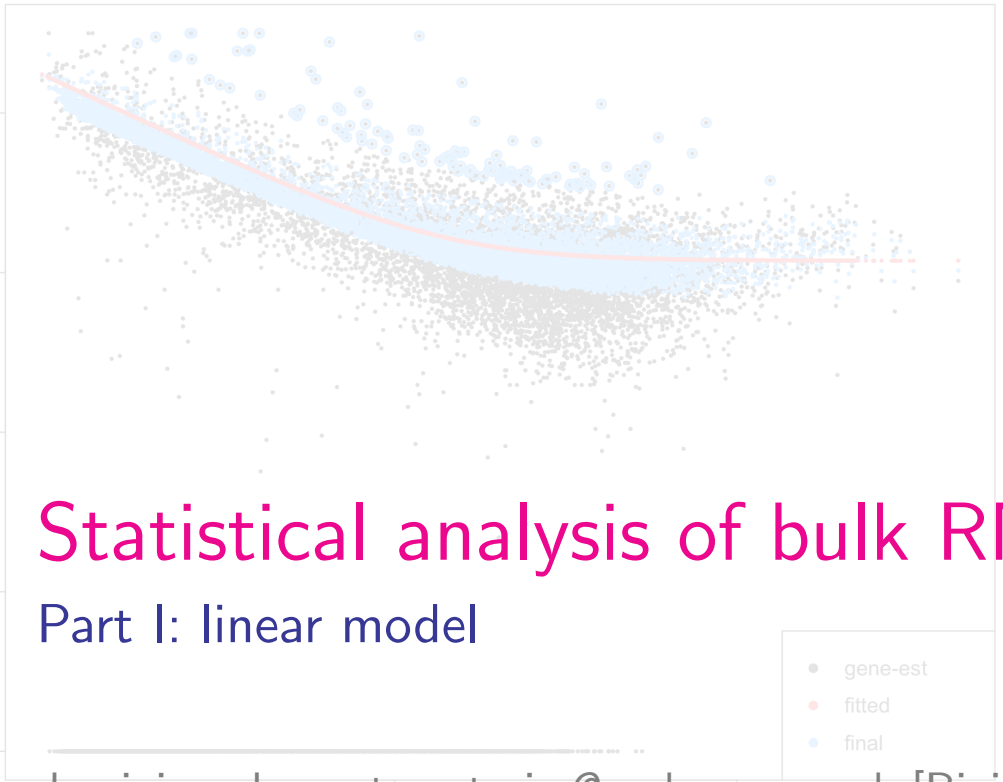


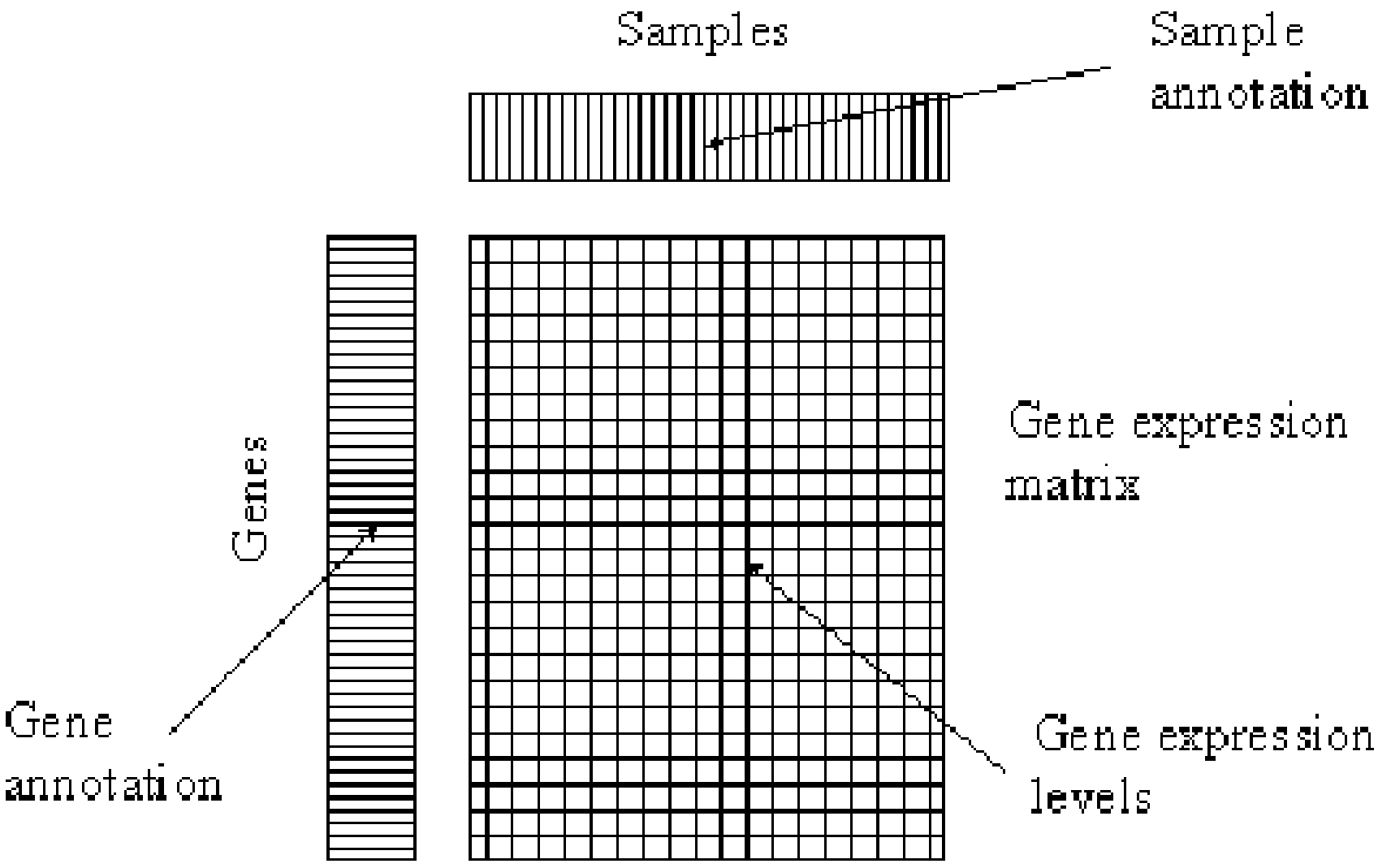
Statistical analysis of bulk RNA-seq data

Part I: linear model

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

Introduction



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

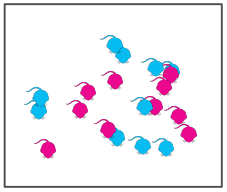
Grand Picture of Statistics

Statistical Hypotheses

H0: $\mu_B = \mu_L$

H1: $\mu_B \neq \mu_L$

Sample/Sequencing/alignment



Idea:

EGF is differentially expressed (DE) in luminal (L) and basal (B) cells

Data: RNASeq counts

$(y_{B,1}; y_{B,2}; ...; y_{B,n_B})$

$(y_{L,1}; y_{L,2}; ...; y_{L,n_L})$

Point estimation

$\widehat{\mu}_B - \widehat{\mu}_L$

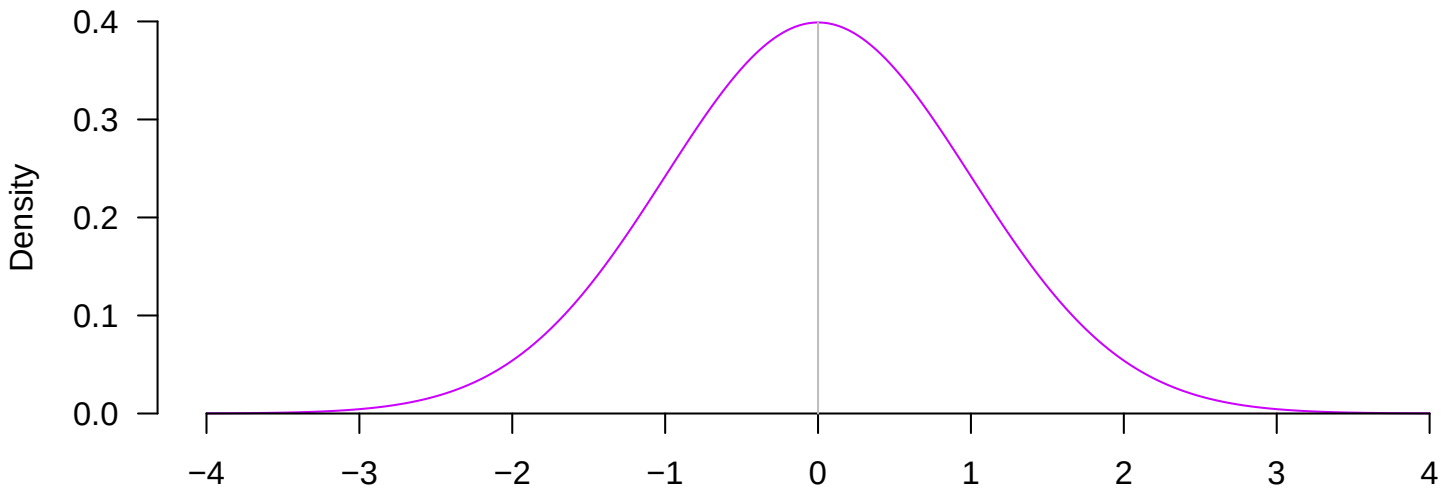
Inference

$$T_{obs} = \frac{\widehat{\mu}_B - \widehat{\mu}_L}{s_p \sqrt{\frac{1}{n_B} + \frac{1}{n_L}}} \sim St_{n_T + n_C - 2}$$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical tests

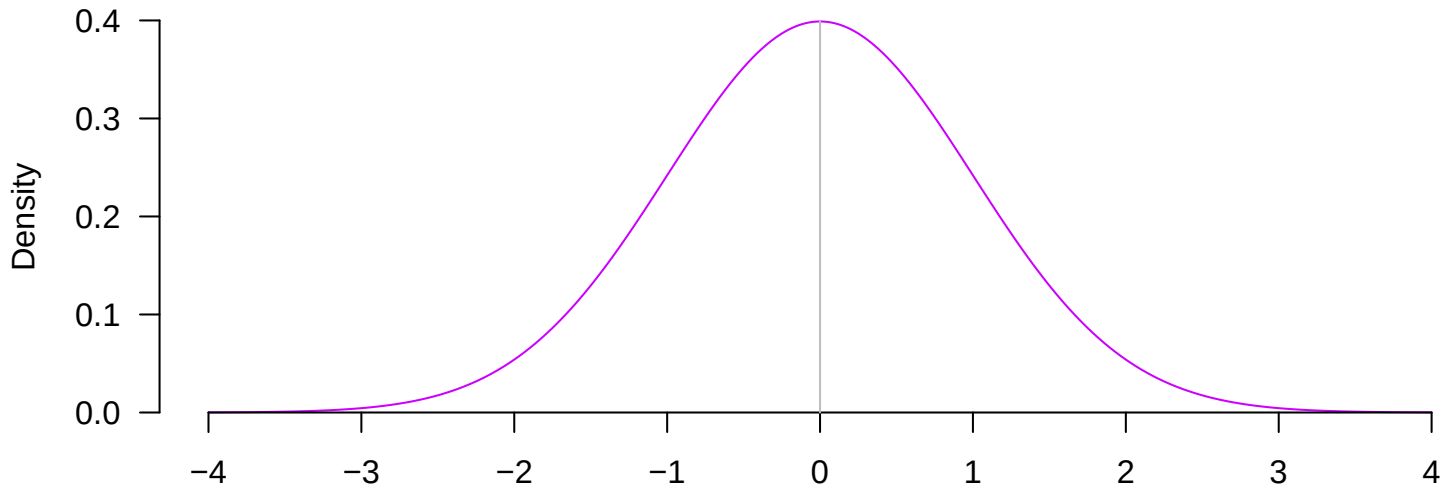
Assess how likely the observed test statistics is compared to the test statistics distribution under H0:



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical tests

Assess how likely the observed test statistics is
compared to the test statistics distribution under H0:



P-value for a two-sided test: p-value $= P(|T| > T_{obs})$

i.e. the probability of getting a test statistic as extreme or more extreme than the calculated test statistic if H0 is true

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical tests

4 possible outcomes

Conclude:

- if p-value $> \alpha$ $\rightarrow$ do not reject H0.
- if p-value $< \alpha$ $\rightarrow$ reject H0 in favour of H1.

		Test Outcome	
		H0 not rejected	H1 accepted
Unknown Truth	H0 true	$1-\alpha$ [TN]	α [FP]
	H1 true	β [FN]	$1-\beta$ [TP]

where

- α is the type I error,
- β is the type II error.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical tests

4 possible outcomes

Conclude:

- if p-value $> \alpha$ $\rightarrow$ do not reject H0.
- if p-value $< \alpha$ $\rightarrow$ reject H0 in favour of H1.

		Test Outcome	
		H0 not rejected	H1 accepted
Unknown Truth	H0 true	$1-\alpha$ [TN]	α [FP]
	H1 true	β [FN]	$1-\beta$ [TP]

where

- α is the type I error,
- β is the type II error.

Want to minimise FP and FN through design

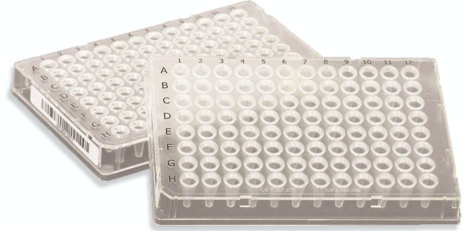
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Experimental design

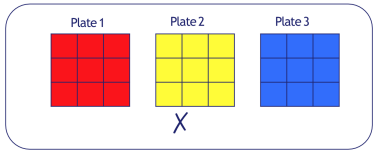
3 fundamental aspects of sounds experiments (Fisher 1935)

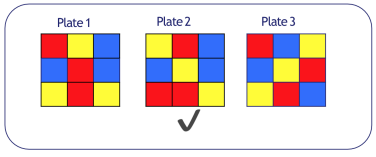
- **Replication**
  Try to capture all sources of variability
  (Biological versus technical variability)

- **Blocking**
  Try to remove technical biases/confounding
  (Lane and batch effects)



- **Randomisation**
  Try to remove confounding due to other factors





CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Experimental design

Sample size per condition

Sample size calculation:

Aim is to define the sample size allowing to detect an effect of a given size at the α level with a given probability (power):

- δ, the effect size: function of μ_L and μ_B (log fold change, standardised difference),
- $1 - \beta$, the power,
- α, the type I error.
- ϕ, nuisance parameters (variability, sequencing depth, multiplicity correction)

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

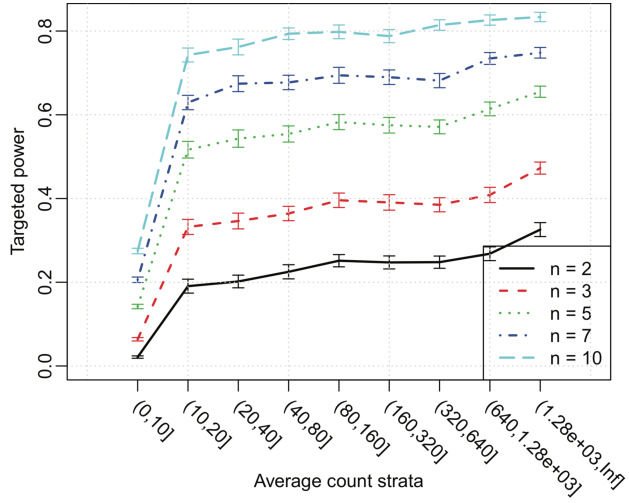
Experimental design

Sample size per condition

Sample size calculation:

Aim is to define the sample size allowing to detect an effect of a given size at the α level with a given probability (power):

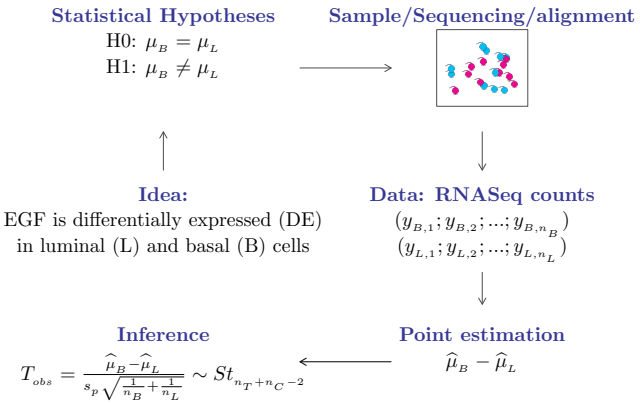
- δ, the effect size: function of μ_L and μ_B (log fold change, standardised difference),
- $1 - \beta$, the power,
- α, the type I error.
- ϕ, nuisance parameters (variability, sequencing depth, multiplicity correction)

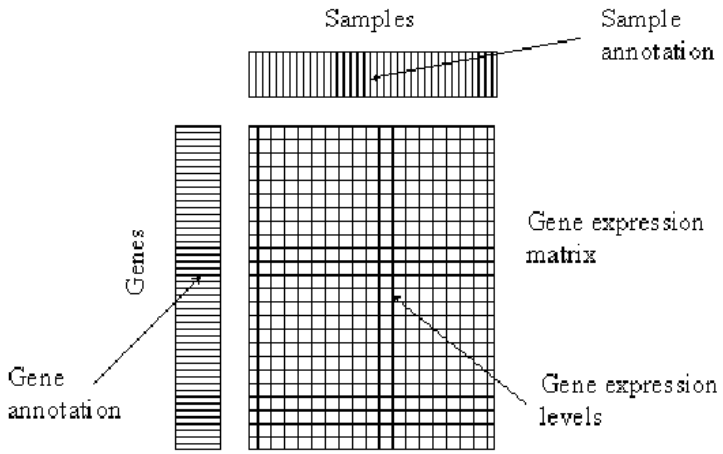


(Wu, Wang and Wu (2015))

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

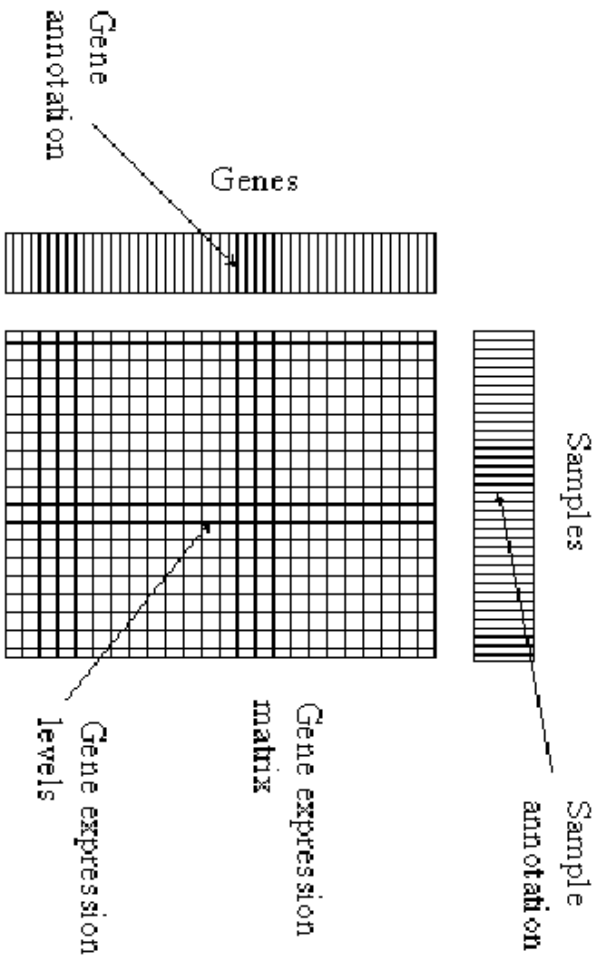
Statistical modelling





CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical modelling



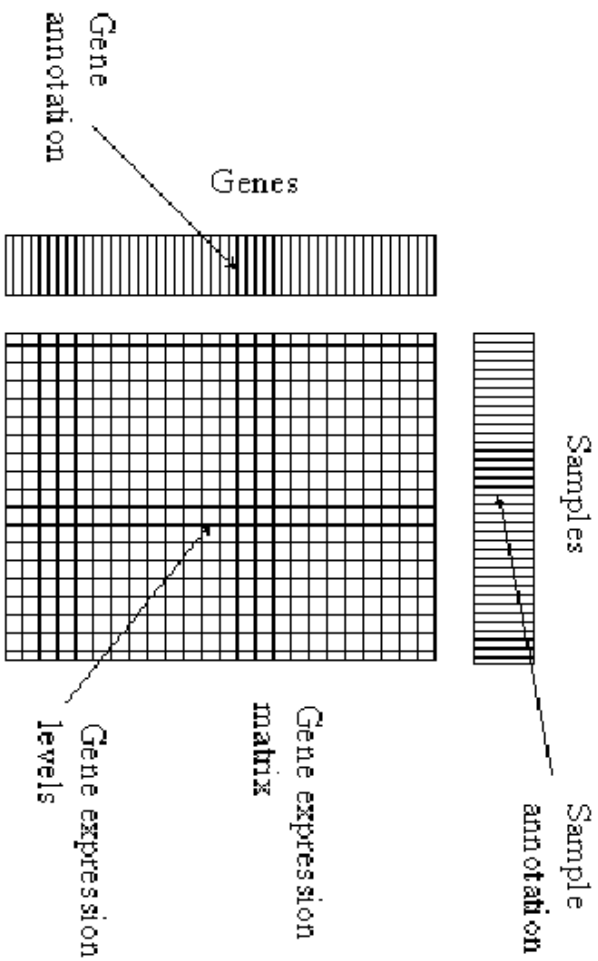
$$\mathbf{y} = f(\mathbf{X}) + \epsilon$$
$$\mathrm{E}[\mathbf{y}] = f(\mathbf{X})$$

where

- $\mathbf{y}$ denotes the $(n \times 1)$ vector of expression intensities of a given gene,
- $\mathbf{X}$ denotes the $(n \times p)$ design/predictor matrix,
- ϵ denotes the $(n \times 1)$ stochastic error vector,
- $\mathrm{E}[\mathbf{y}]$ denotes the expectation of $\mathbf{y}$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical modelling : Linear regression



$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

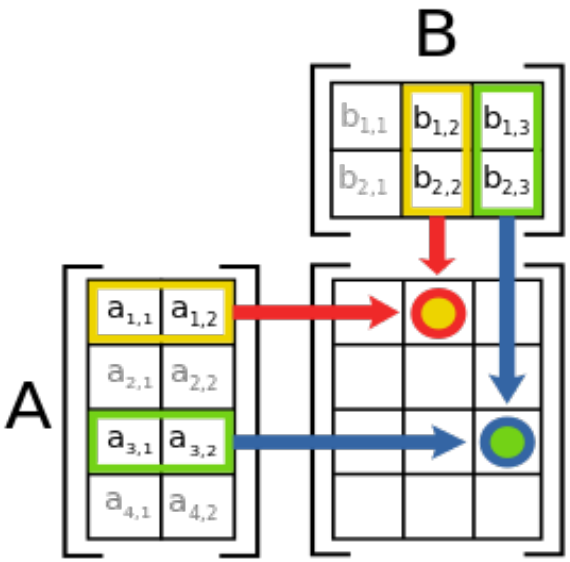
$$\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$$

where

- $\mathbf{y}$ denotes the $(n \times 1)$ vector of expression intensities of a given gene,
- $\mathbf{X}$ denotes the $(n \times p)$ design/predictor matrix,
- $\boldsymbol{\beta}$ denotes the $(p \times 1)$ parameter vector,
- $\epsilon \sim N(0, \sigma^2)$ denotes the $(n \times 1)$ stochastic error vector,
- $\mathrm{E}[\mathbf{y}]$ denotes the expectation of $\mathbf{y}$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical modelling : Linear regression



(Wikipedia)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

$$\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$$

where

- $\mathbf{y}$ denotes the $(n \times 1)$ vector of expression intensities of a given gene,
- $\mathbf{X}$ denotes the $(n \times p)$ design/predictor matrix,
- $\boldsymbol{\beta}$ denotes the $(p \times 1)$ parameter vector,
- $\epsilon \sim N(0, \sigma^2)$ denotes the $(n \times 1)$ stochastic error vector,
- $\mathrm{E}[\mathbf{y}]$ denotes the expectation of $\mathbf{y}$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical modelling : Strategy

- Collect the information related to each sample for the predictors of interest,
- define $\boldsymbol{\beta}$, the sets of parameters we are interested in,
- build the $\mathbf{X}$ matrix that relates the sample information with the $\boldsymbol{\beta}$,
- estimate the $\boldsymbol{\beta}$,
- use statistical inference to assess significance (p-values).

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Statistical modelling : Contrast matrices

Contrast matrices for models with

- one factor / categorical predictor,
  - two experimental conditions (dichotomous predictor), t-test
  - several experimental conditions, ANOVA
- two factors / categorical predictors,
  - without interaction,
  - with interaction,

  Two-way ANOVA
- categorical and continuous factors.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a two-level factor

Sample	Treatment
Sample1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

Number of samples: 6
Number of factors: 1 with 2 levels (Control and Treatment A)

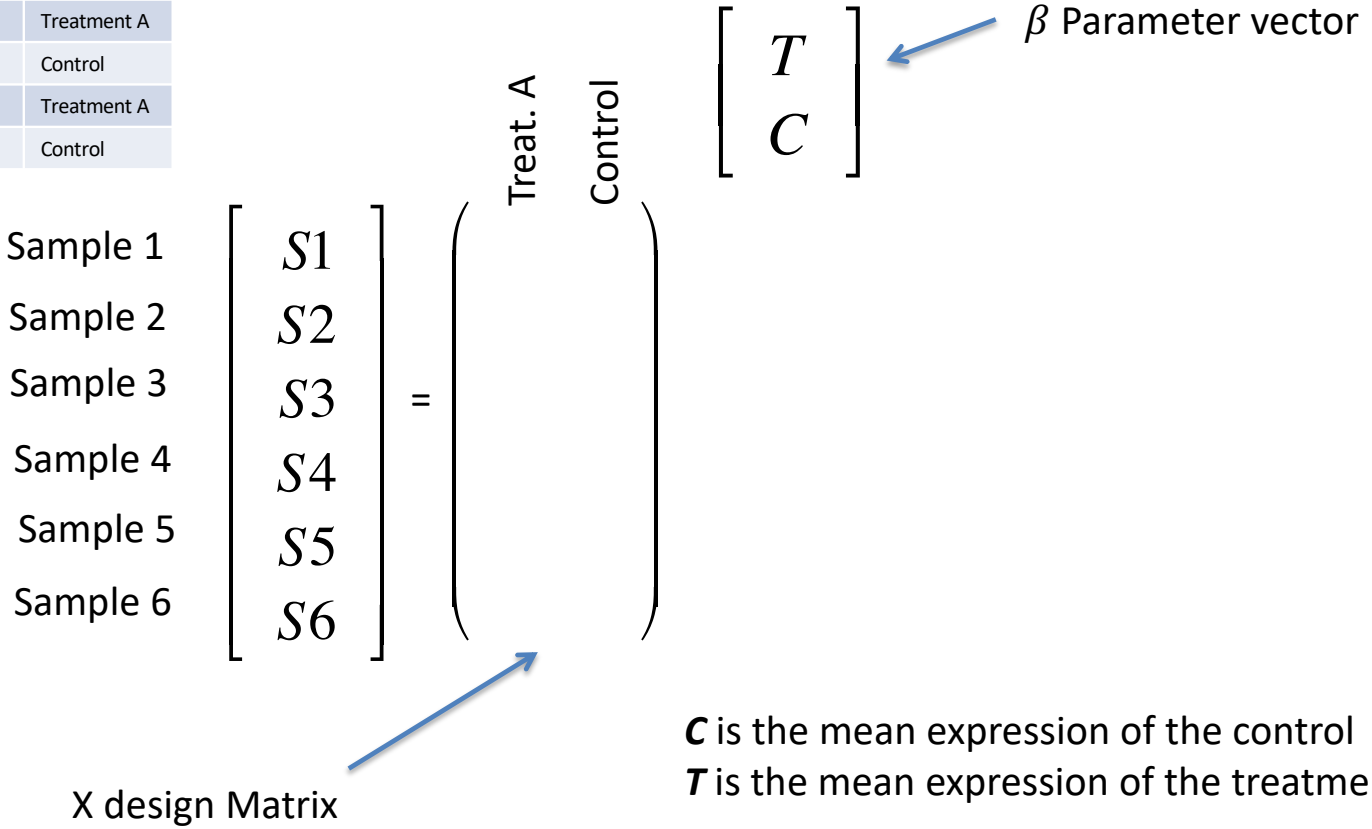
Possible parameters (What differences are important)?

- Effect of Treatment A
- Effect of Control

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a two-level factor

Sample	Treatment
Sample1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control



C is the mean expression of the control
T is the mean expression of the treatment

Design matrix for models with a two-level factor

Different parameterisation: using intercept

Sample	Treatment
Sample1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

Let's now consider this parameterization:

C= Baseline expression
T_A= Baseline expression + effect of treatment

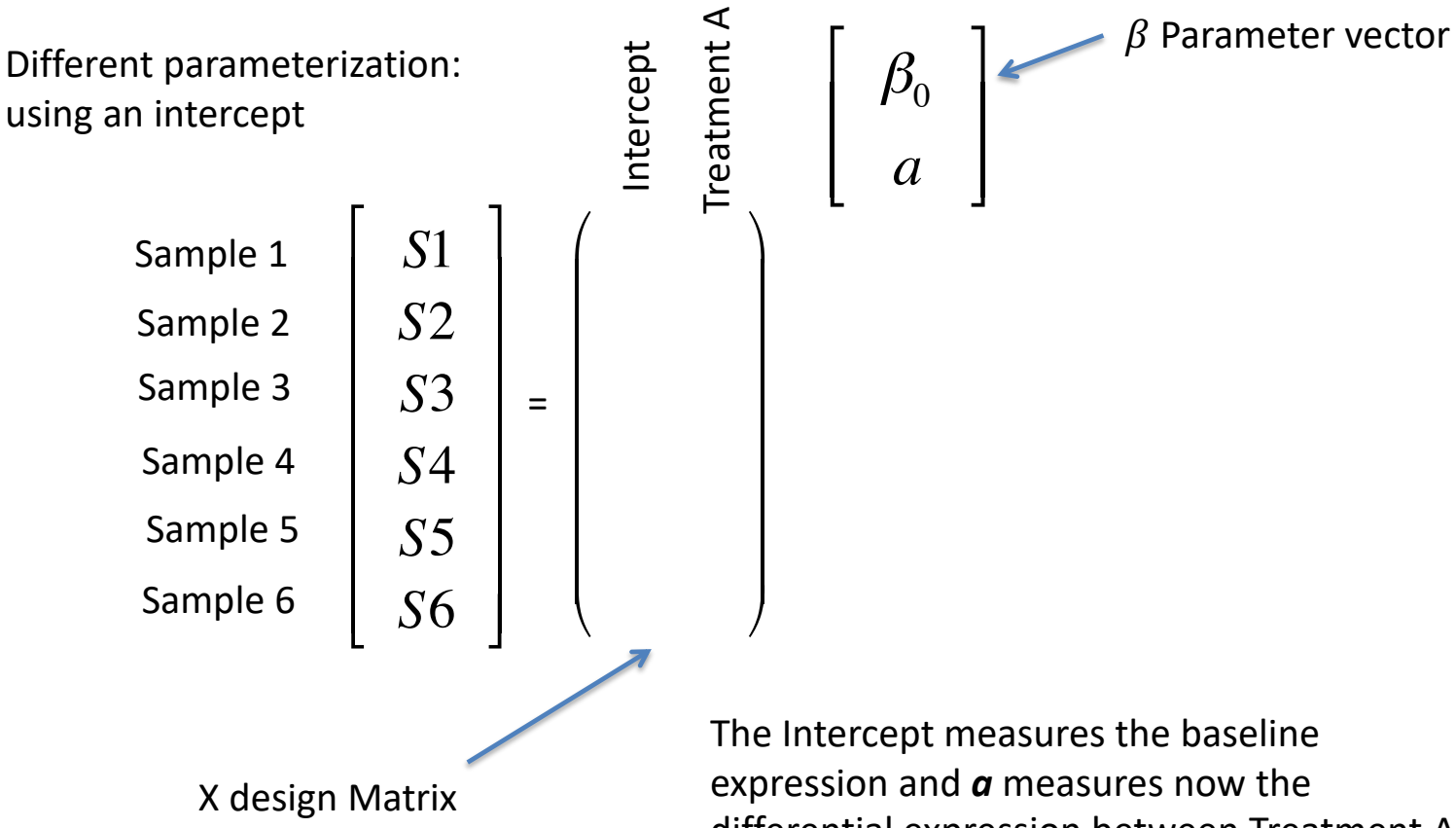
So the set of parameters are:

C = Control (mean expression of the control)
a = T_A – Control (mean change in expression under treatment

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a two-level factor

Different parameterization: using an intercept



The Intercept measures the baseline expression and ***a*** measures now the differential expression between Treatment A and Control

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a two-level factor

The two parameterizations are equivalent but allows to test different contrasts/parameters

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \hat{T} \\ \hat{C} \end{bmatrix} = \widehat{T - C}$$

Contrast matrix

Contrast matrices allow us to estimate (and test) linear combinations of our coefficients.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a three-level factor

Sample	Treatment
Sample1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

Number of samples: 6
Number of factors: 1 with 3 levels (Control, Treatment A, Treatment B)

Possible parameters (What differences are important)?

- Effect of Treatment A
- Effect of Treatment B
- Effect of Control
- Differences between treatments?

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a three-level factor

Sample	Treatment
Sample1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} \end{pmatrix} \begin{bmatrix} T_A \\ T_B \\ C \end{bmatrix}$$

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ b \end{bmatrix}$$

Control = Baseline
T_A = Baseline + a
T_B = Baseline + b

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a three-level factor

The model with intercept always take one level as a reference group:

The reference group here is treatment A, the coefficients are comparisons against it!

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{bmatrix} \beta_0 \\ b \\ c \end{bmatrix}$$

By default, R uses the first level as baseline

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a three-level factor: R code

```
> one3levelfactor = data.frame(condition =
                    rep(c("TreatmentA", "TreatmentB", "Control"), 2))

# model without intercept and default levels:
> X1 = model.matrix(~ condition - 1, data = one3levelfactor)

# model with intercept and default levels
> X2 = model.matrix(~ condition, data = one3levelfactor)

# model with intercept and self-defined levels
> levels(one3levelfactor$condition)
> levels(one3levelfactor$condition) = c("TreatmentB", "TreatmentA", "Control")
> X3 = model.matrix(~ condition, data = one3levelfactor)
```


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a three-level factor: Exercise

Build contrast matrices for all pairwise comparisons for this design:

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} T_A \\ T_B \\ C \end{bmatrix} \qquad \begin{pmatrix} \end{pmatrix} \begin{bmatrix} \hat{T}_A \\ \hat{T}_B \\ \hat{C} \end{bmatrix}$$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design matrix for models with a three-level factor: Exercise

Build contrast matrices for all pairwise comparisons for these designs:

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ b \end{bmatrix}$$

$$\begin{pmatrix} \end{pmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{a} \\ \hat{b} \end{bmatrix}$$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Models with 2 factors

Sample	Treatment	ER status
Sample1	Treatment A	+
Sample 2	No Treatment	+
Sample 3	Treatment A	+
Sample 4	No Treatment	+
Sample 5	Treatment A	-
Sample 6	No Treatment	-
Sample 7	Treatment A	-
Sample 8	No Treatment	-

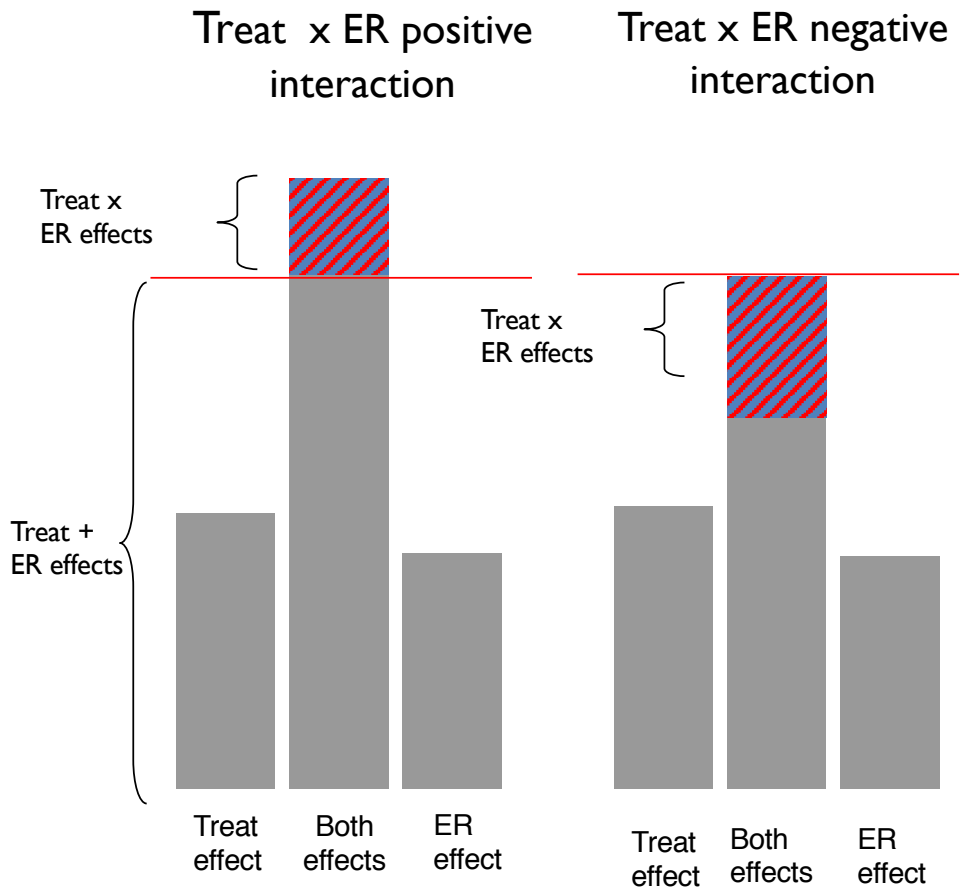
Number of samples: 8
Number of factors: 2 two-level factors

```
> two2levelfactor = data.frame(treatment = rep(c("TreatA","NoTreat"),4),
                               er = rep(c("+","-"),each=4))
```


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Models with 2 factors: interactions

	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3



(Adapted from Natalie Thorne, Nuno L. Barbosa Morais)

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Models with 2 factors: no interaction

```
X1 = model.matrix(~ treatment + er, data=two2levelfactor)
```

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \end{bmatrix} = \begin{pmatrix} \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ er+ \end{bmatrix}$$

	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Models with 2 factors: with interaction

```
> X2 = model.matrix(~ treatment * er, data=two2levelfactor)
> X3 = model.matrix(~ treatment + er + treatment:er, data=two2levelfactor)
```

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ er + \\ a.er + \end{bmatrix}$$

Interaction effect of Treatment A on ER+ samples

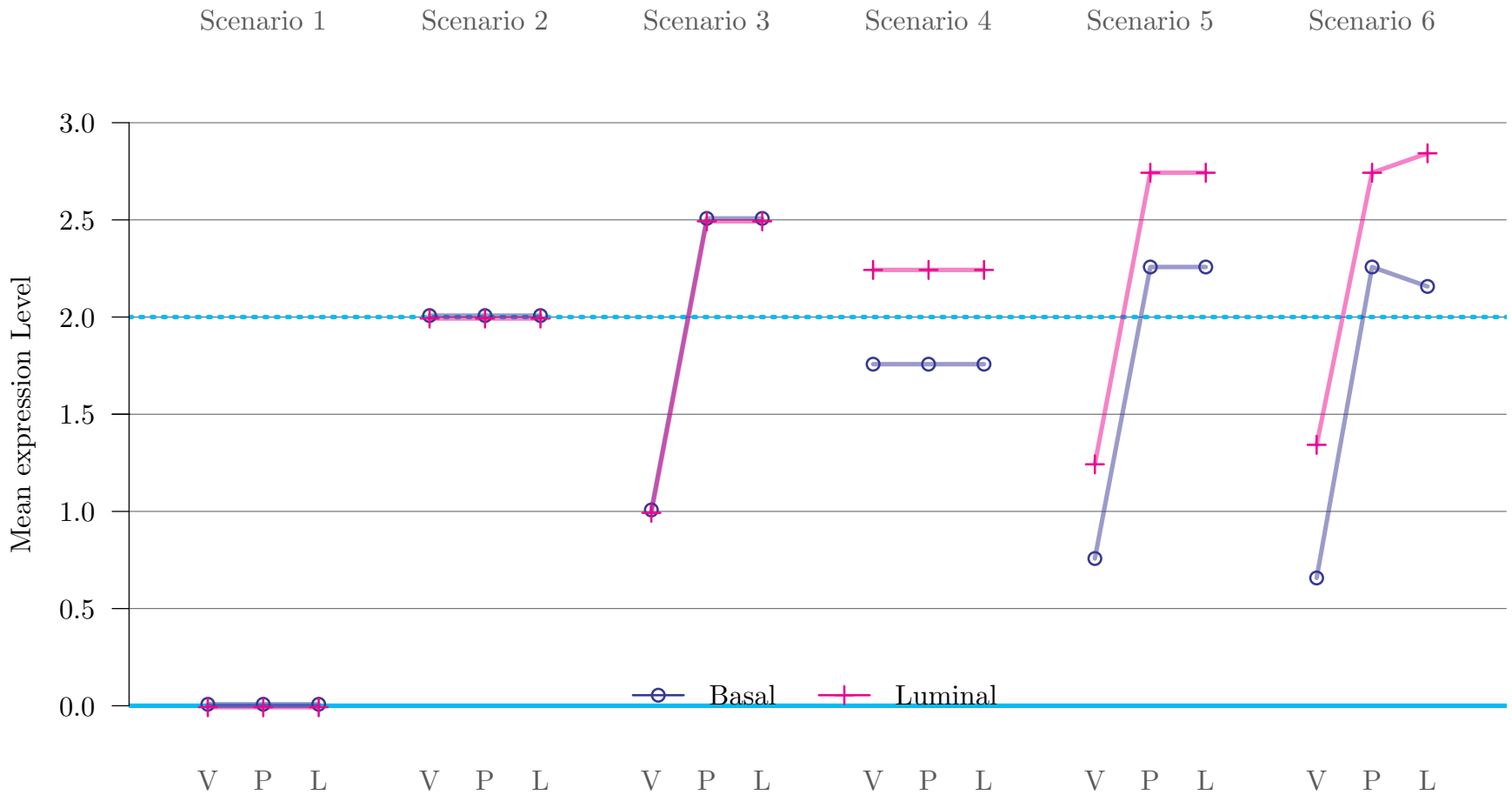
	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Models with 2 factors: possible scenarios

2 factors:

- cell type (2 levels): luminal versus basal
- mouse type (3 levels): virgin, pregnant, lactating



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Models with 2 predictors: a factor and a continuous one

Sample	ER	Dose
Sample 1	+	37
Sample 2	-	52
Sample 3	+	65
Sample 4	-	89
Sample 5	+	24
Sample 6	-	19
Sample 7	+	54
Sample 8	-	67

Number of samples: 8
2 predictors: ER (a two-level factor) and Dode (a continuous predictor)

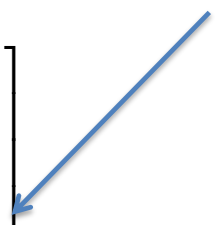
```
> mixedpredictors = data.frame(er = rep(c("+","-"),4),
                               dose = c(37,52,65,89,24,19,54,67))
```


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Models with 2 predictors: a factor and a continuous one

```
X = model.matrix(~ er + dose, data= mixedpredictors)
```

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 37 \\ 1 & 0 & 52 \\ 1 & 1 & 65 \\ 1 & 0 & 89 \\ 1 & 1 & 24 \\ 1 & 0 & 19 \\ 1 & 1 & 54 \\ 1 & 0 & 67 \end{pmatrix} \begin{bmatrix} \beta_0 \\ er + \\ d \end{bmatrix}$$



If we consider the effect of dose ***linear*** we use 1 coefficient (degree of freedom). We can also model it as non-linear (using splines, for example).

Sample	ER	Dose
Sample 1	+	37
Sample 2	-	52
Sample 3	+	65
Sample 4	-	89
Sample 5	+	24
Sample 6	-	19
Sample 7	+	54
Sample 8	-	67

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Model Estimation and inference

$$Y = X\beta + \varepsilon$$

β → Parameter of interest

$\hat{\beta}$ → **Estimate** of the parameter of interest

$se(\hat{\beta})$ → **Standard Error** of the estimator of the parameter of interest

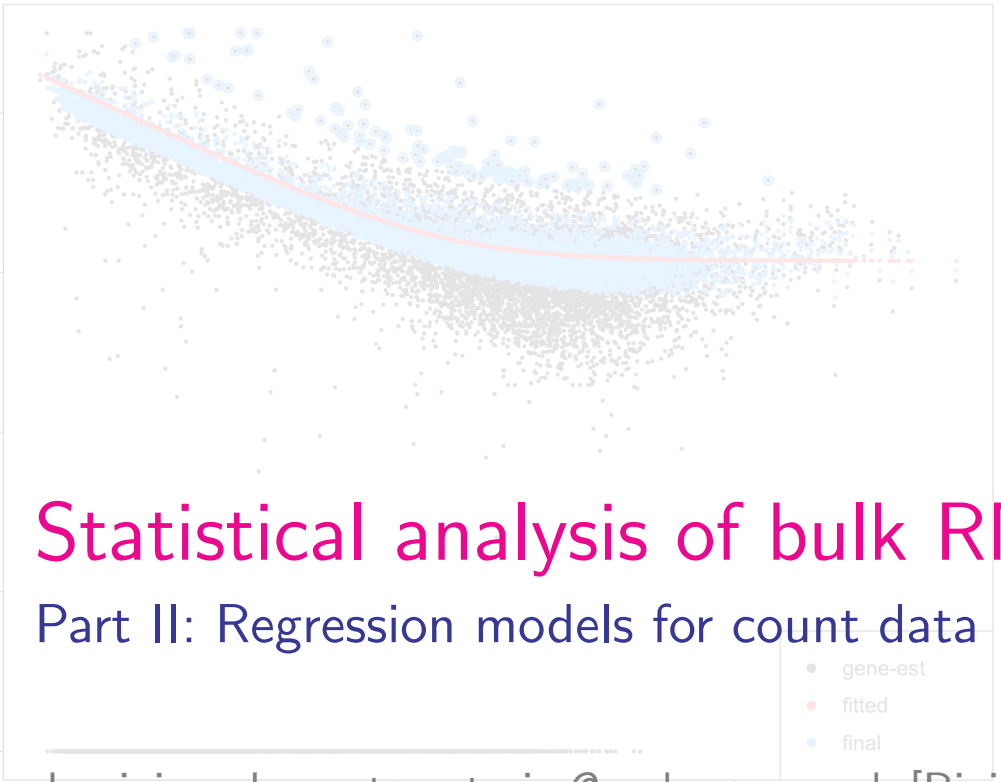
$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad MLE : \hat{\beta} = \arg\max\{L(\beta \mid x)\}$$

$$se(\hat{\beta}_i) = \sigma\sqrt{c_i} \quad \text{where } c_i \text{ is the } i^{\text{th}} \text{ diagonal element of } \left(X^T X\right)^{-1}$$

$\hat{y} = X\hat{\beta}$ → Fitted values (predicted by the model)

$e = y - \hat{y}$ → Residuals (observed errors)

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE



Statistical analysis of bulk RNA-seq data

Part II: Regression models for count data

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

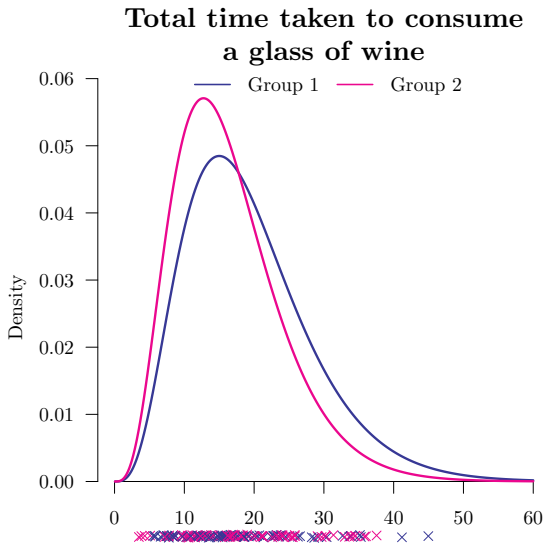
The mean is taken as "normalized counts" scaled by a normalization factor

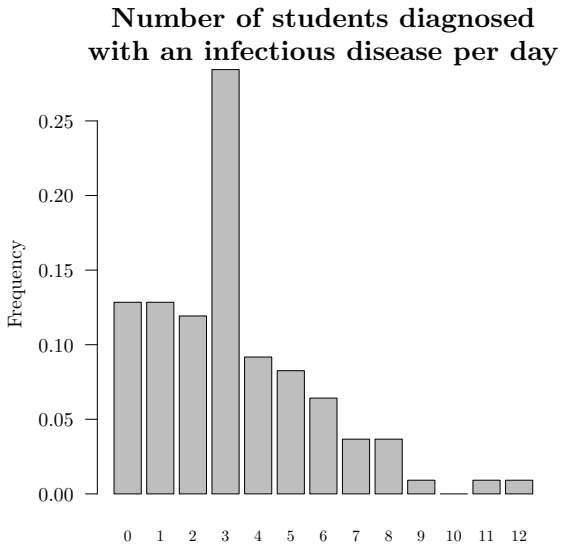
raw count for gene i, sample j

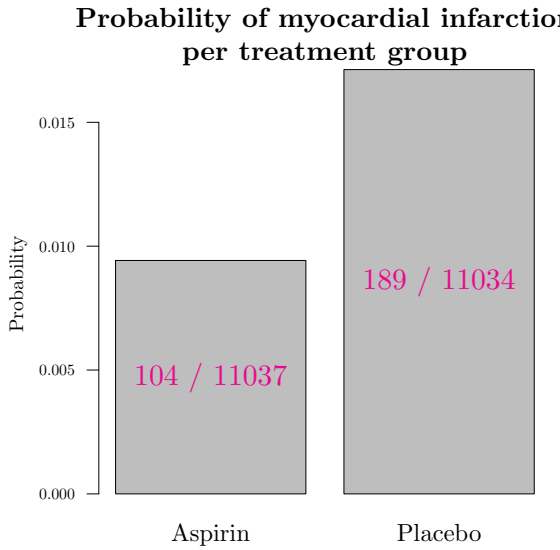
one dispersion per gene

$$K_{ij} \sim \mathrm{NB}(s_{ij}q_{ij}, \alpha_i)$$

Examples of data with non-normal conditional distributions

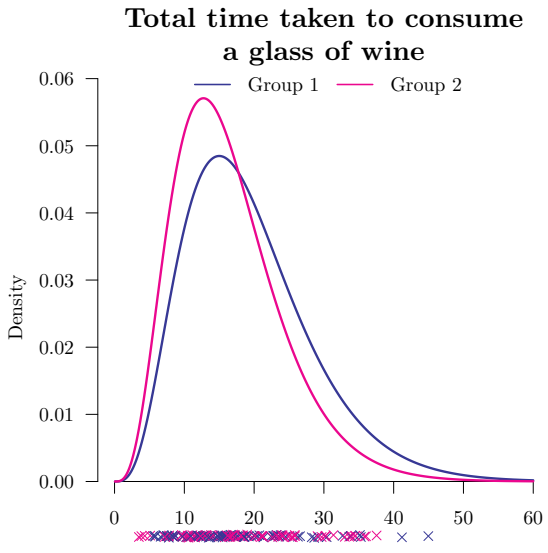


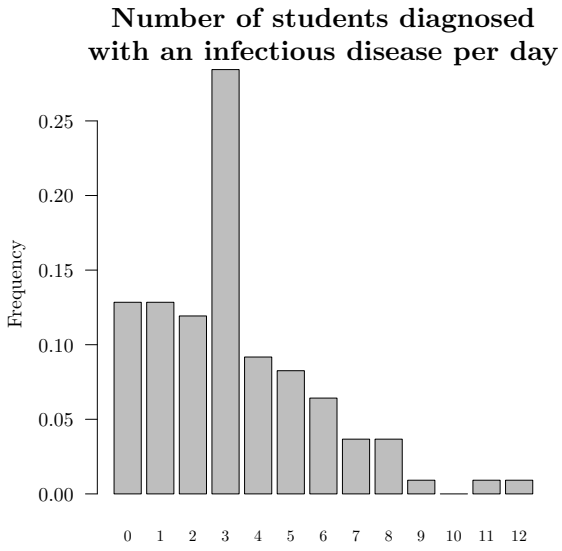


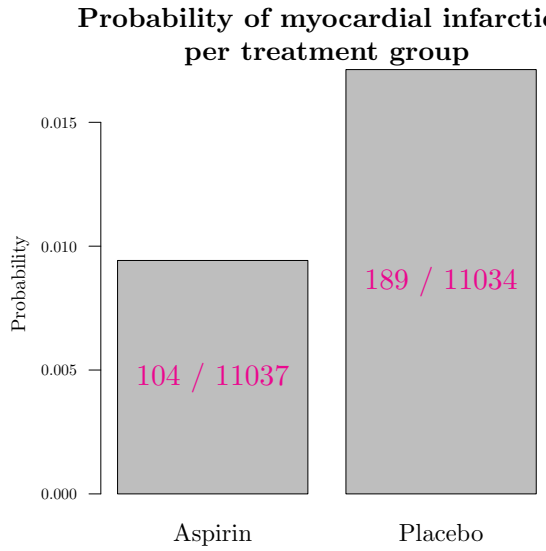


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Examples of data with non-normal conditional distributions







Linear model not suitable:

- Assumed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2),$$

  - theoretical range of $\epsilon = [-\infty, +\infty]$,
  - $\mathbf{X}\boldsymbol{\beta}$ not bounded to $[0, \infty]$ or $[0, 1]$,
  - Var$[\mathbf{y}]$ independent of E$[\mathbf{y}]$.

- Solution:

$$\mathbf{y}|(\mathbf{X}, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{X}\boldsymbol{\beta}), \phi),$$

where $distribution$ belongs to the exponential family and $function$ is monotonically increasing.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

GLM: conditional distributions

$$\mathbf{y}|(\mathbf{X}, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{X}\boldsymbol{\beta}), \phi),$$

- Some possible conditional *distributions* :
  statistical probability mass functions & density functions
  - Within the exponential family ['classical' GLM framework]

normal	chi-squared	Poisson	Inverse Wishart
exponential	beta	Negative Binomial	
gamma	Dirichlet	Bernoulli	...

  - Outside the exponential family ['extended' GLM framework]

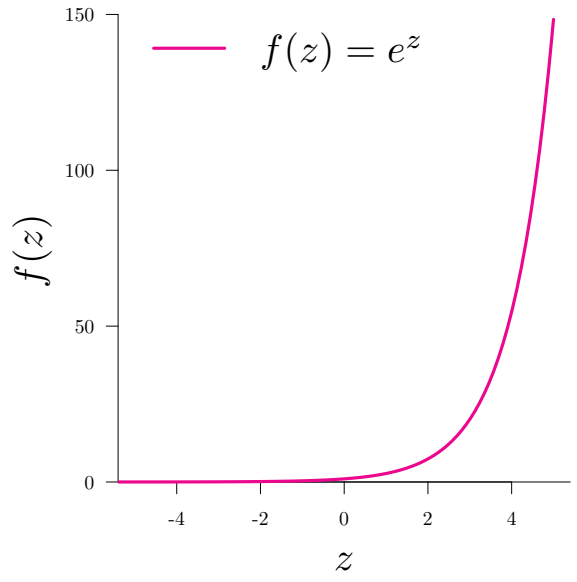
Box-Cox power	Gaussian	Weibull
exponential	inverse Gaussian	Pareto type I, II, III
exponential Gaussian	logistic	Poisson inverse Gaussian
generalized beta	power exponential	
generalized gamma	reverse Gumbel	...
generalized inverse	skew power exponential	

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

GLM: link functions

$$\mathbf{y}|(\mathbf{X}, \boldsymbol{\beta}, \phi) \sim distribution(function(\mathbf{X}\boldsymbol{\beta}), \phi),$$

- Most used link $functions$:
  connection between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\beta}$
  - to restrict $f(\mathbf{X}\boldsymbol{\beta})$ to belong to $[0, \infty[$:
    - log link: $f(z) = e^z$



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Distribution for count data: Poisson

Example:
Interest for the number of reads/counts for gene 'X' for a sample basal cells of n mice

Sample of n mice:	$i = 1$	$i = 2$	$i = 3$	$\cdots$	$i = 115$
y_i	607	873	1218	$\cdots$	2715

If, during a time interval or in a given area,

- events occur independently,
- at the same rate,
- and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area),

then,

- a count occurring in a fixed time interval or in a given area, Y, may be modelled by means of a Poisson distribution with parameter μ:

$$Y \sim Poisson(\mu) \text{ where } \mu = \mathsf{E}[Y] = \mathsf{Var}[Y],$$

- the probability of observing x events during a fixed time interval or in a given area is given by

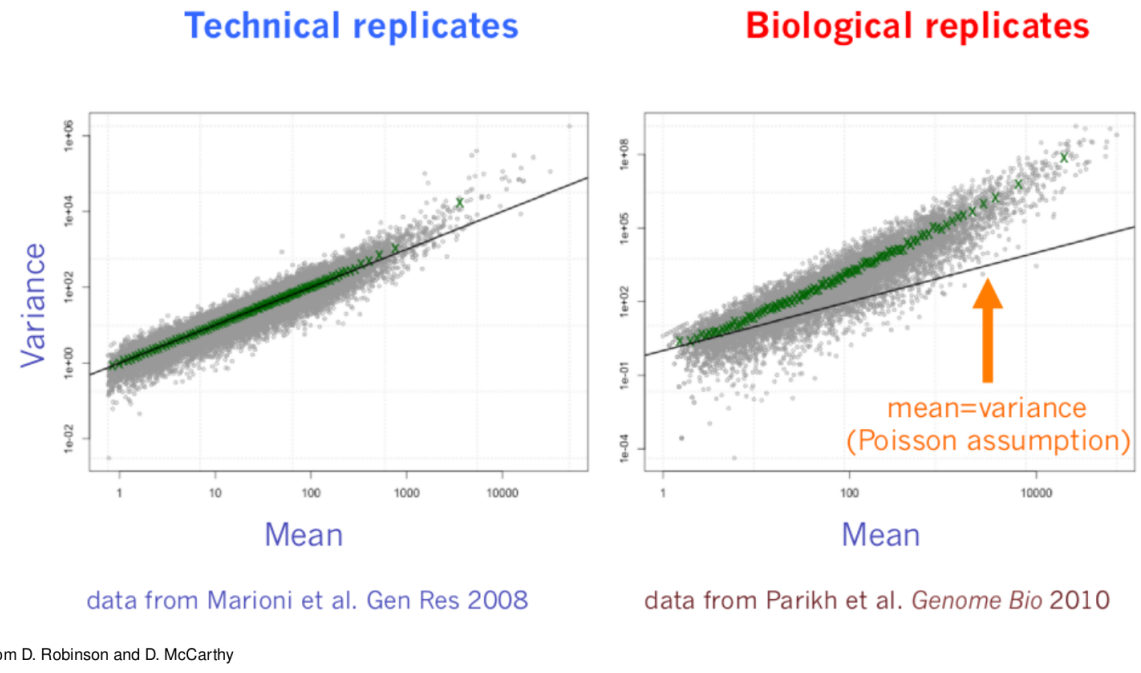
$$P(Y = y|\mu) = \frac{\mu^y e^{-\mu}}{y!}.$$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Distribution for count data: Poisson vs Neg. Bin.

Exploratory data analysis

scores between 0 and 1 $\Rightarrow$ underdispersion (variance smaller than mean)



From D. Robinson and D. McCarthy

scores greater than 1 : overdispersion $\Rightarrow$ adapted to biological replicates

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Distribution for count data: Poisson vs Neg. Bin.

Available tests

Models of count data

- Data transformation and gaussian-based model : limma - voom
- Poisson : TSPM
- Negative Binomial : edgeR, DESeq(2), NBPSeq, baySeq, ShrinkSeq, ...

Statistical approaches

- Frequentist Approach : edgeR, DESeq(2), NBPSeq, TSPM, ...
- Bayesian Approach : baySeq, ShrinkSeq, EBSeq, ...
- Non-parametric approach : SAMSeq, NOISeq, ...

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

2a/ Negative binomial

- General form:

$$Y_i \sim \mathsf{NB}(\mu_i, \phi)$$

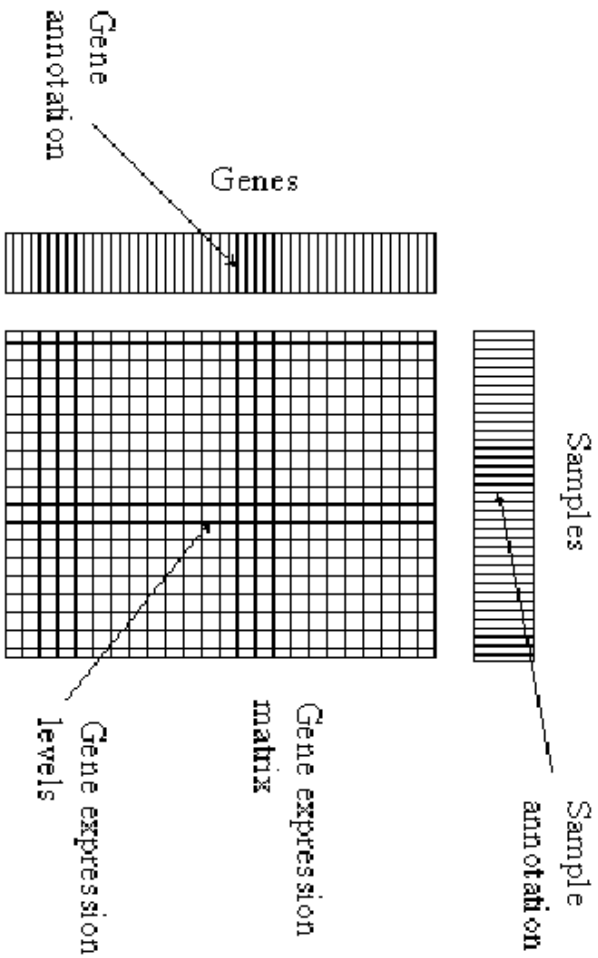
$$f_{Y_i}(y_i|\mu_i, \phi) = \frac{\Gamma(y + \frac{1}{\phi})}{\Gamma(\frac{1}{\phi})\Gamma(y+1)} \left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^y \left(\frac{1}{1+\phi\mu_i}\right)^{\frac{1}{\phi}}$$

with expectation and variance given by

  - $\mathsf{E}[Y_i] = \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$
  - $\mathsf{Var}[Y_i] = \mu_i(1 + \phi\mu_i)$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

2b/ Negative binomial regression



$$\mathbf{y} \sim \mathrm{NB}(\mu, \phi)$$

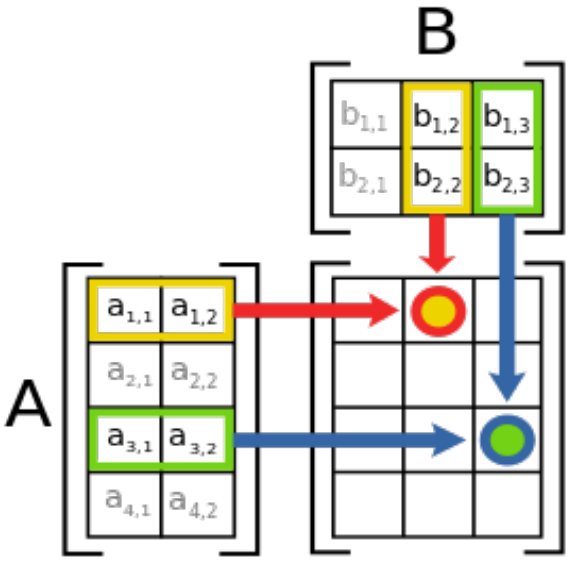
$$\mathrm{E}[\mathbf{y}] = \mu = \exp(\mathbf{X}\boldsymbol{\beta})$$

where

- $\mathbf{y}$ denotes the $(n \times 1)$ vector of expression intensities of a given gene,
- $\mathbf{X}$ denotes the $(n \times p)$ design/predictor matrix,
- $\boldsymbol{\beta}$ denotes the $(p \times 1)$ parameter vector,
- $\mathrm{E}[\mathbf{y}] = \mu$ denotes the expectation of $\mathbf{y}$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

2b/ Negative binomial regression



(Wikipedia)

$$\mathbf{y} \sim \mathrm{NB}(\mu, \phi)$$

$$\mathrm{E}[\mathbf{y}] = \mu = \exp(\mathbf{X}\boldsymbol{\beta})$$

where

- $\mathbf{y}$ denotes the $(n \times 1)$ vector of expression intensities of a given gene,
- $\mathbf{X}$ denotes the $(n \times p)$ design/predictor matrix,
- $\boldsymbol{\beta}$ denotes the $(p \times 1)$ parameter vector,
- $\mathrm{E}[\mathbf{y}] = \mu$ denotes the expectation of $\mathbf{y}$

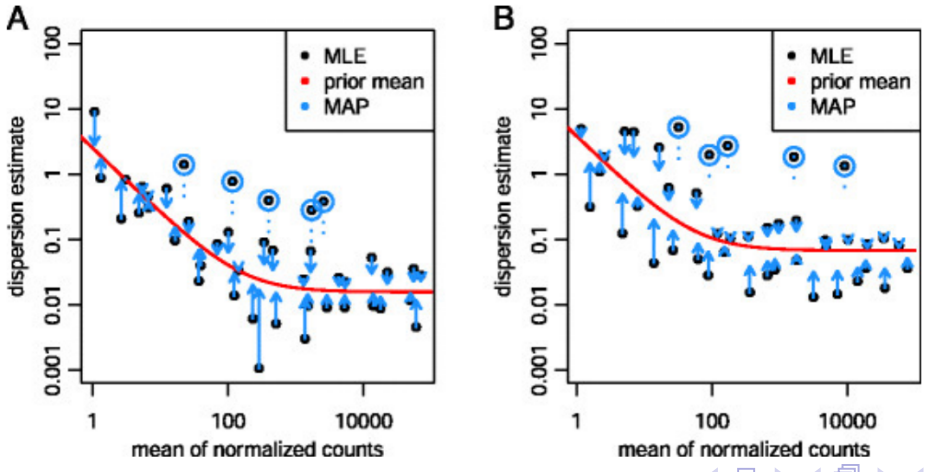
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

2c/ Negative binomial: Estimation

Dispersion estimation with DESeq2

Hypothesis : genes of similar average expression strength have similar dispersion

1. Estimate gene-wise dispersion estimates using maximum likelihood (ML) (black dots)
2. Fit a smooth curve (red line)
3. Shrink the gene-wise dispersion estimates (empirical Bayes approach) toward the values predicted by the curve to obtain final dispersion values (blue arrow heads).



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

2d/ Negative binomial: Controlling for library size

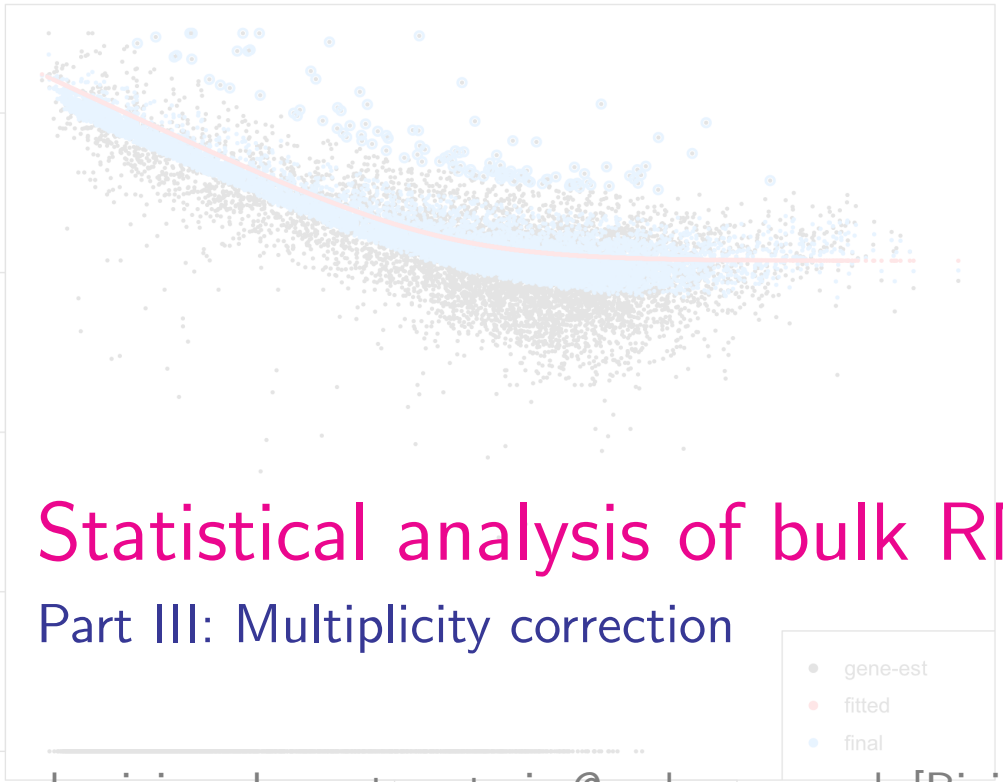
- For a given gene, the variance of the Negative Binomial for the ith sample is given by

$$\text{Var}(Y_i) = \mu_i(1 + \phi\mu_i)$$

- To control for the library size S_i of the ith sample, DESeq2 uses

$$\text{Var}(Y_i) = S_i\mu_i(1 + \phi S_i\mu_i)$$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE



Statistical analysis of bulk RNA-seq data

Part III: Multiplicity correction

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

3/ Multiplicity correction

Multiple Testing

False positive (FP) : A non differentially expressed (DE) gene which is declared DE.

For all 'genes', we test H_0 (gene i is not DE) vs H_1 (the gene is DE) using a statistical test

Problem

Let assume all the G genes are not DE. Each test is realized at α level
Ex : $G = 10000$ genes and $\alpha = 0.05 \rightarrow E(FP) = 500$ genes.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

3/ Multiplicity correction

The Family Wise Error Rate (FWER)

Definition

Probability of having at least one Type I error (false positive), of declaring DE at least one non DE gene.

$$FWER = \mathbb{P}(FP \leq 1)$$

The Bonferroni procedure

Either each test is realized at $\alpha = \alpha^*/G$ level
or use of adjusted pvalue $pBonf_i = min(1, p_i * G)$ and FWER $\leq \alpha^*$.
For $G = 2000$, $\leq \alpha^* = 0.05$, $\alpha = 2.510^{-5}$.

Easy but conservative and not powerful.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

The False Discovery Rate (FDR)

Idea : Do not control the error rate but the proportion of error
$\Rightarrow$ less conservative than control of the FWER.

Definition

The false discovery rate of [Benjamini and Hochberg, 1995] is the expected proportion of Type I errors among the rejected hypotheses

$$\text{FDR} = \mathbb{E}(FP/P) \text{ if } P > 0 \text{ and } 0 \text{ if } P = 0$$

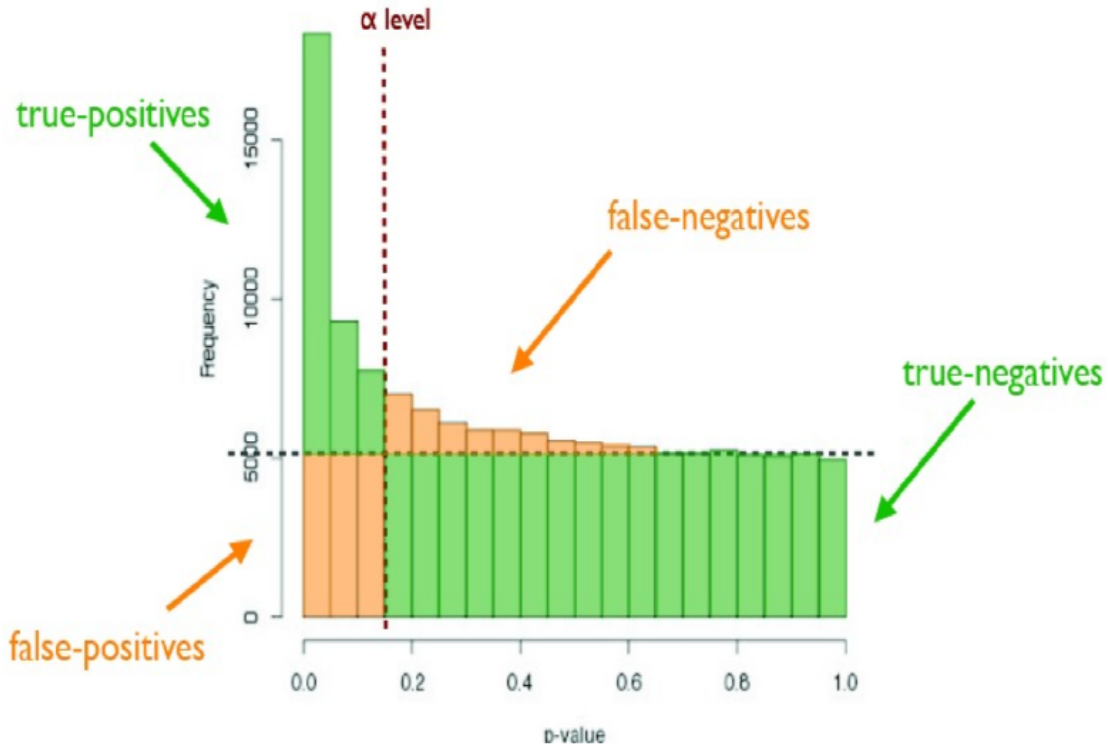
Prop

$$\text{FDR} \leq \text{FWER}$$

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

3/ Multiplicity correction

Standard assumption for p-value distribution



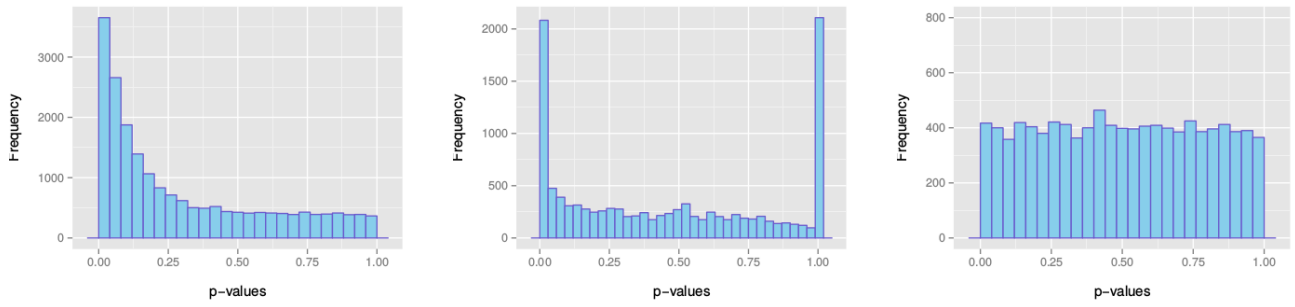
Source : M. Guedj, Pharnext

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

3/ Multiplicity correction

p-values histograms for diagnosis

Examples of expected overall distribution



(a) : the most desirable shape

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

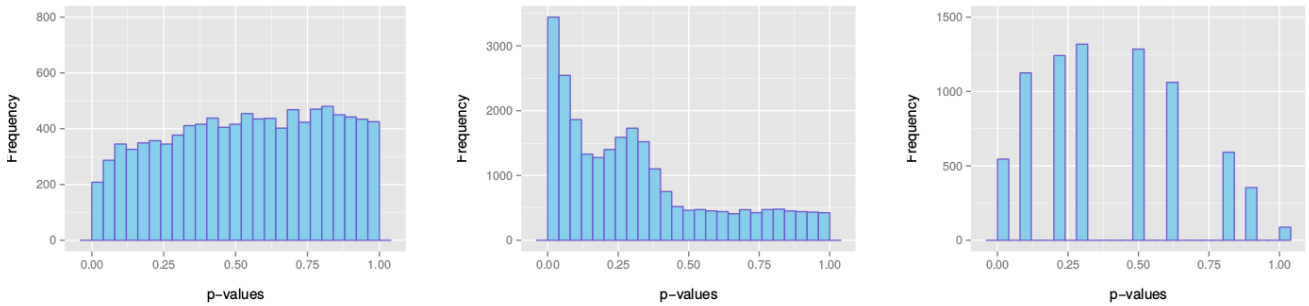
(c) : do not expect positive tests after correction

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

3/ Multiplicity correction

p-values histograms for diagnosis

Examples of not expected 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

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

3/ Multiplicity correction

Multiple testing : key points

- Important to control for multiple tests
- FDR or FWER depends on the cost associated to FN and FP

Controlling the FWER :

Having a great confidence on the DE elements (strong control). Accepting to not detect some elements (lack of sensitivity $\Leftrightarrow$ a few DE elements)

Controlling the FDR :

Accepting a proportion of FP among DE elements. Very interesting in exploratory study.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE