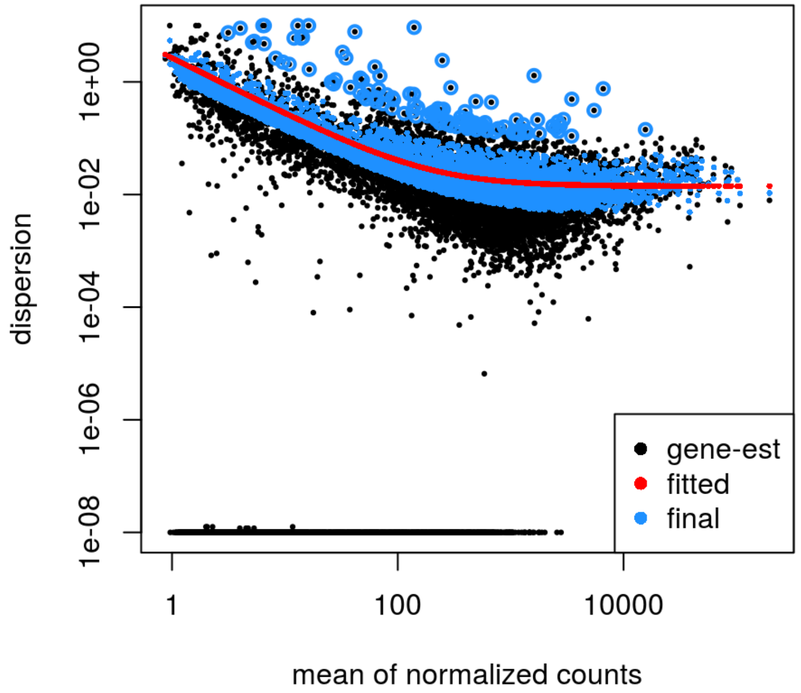
STATISTICAL ASPECTS OF DIFFERENTIAL EXPRESSION ANALYSIS

How we estimate dispersion in DESEQ2

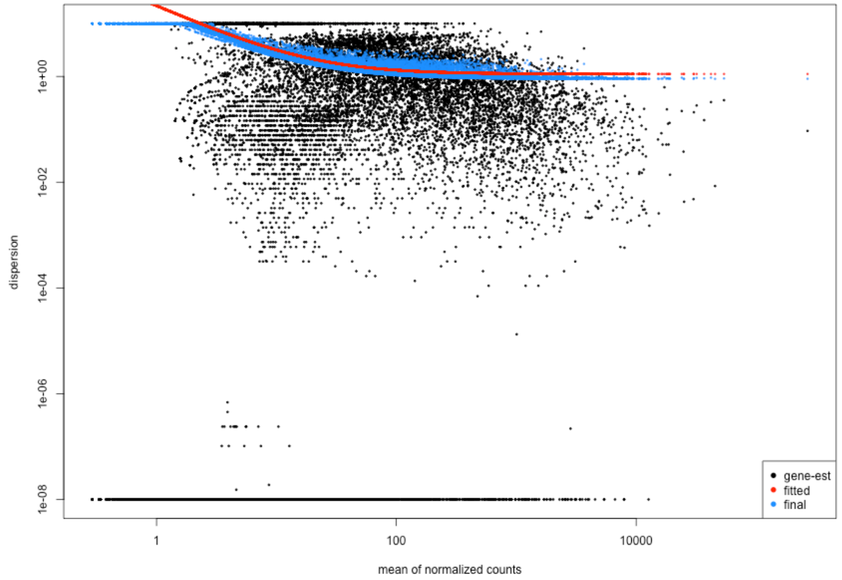


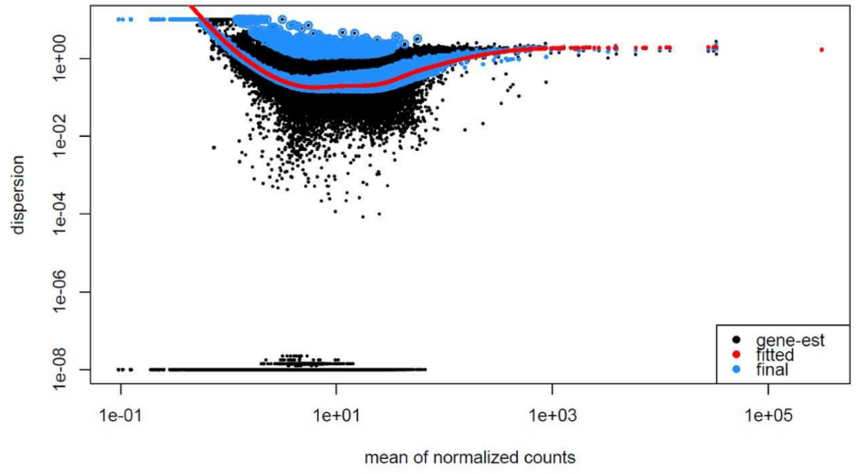


Good:


dispersion
1e+00
1e-02
1e-04
1e-06
1e-08
1
100
10000
mean of normalized counts
gene-est
fitted
final

Bad:


dispersion
1e+00
1e-02
1e-04
1e-06
1e-08
1
100
10000
mean of normalized counts
gene-est
fitted
final


dispersion
1e+00
1e-02
1e-04
1e-06
1e-08
1e-01
1e+01
1e+03
1e+05
mean of normalized counts
gene-est
fitted
final

STATISTICAL ASPECTS OF DIFFERENTIAL EXPRESSION ANALYSIS

$$Counts \sim NB(\mu, \phi)$$

$$\mu = sq$$

$$log2(q) = \beta_0 + \beta_1 * treatment + \beta_2 * age + \ldots$$

counts - expression of the gene

$\boldsymbol{\beta}_i$ - parameters we want to estimate from the data

$\boldsymbol{\beta}_0$ - the “intercept” (the value of expression when all other parameters are set at a reference level)

ϕ - the “dispersion” (uncertainty) of our model (also estimated from the data)

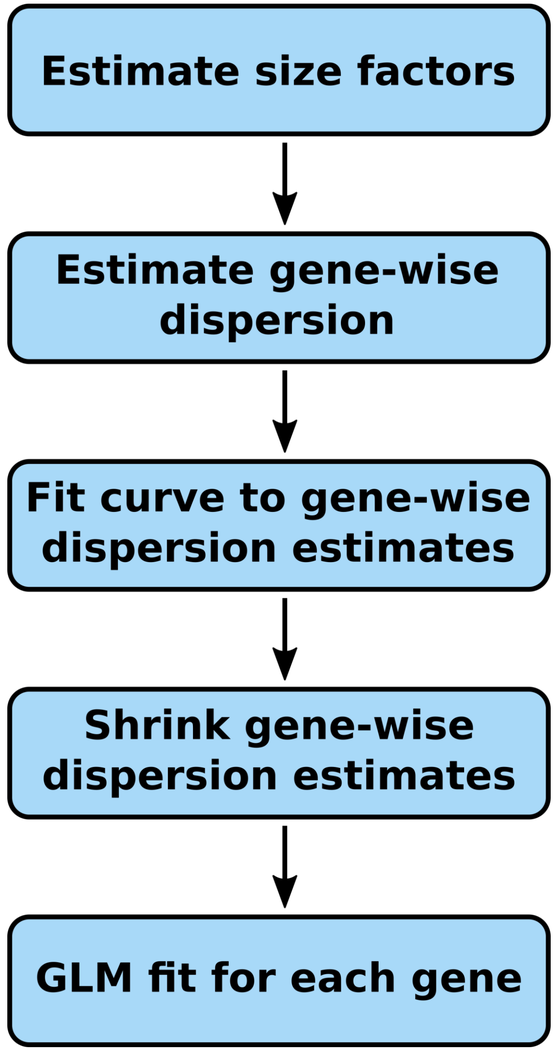
s - scaling factor (sequencing depth and transcript composition)

Summary:

- Use **negative binomial linear regression** to model gene expression in RNA-seq
- Calculate **size factors** for each sample to account for differences in sequencing depth and transcript composition between samples
- Estimate **dispersion** for each gene by “borrowing” information across genes for more precise estimates when sample sizes are small (as is typical in RNA-seq experiments)
- Estimate model **coefficients** which are used to define test hypothesis **($\boldsymbol{\beta}_i = 0$)**

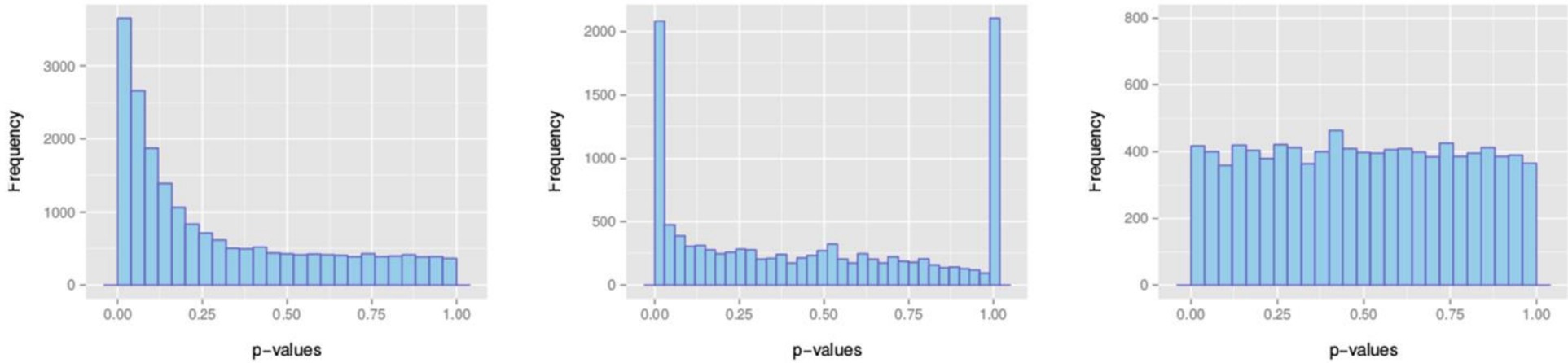
Description of DESeq2 model: Love, Huber & Anders (2014)

DESEQ2 WORKFLOW



P-VALUE HISTOGRAMS

Examples of expected overall distribution



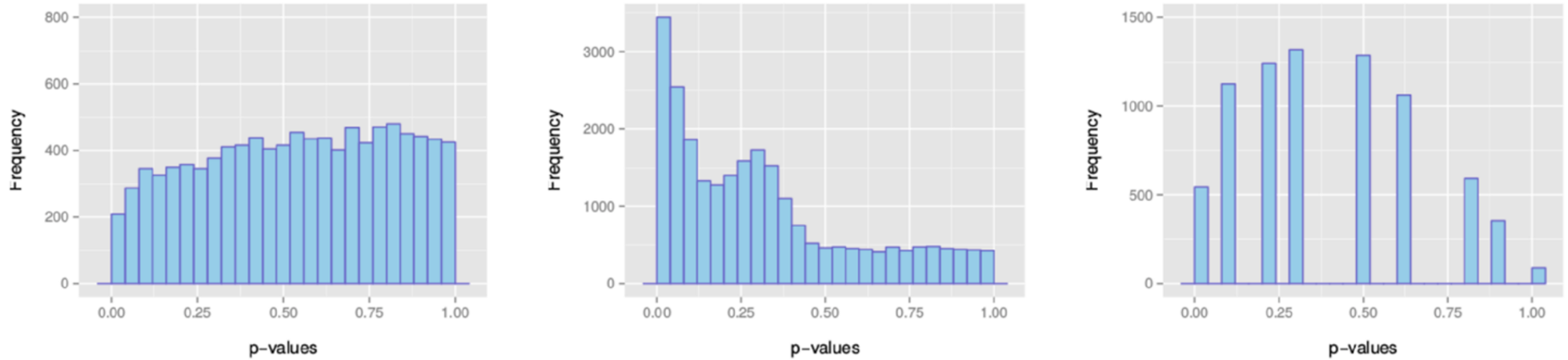
(a) : the most desirable shape

(b) : very low counts genes usually have large p-values

(c) : do not expect positive tests after correction

P-VALUE HISTOGRAMS

Examples of unexpected overall distribution



(a) : indicates a batch effect (confounding hidden variables)

(b) : the test statistics may be inappropriate (due to strong correlation structure for instance)

(c) : discrete distribution of p-values : unexpected

MULTIPLICITY CORRECTION

- A gene with a significance cut-off of $\alpha = 0.05$, means there is a 5% chance it is a false positive.
- If we test for 20,000 genes for differential expression at $\alpha = 0.05$, we would expect to find 1,000 genes by chance
- If we found 3000 genes to be differentially expressed total, roughly one third of our genes are false positives!
- The more genes we test, the more we inflate the false positive rate. This is the multiple testing problem.

MULTIPLICITY CORRECTION

- Bonferroni: The adjusted p-value is calculated by: α^{k} (k = total number of tests). This is a very conservative approach

- FDR/Benjamini-Hochberg: Benjamini and Hochberg (1995) defined the concept of FDR and created an algorithm to control the expected FDR below a specified level given a list of independent p-values.

CONCLUSIONS

- Assumptions assumptions assumptions