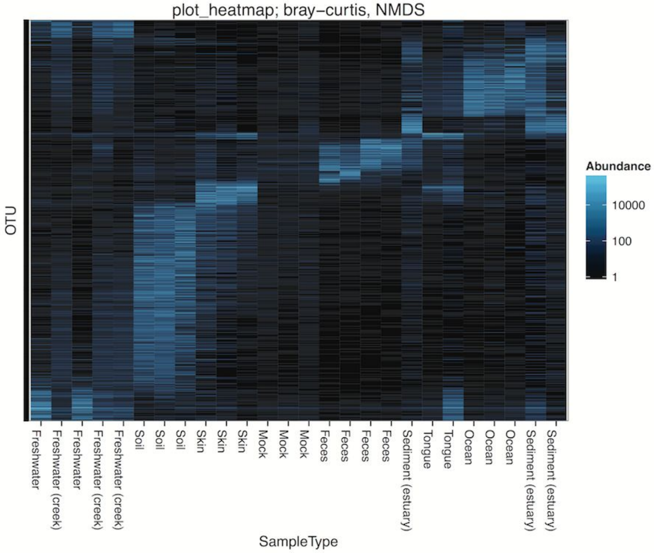
Deconstructing DESeq2

Building Blocks of Differential Abundance Testing

Goal: Characterizing Variation

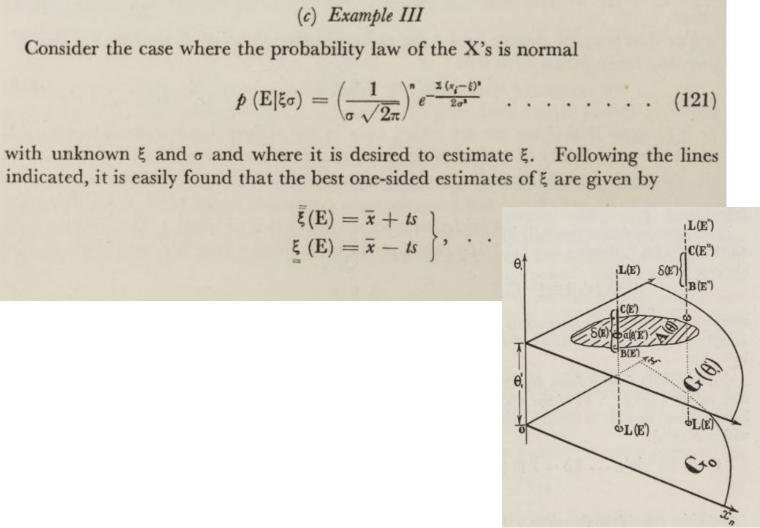
- How does variation in RSV counts reflect sample characteristics?
  - Is it consistent with existing theories?
  - Does it suggest new hypothesis?
- How are characteristics of columns associated with characteristics of rows?



Goal: Characterizing Variation

- Statistical Tools
  - **Inference**: Quantify degree of uncertainty in associations
  - **Visualization**: Compress complexity into interpretable representation

Confidence Intervals (from Neyman’s 1937 paper)



(*c*) *Example III*

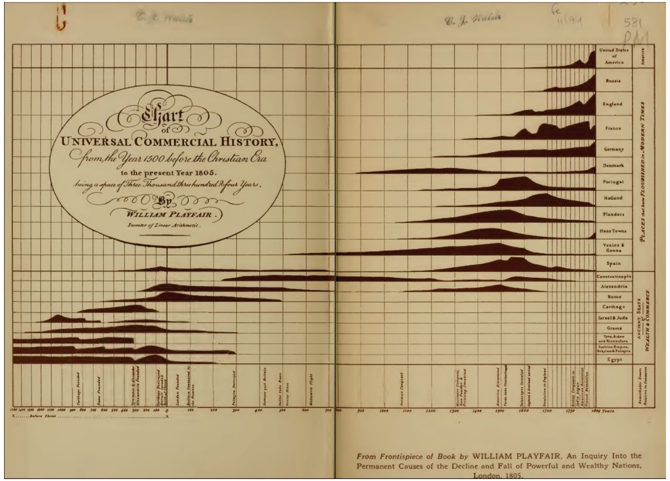
Consider the case where the probability law of the X’s is normal

$$p\,(\mathrm{E}|\xi\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\Sigma(x_i-\xi)^2}{2\sigma^2}} \quad . \; . \; . \; . \; . \; . \; . \; . \quad (121)$$

with unknown ξ and σ and where it is desired to estimate ξ. Following the lines indicated, it is easily found that the best one-sided estimates of ξ are given by

$$\left.\begin{aligned}\overline{\overline{\xi}}(\mathrm{E}) &= \bar{x} + ts\\ \underline{\underline{\xi}}\,(\mathrm{E}) &= \bar{x} - ts\end{aligned}\right\}, \quad . \; .$$

Time Series (from Playfair, 1805)

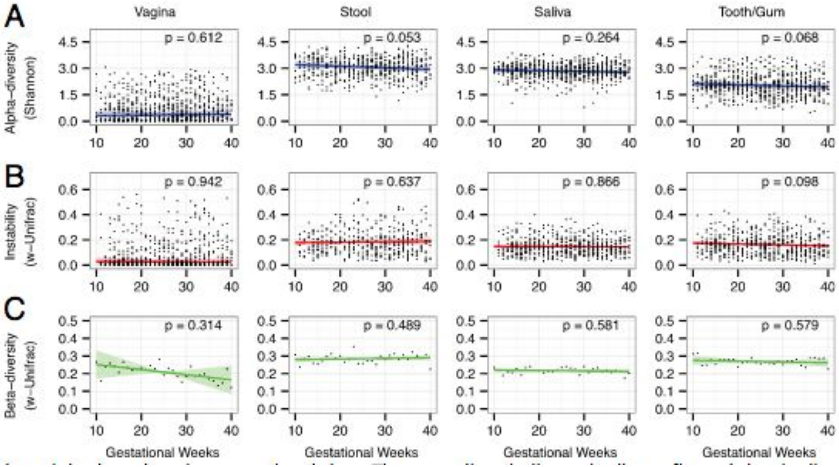


From Frontispiece of Book by WILLIAM PLAYFAIR, An Inquiry Into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations, London, 1805.

Goal: Characterizing Variation

- Statistical Tools
  - **Inference**: Quantify degree of uncertainty in associations
  - **Visualization**: Compress complexity into interpretable representation

Confidence intervals **AND** Time Series (Callahan 2015)



Goal: Characterizing Variation

- Statistical Tools
  - **Inference**: Quantify degree of uncertainty in associations
  - **Visualization**: Compress complexity into interpretable representation

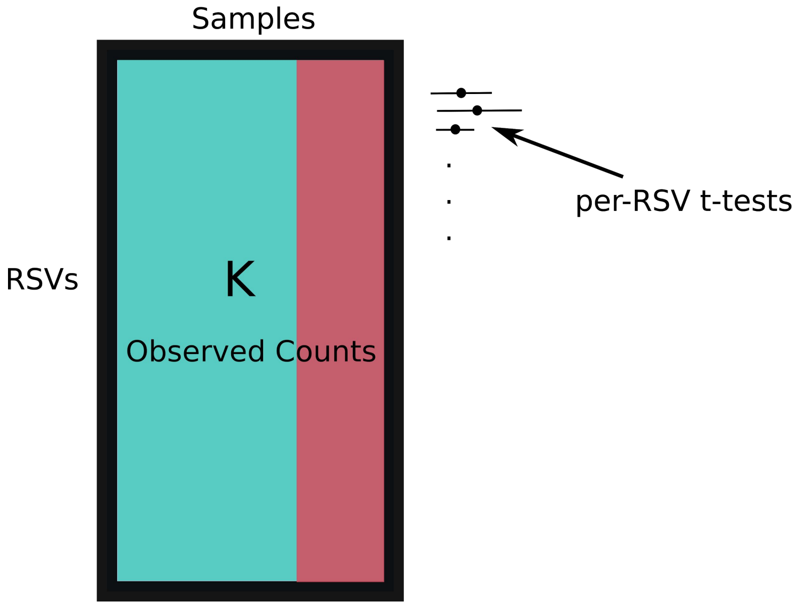
Our focus here will (mostly) be inference.

Goal: Characterizing Variation

- Statistical Tools
  - **Inference**: Quantify degree of uncertainty in associations
  - **Visualization**: Compress complexity into interpretable representation

Our focus here will (mostly) be inference.
A (naive) starting point:

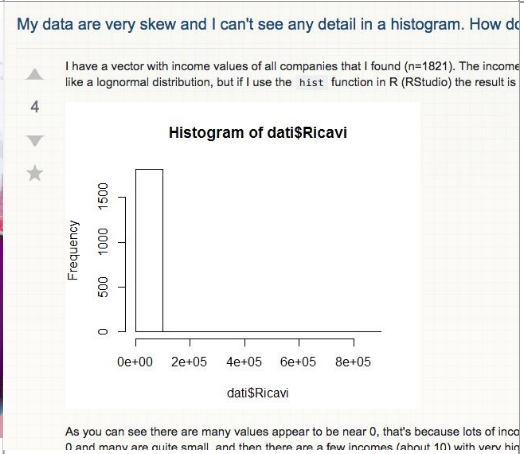
Control
Treatment



Challenges

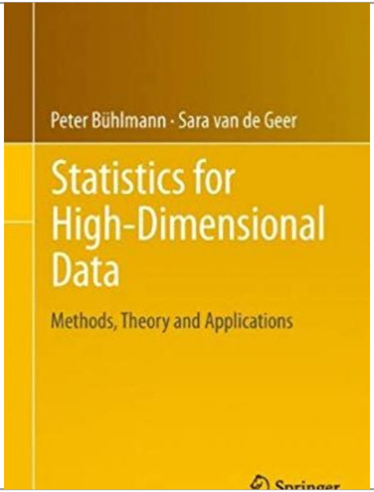
- A few characteristics of microbiome data make it challenging to analyze
- We'll discuss techniques for dealing with these issues
- Especially in relation to DESeq2

<table>
<tr>
<th>Batch Effects ("normalization")</th>
<th>Count structure / Skewness</th>
<th>High-Dimensionality
(few samples + multiple testing)</th>
</tr>
<tr>
<td>

</td>
<td>


My data are very skew and I can't see any detail in a histogram. How do

I have a vector with income values of all companies that I found (n=1821). The income
like a lognormal distribution, but if I use the `hist` function in R (RStudio) the result is

As you can see there are many values appear to be near 0, that's because lots of inco
0 and many are quite small, and then there are a few incomes (about 10) with very hig
</td>
<td>


Peter Bühlmann · Sara van de Geer

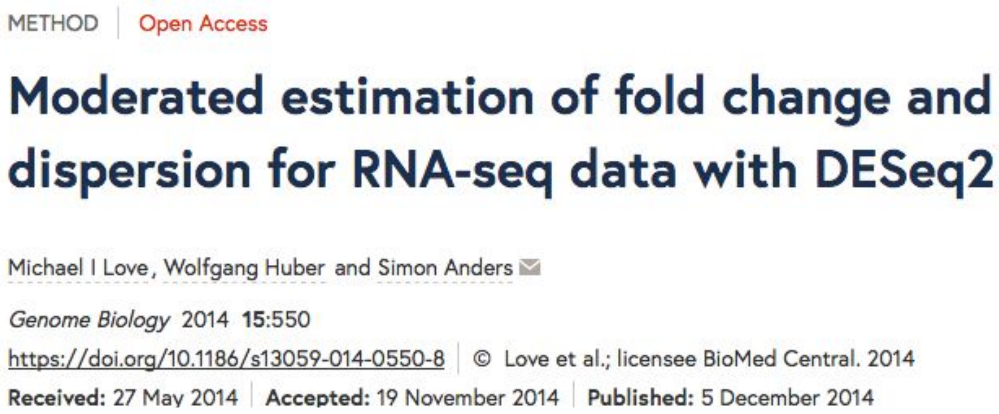
Statistics for High-Dimensional Data

Methods, Theory and Applications

</td>
</tr>
</table>

DESeq2 Overview

- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
    - Microbiome-specific adaptation still open research problem (as far as I am aware)



METHOD | Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber and Simon Anders

Genome Biology 2014 **15**:550

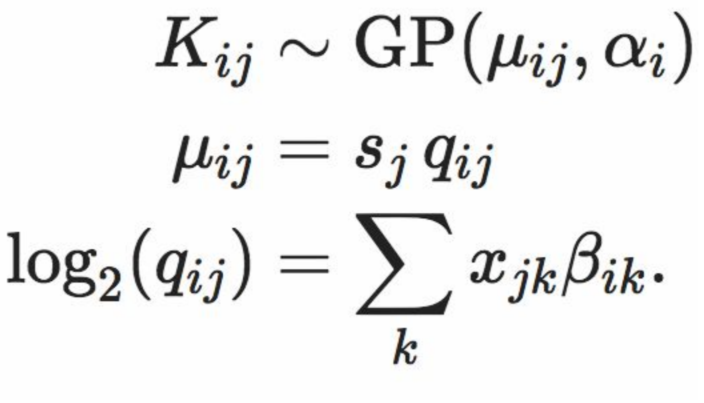
https://doi.org/10.1186/s13059-014-0550-8 | © Love et al.; licensee BioMed Central. 2014

Received: 27 May 2014 | **Accepted:** 19 November 2014 | **Published:** 5 December 2014

DESeq2 Overview

- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem (as far as I am aware)

The underlying math:



$$K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j \, q_{ij}$$

$$\log_2(q_{ij}) = \sum_k x_{jk} \beta_{ik}.$$

DESeq2 Overview

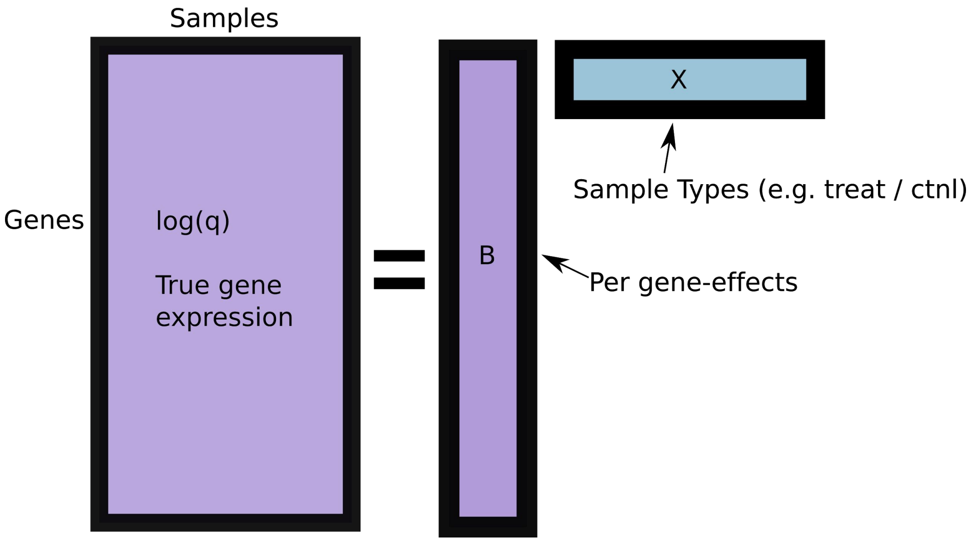
- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

The underlying math:

$$K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j \, q_{ij}$$

$$\boxed{\log_2(q_{ij}) = \sum_k x_{jk}\beta_{ik}.}$$



DESeq2 Overview

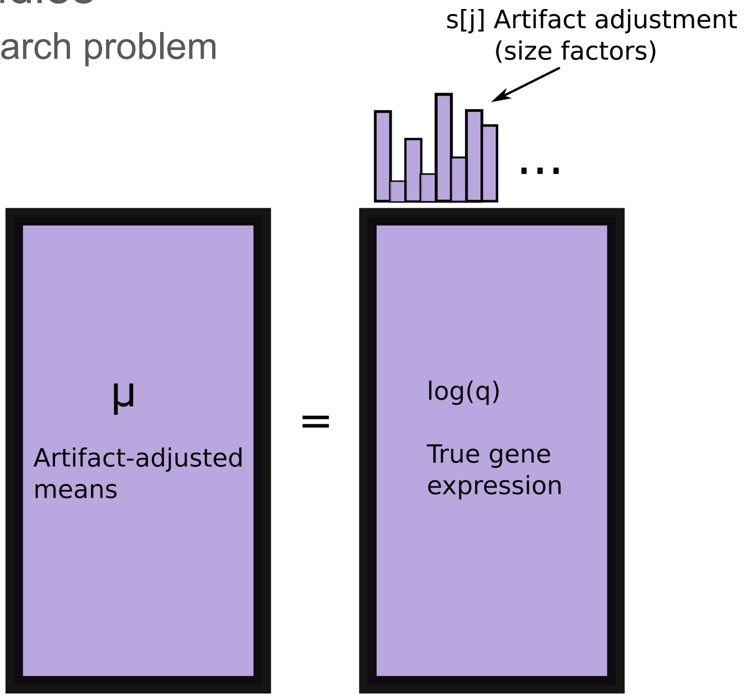
- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

The underlying math:

$$K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)$$

$$\boxed{\mu_{ij} = s_j \, q_{ij}}$$

$$\log_2(q_{ij}) = \sum_k x_{jk} \beta_{ik}.$$



DESeq2 Overview

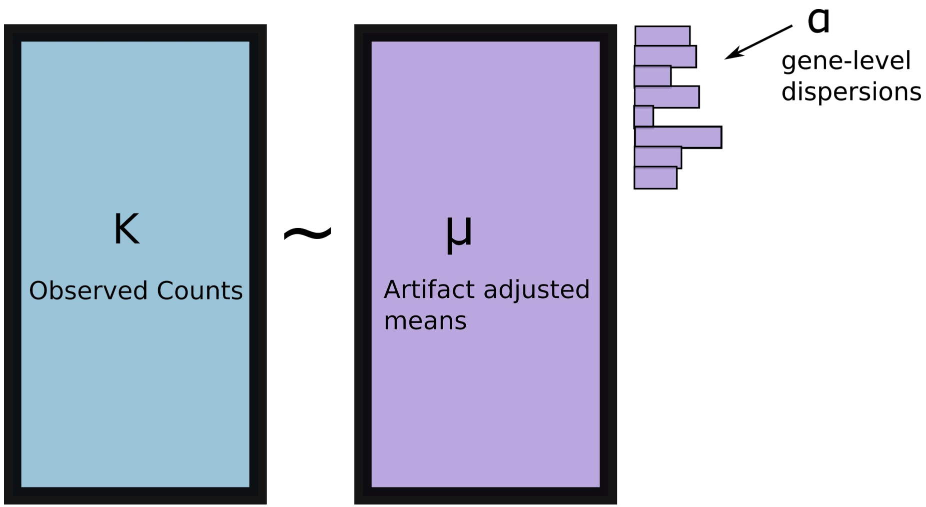
- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

The underlying math:

$$\boxed{K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)}$$

$$\mu_{ij} = s_j \, q_{ij}$$

$$\log_2(q_{ij}) = \sum_k x_{jk} \beta_{ik}.$$



DESeq2 Overview

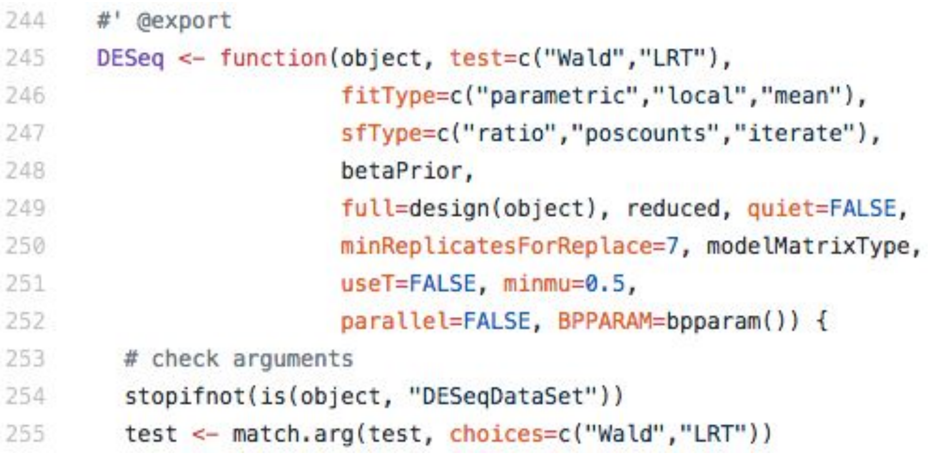
- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
    - Microbiome-specific adaptation still open research problem

The underlying math:

$$K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j \, q_{ij}$$

$$\log_2(q_{ij}) = \sum_k x_{jk} \beta_{ik}.$$



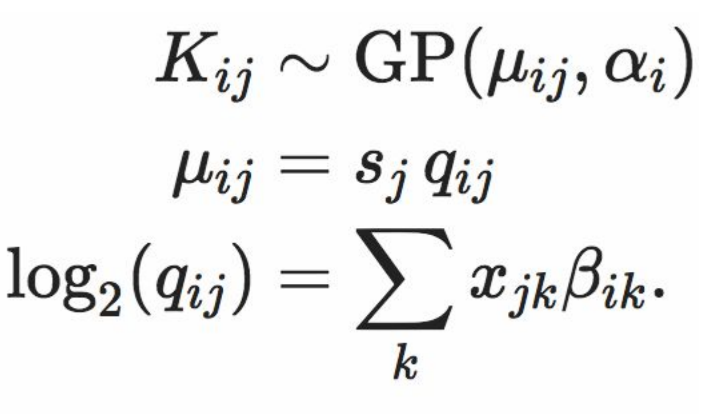
```
#' @export
DESeq <- function(object, test=c("Wald","LRT"),
                  fitType=c("parametric","local","mean"),
                  sfType=c("ratio","poscounts","iterate"),
                  betaPrior,
                  full=design(object), reduced, quiet=FALSE,
                  minReplicatesForReplace=7, modelMatrixType,
                  useT=FALSE, minmu=0.5,
                  parallel=FALSE, BPPARAM=bpparam()) {
  # check arguments
  stopifnot(is(object, "DESeqDataSet"))
  test <- match.arg(test, choices=c("Wald","LRT"))
```

https://github.com/mikelove/DESeq2/blob/master/R/core.R

DESeq2 Overview

- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

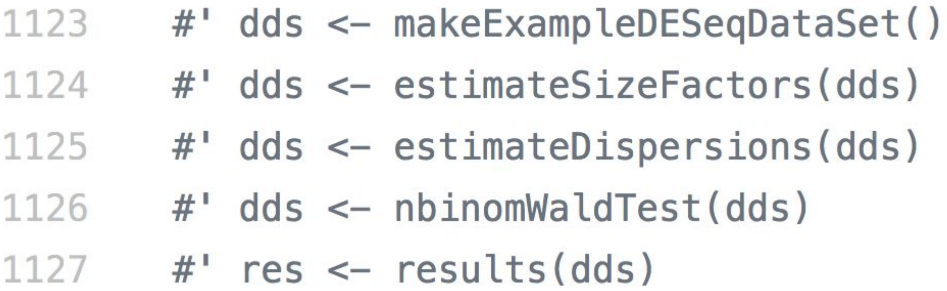
The underlying math:



$$K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j \, q_{ij}$$

$$\log_2(q_{ij}) = \sum_k x_{jk} \beta_{ik}.$$



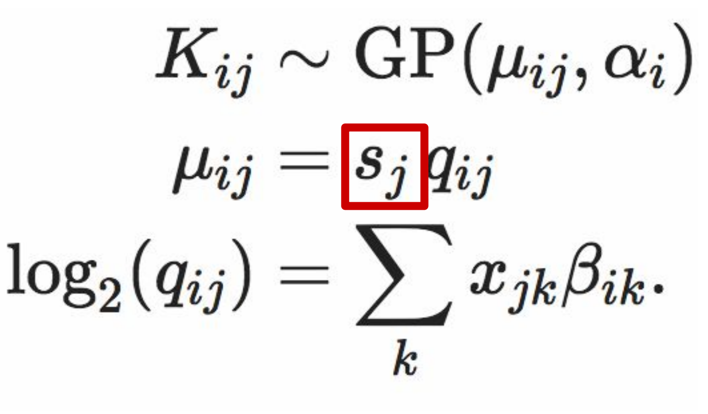
```
#' dds <- makeExampleDESeqDataSet()
#' dds <- estimateSizeFactors(dds)
#' dds <- estimateDispersions(dds)
#' dds <- nbinomWaldTest(dds)
#' res <- results(dds)
```

https://github.com/mikelove/DESeq2/blob/master/R/core.R

DESeq2 Overview

- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

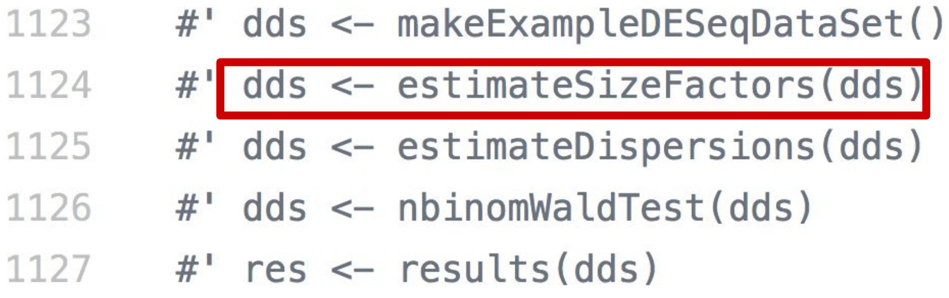
The underlying math:



$$K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = \boxed{s_j} q_{ij}$$

$$\log_2(q_{ij}) = \sum_k x_{jk}\beta_{ik}.$$



```
#' dds <- makeExampleDESeqDataSet()
#' dds <- estimateSizeFactors(dds)
#' dds <- estimateDispersions(dds)
#' dds <- nbinomWaldTest(dds)
#' res <- results(dds)
```

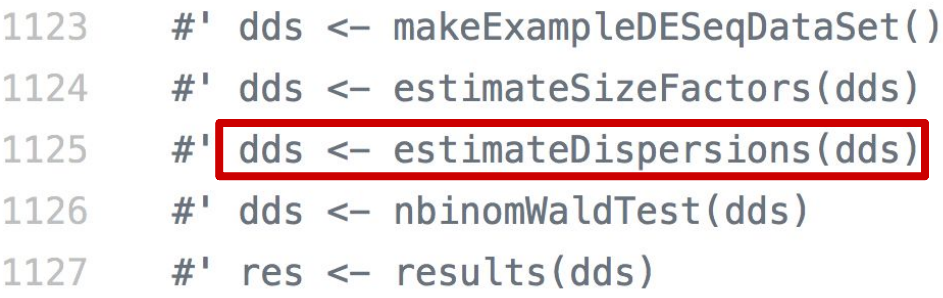
https://github.com/mikelove/DESeq2/blob/master/R/core.R

DESeq2 Overview

- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

The underlying math:

$$
\begin{aligned}
K_{ij} &\sim \mathrm{GP}(\mu_{ij}, \boxed{\alpha_i}) \\
\mu_{ij} &= s_j \, q_{ij} \\
\log_2(q_{ij}) &= \sum_k x_{jk} \beta_{ik}.
\end{aligned}
$$



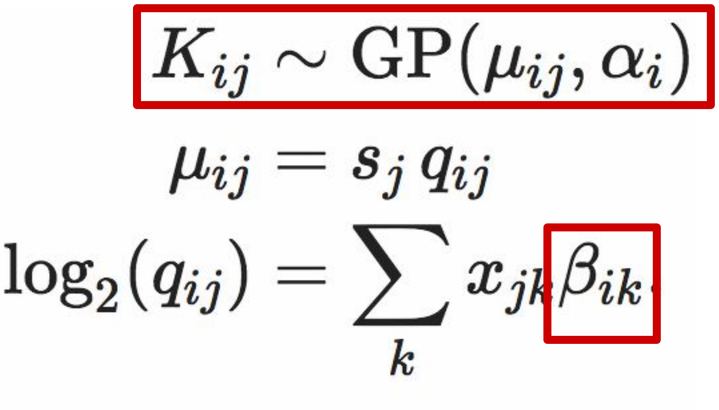
```
#' dds <- makeExampleDESeqDataSet()
#' dds <- estimateSizeFactors(dds)
#' dds <- estimateDispersions(dds)
#' dds <- nbinomWaldTest(dds)
#' res <- results(dds)
```

https://github.com/mikelove/DESeq2/blob/master/R/core.R

DESeq2 Overview

- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

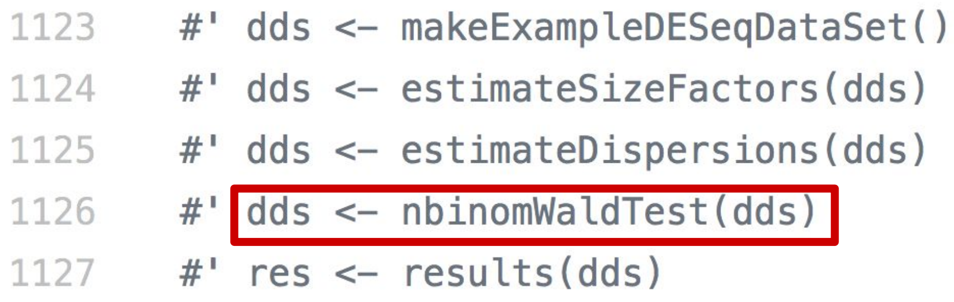
The underlying math:



$$\boxed{K_{ij} \sim \mathrm{GP}(\mu_{ij}, \alpha_i)}$$

$$\mu_{ij} = s_j \, q_{ij}$$

$$\log_2(q_{ij}) = \sum_k x_{jk} \boxed{\beta_{ik}}$$



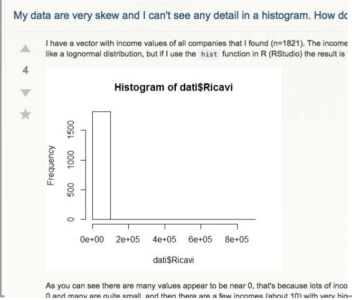
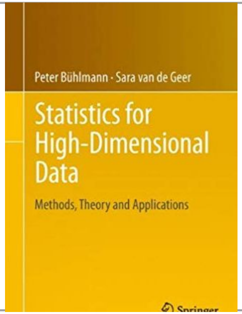
```
#' dds <- makeExampleDESeqDataSet()
#' dds <- estimateSizeFactors(dds)
#' dds <- estimateDispersions(dds)
#' dds <- nbinomWaldTest(dds)
#' res <- results(dds)
```

https://github.com/mikelove/DESeq2/blob/master/R/core.R

DESeq2 Overview

- Method designed for RNA-seq differential expression analysis
- Has been used widely in microbiome studies
  - Microbiome-specific adaptation still open research problem

Let’s try to motivate each component.

Batch Effects (“normalization”)	Count structure / Skewness	High-Dimensionality (few samples + multiple testing)
	 My data are very skew and I can't see any detail in a histogram. How dc 4 I have a vector with income values of all companies that I found (n=1821). The income like a lognormal distribution, but if I use the `hist` function in R (RStudio) the result is Histogram of dati$Ricavi Frequency 0 500 1000 1500 0e+00 2e+05 4e+05 6e+05 8e+05 dati$Ricavi As you can see there are many values appear to be near 0, that's because lots of inco 0 and many are quite small, and then there are a few incomes (about 10) with very hig	 Peter Bühlmann · Sara van de Geer Statistics for High-Dimensional Data Methods, Theory and Applications Springer

Normalization

Why do we need normalization?

- Sources of technical variation resulting from experimental setup
- Confounds true biological variation of interest

Normalization

Why do we need normalization?

- Sources of technical variation resulting from experimental setup
- Confounds true biological variation of interest

Examples

- Differences in sample prep or sequencing protocol
  - Sequencing depth

Normalization

Why do we need normalization?

- Sources of technical variation resulting from experimental setup
- Confounds true biological variation of interest

Examples

- Differences in sample prep or sequencing protocol
    - Sequencing depth
- True biological effects, unrelated to what you care about
    - Age of person sample was collected from

Simple (but problematic) Solutions

- Rarefaction
  - Subsample counts across samples down to the minimum observed in any
- Convert to proportions
  - Divide all samples by their total counts
- Quantile normalization
  - Divide samples by their value at a particular quantile (e.g., 90%)

Why are these problematic?

Simple (but problematic) Solutions

- Rarefaction
  - Subsample counts across samples down to the minimum observed in any
- Convert to proportions
  - Divide all samples by their total counts
- Quantile normalization
  - Divide samples by their value at a particular quantile (e.g., 90%)

Why are these problematic?

Overall counts are informative

- (4, 7, 2) ≠ (400, 700, 200)
- Larger Counts → Lower Uncertainty

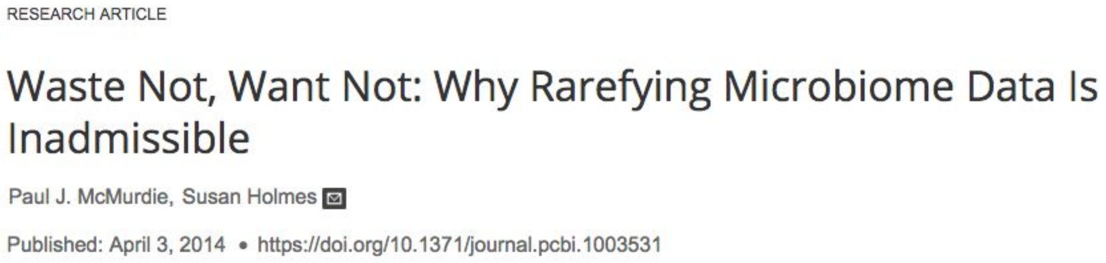
Simple (but problematic) Solutions

- Rarefaction
    - Subsample counts across samples down to the minimum observed in any
- Convert to proportions
    - Divide all samples by their total counts
- Quantile normalization
    - Divide samples by their value at a particular quantile (e.g., 90%)

Why are these problematic?

Overall counts are informative

- (4, 7, 2) ≠ (400, 700, 200)
- Larger Counts → Lower Uncertainty



RESEARCH ARTICLE

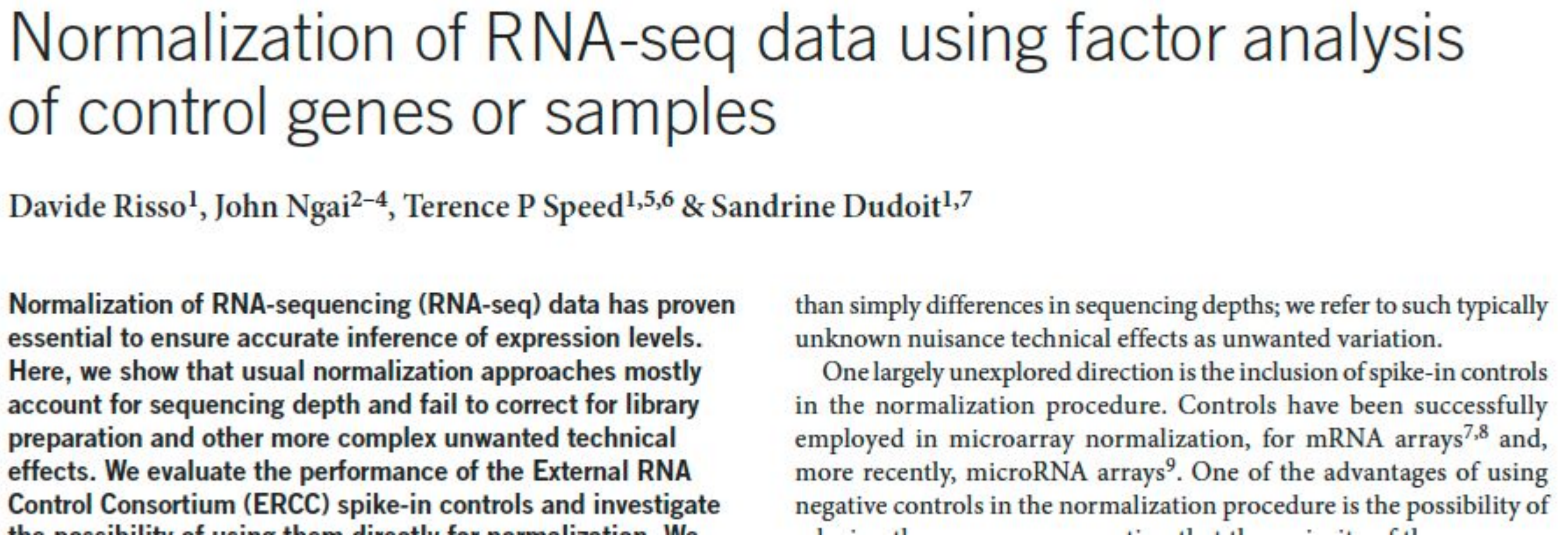
Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org/10.1371/journal.pcbi.1003531

Factors of Technical Variation

General theme: Remove latent factors likely due to technical variation.



Normalization of RNA-seq data using factor analysis of control genes or samples

Davide Risso[1], John Ngai[2–4], Terence P Speed[1,5,6] & Sandrine Dudoit[1,7]

Normalization of RNA-sequencing (RNA-seq) data has proven essential to ensure accurate inference of expression levels. Here, we show that usual normalization approaches mostly account for sequencing depth and fail to correct for library preparation and other more complex unwanted technical effects. We evaluate the performance of the External RNA Control Consortium (ERCC) spike-in controls and investigate the possibility of using them directly for normalization. We

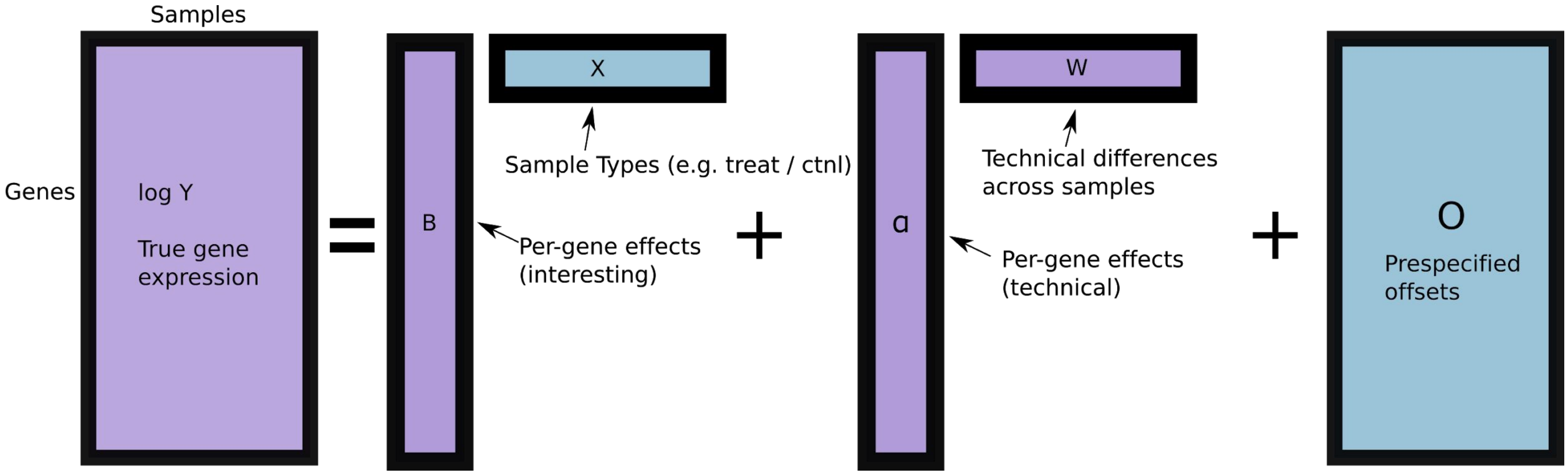
than simply differences in sequencing depths; we refer to such typically unknown nuisance technical effects as unwanted variation.

One largely unexplored direction is the inclusion of spike-in controls in the normalization procedure. Controls have been successfully employed in microarray normalization, for mRNA arrays[7,8] and, more recently, microRNA arrays[9]. One of the advantages of using negative controls in the normalization procedure is the possibility of

Factors of Technical Variation

General theme: Remove latent factors likely due to technical variation.

$$\log E\,[Y|\; W, X, O] = W\alpha + X\beta + O$$


Samples
Genes
log Y
True gene expression
=
B
X
Sample Types (e.g. treat / ctnl)
Per-gene effects (interesting)
+
α
W
Technical differences across samples
Per-gene effects (technical)
+
O
Prespecified offsets

Refinement: Negative Controls

Suppose a gene had two characteristics,

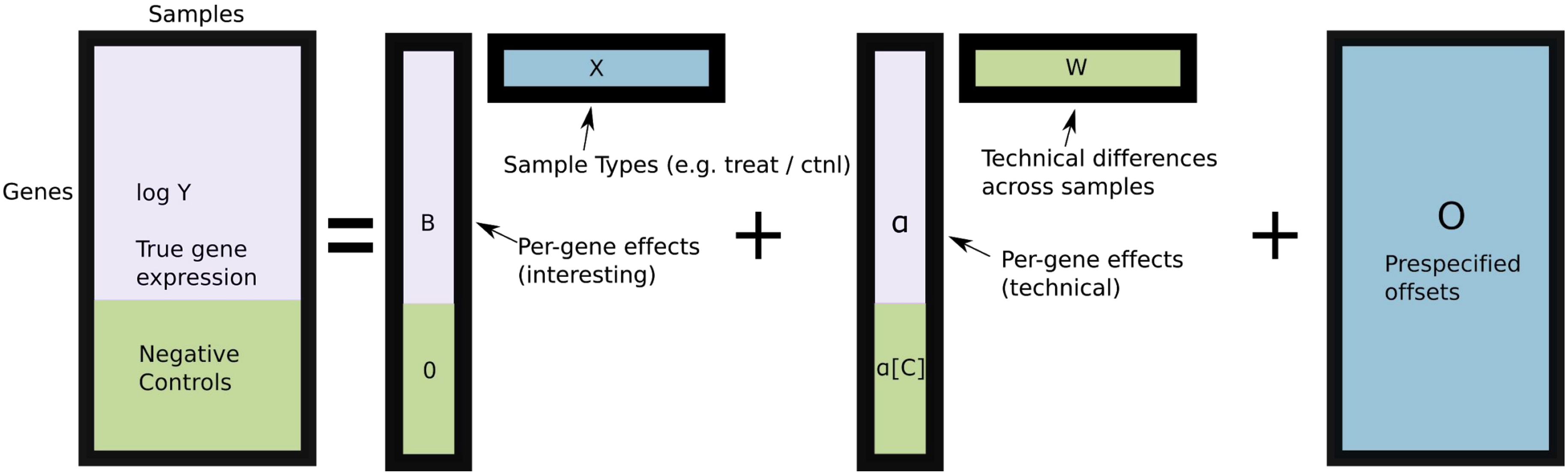
- Gene is unaffected by treatment / control
- Technical variation affects this gene in the same way it affects all others

This gene can be used to “correct” for technical variation in the RUV setup.

Refinement: Negative Controls

Suppose a gene had two characteristics,

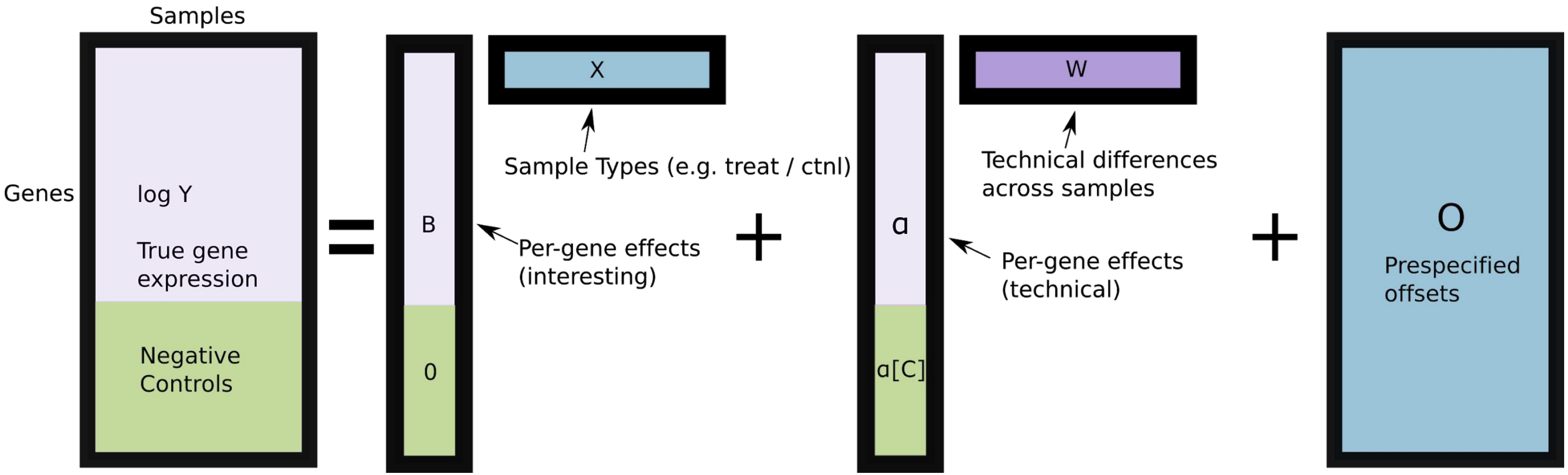
- Gene is unaffected by treatment / control
- Technical variation affects this gene in the same way it affects all others


Samples
Genes
log Y
True gene expression
Negative Controls
=
B
0
X
Sample Types (e.g. treat / ctnl)
Per-gene effects (interesting)
+
ɑ
ɑ[C]
W
Technical differences across samples
Per-gene effects (technical)
+
O
Prespecified offsets

Refinement: Negative Controls

Suppose a gene had two characteristics,

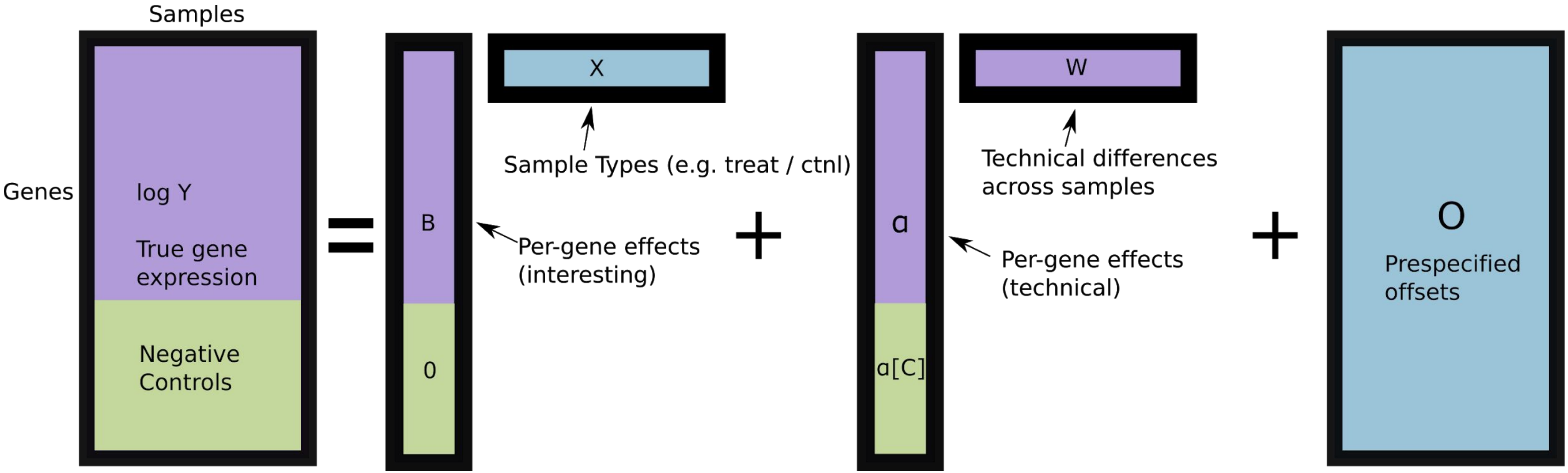
- Gene is unaffected by treatment / control
- Technical variation affects this gene in the same way it affects all others


Samples
Genes
log Y
True gene expression
Negative Controls
=
B
0
X
Sample Types (e.g. treat / ctnl)
Per-gene effects (interesting)
+
α
α[C]
W
Technical differences across samples
Per-gene effects (technical)
+
O
Prespecified offsets

Refinement: Negative Controls

Suppose a gene had two characteristics,

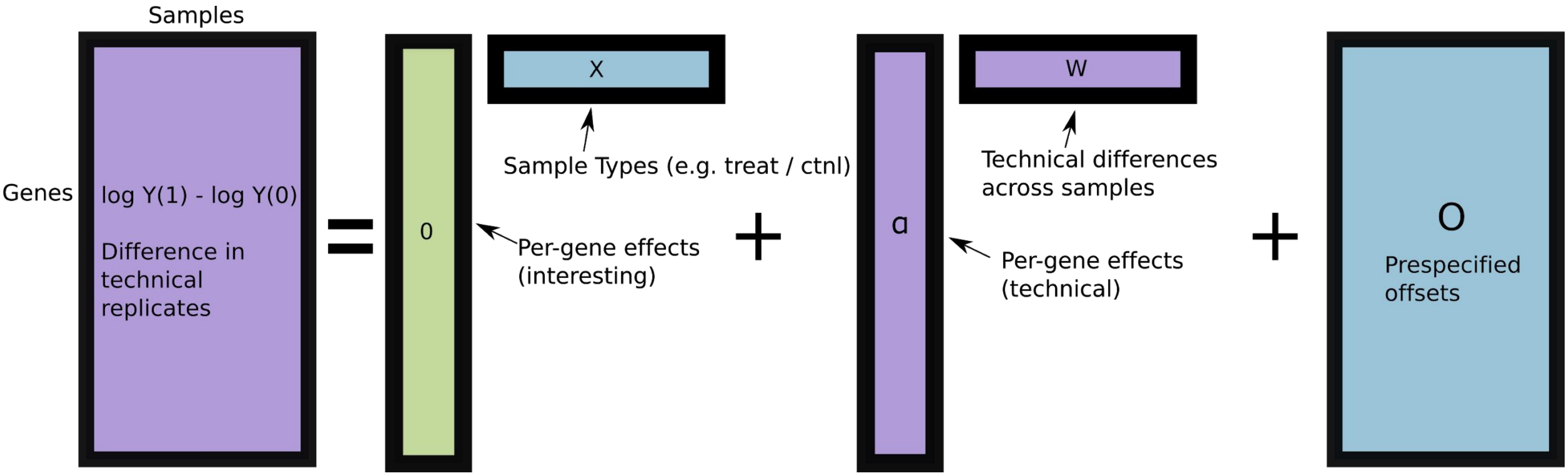
- Gene is unaffected by treatment / control
- Technical variation affects this gene in the same way it affects all others


Samples
Genes
log Y
True gene expression
Negative Controls
=
B
0
X
Sample Types (e.g. treat / ctnl)
Per-gene effects (interesting)
+
α
α[C]
W
Technical differences across samples
Per-gene effects (technical)
+
O
Prespecified offsets

Refinement: Technical Replicates

Negative control correction is sensitive to the choice of control genes.

Alternatively, we can use technical replicates → should exhibit no true variation

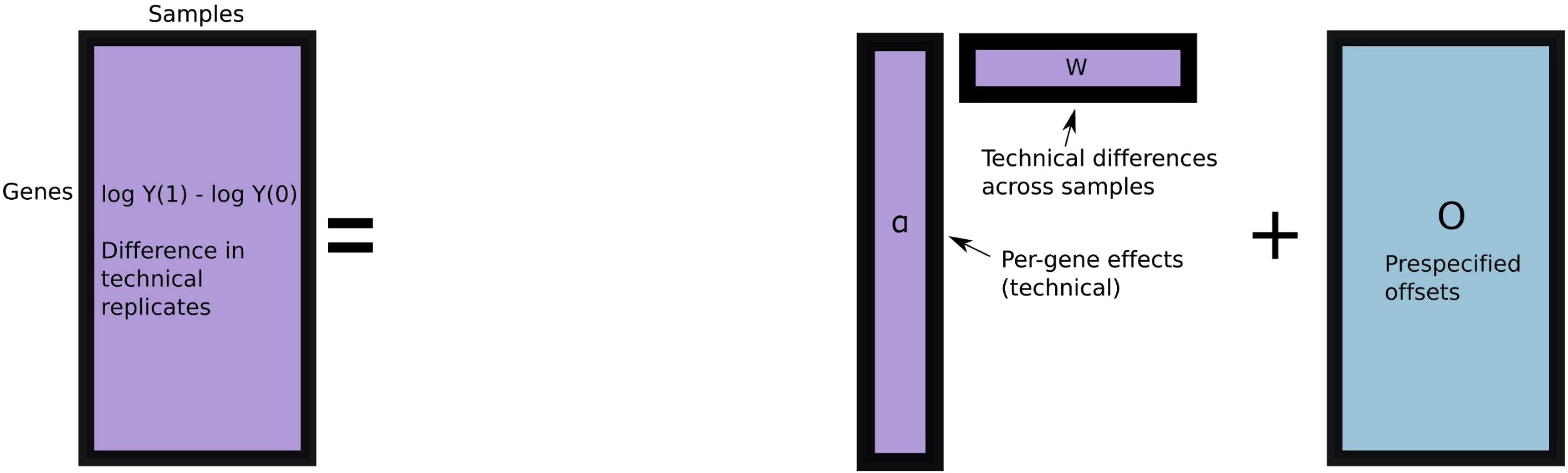


Samples
Genes
log Y(1) - log Y(0)
Difference in technical replicates
=
0
X
Sample Types (e.g. treat / ctnl)
Per-gene effects (interesting)
+
α
W
Technical differences across samples
Per-gene effects (technical)
+
O
Prespecified offsets

Refinement: Technical Replicates

Negative control correction is sensitive to the choice of control genes.

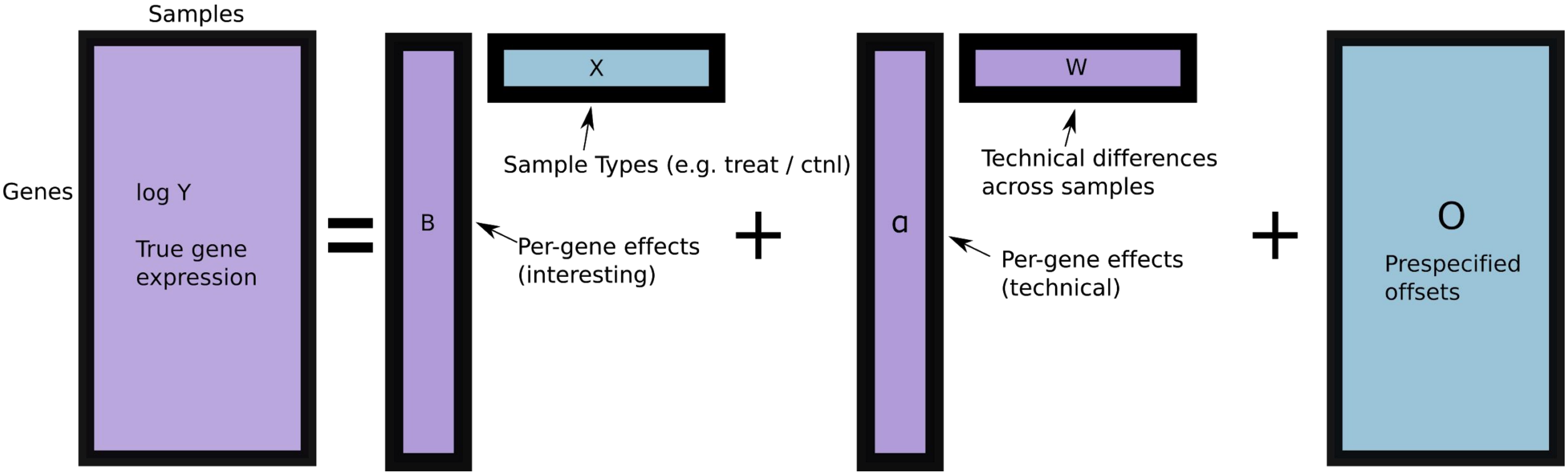
Alternatively, we can use technical replicates → should exhibit no true variation


Samples
Genes
log Y(1) - log Y(0)
Difference in technical replicates
=
α
W
Technical differences across samples
Per-gene effects (technical)
+
O
Prespecified offsets

Refinement: Technical Replicates

Negative control correction is sensitive to the choice of control genes.

Alternatively, we can use technical replicates → should exhibit no true variation


Samples
Genes
log Y
True gene expression
=
B
X
Sample Types (e.g. treat / ctnl)
Per-gene effects (interesting)
+
α
W
Technical differences across samples
Per-gene effects (technical)
+
O
Prespecified offsets

Estimating Factors

- DESeq2 does something closer to upper quantile normalization

$$s_j = \underset{i:\ K_i^{\mathrm{R}} \neq 0}{\operatorname{median}} \frac{K_{ij}}{K_i^{\mathrm{R}}} \quad \text{with} \quad K_i^{\mathrm{R}} = \left(\prod_{j=1}^{m} K_{ij} \right)^{1/m}.$$

Estimating Factors

- DESeq2 does something closer to upper quantile normalization

$$s_j = \underset{i:\ K_i^{\mathrm{R}} \neq 0}{\mathrm{median}} \frac{K_{ij}}{\boxed{K_i^{\mathrm{R}}}} \quad \text{with} \quad K_i^{\mathrm{R}} = \left(\prod_{j=1}^{m} K_{ij}\right)^{1/m}.$$

Typical value for that RSV

Estimating Factors

- DESeq2 does something closer to upper quantile normalization

How much larger is this sample, for that RSV?

$$s_j = \underset{i:\ K_i^{\mathrm{R}} \neq 0}{\operatorname{median}} \boxed{\frac{K_{ij}}{\boxed{K_i^{\mathrm{R}}}}} \quad \text{with} \quad K_i^{\mathrm{R}} = \left(\prod_{j=1}^{m} K_{ij} \right)^{1/m} .$$

Typical value for that RSV

Estimating Factors

- DESeq2 does something closer to upper quantile normalization

How much larger is this sample, across RSVs?

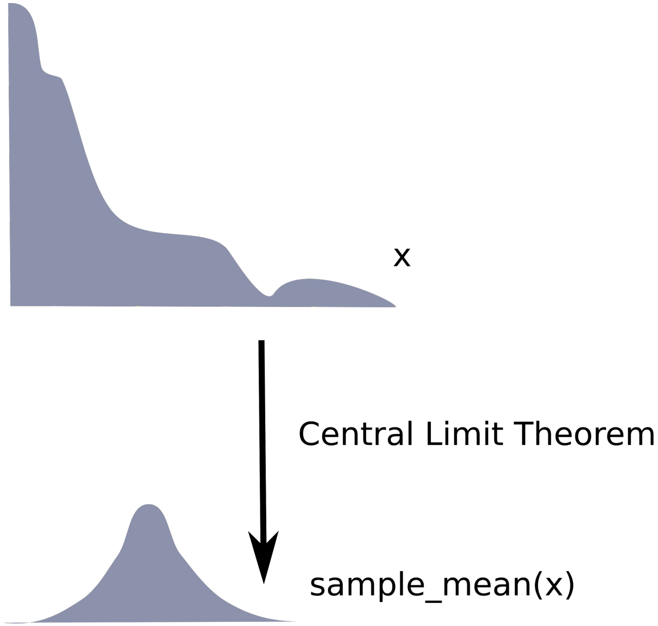
How much larger is this sample, for that RSV?

$$s_j = \boxed{\operatorname*{median}_{i:\ K_i^{\mathrm{R}} \neq 0} \boxed{\frac{K_{ij}}{\boxed{K_i^{\mathrm{R}}}}}} \quad \text{with} \quad K_i^{\mathrm{R}} = \left(\prod_{j=1}^{m} K_{ij}\right)^{1/m}.$$

Typical value for that RSV

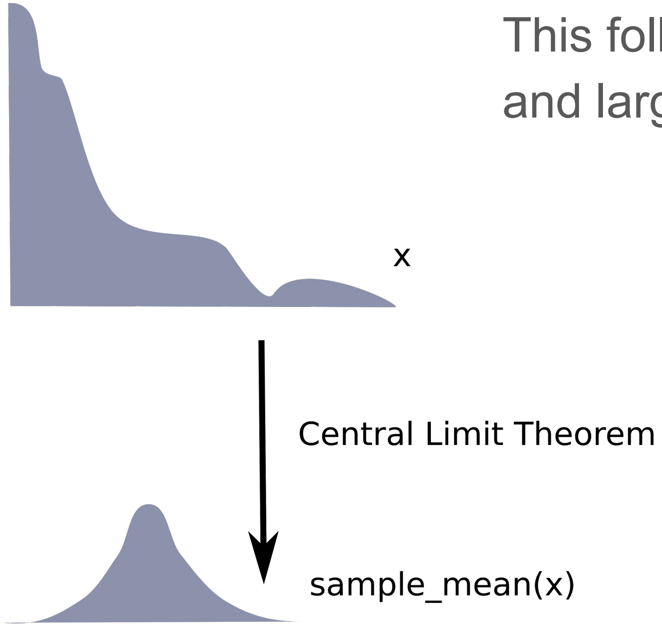
Count Structure (and skewness)

- Misconception: To use a t-test, you need normally distributed data.
- Reality: You only need normality in means



Count Structure (and skewness)

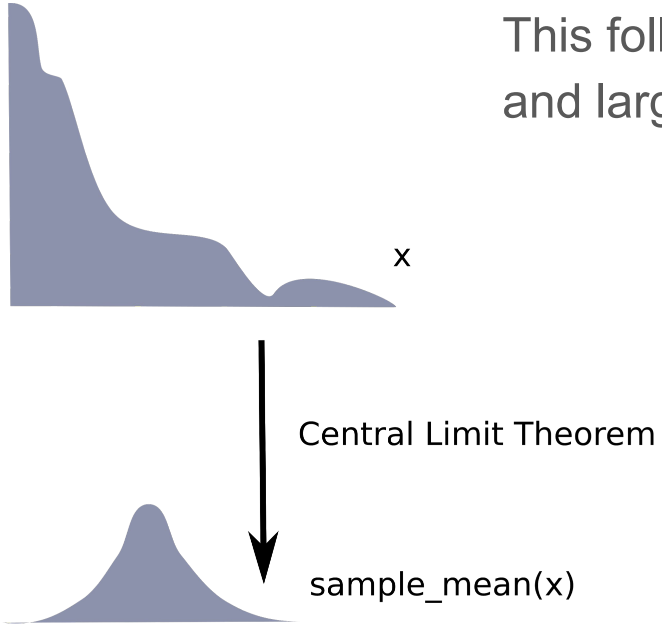
- Misconception: To use a t-test, you need normally distributed data.
- Reality: You only need normality in means



This follows from the central limit theorem and large enough sample sizes.

Count Structure (and skewness)

- Misconception: To use a t-test, you need normally distributed data.
- Reality: You only need normality in means

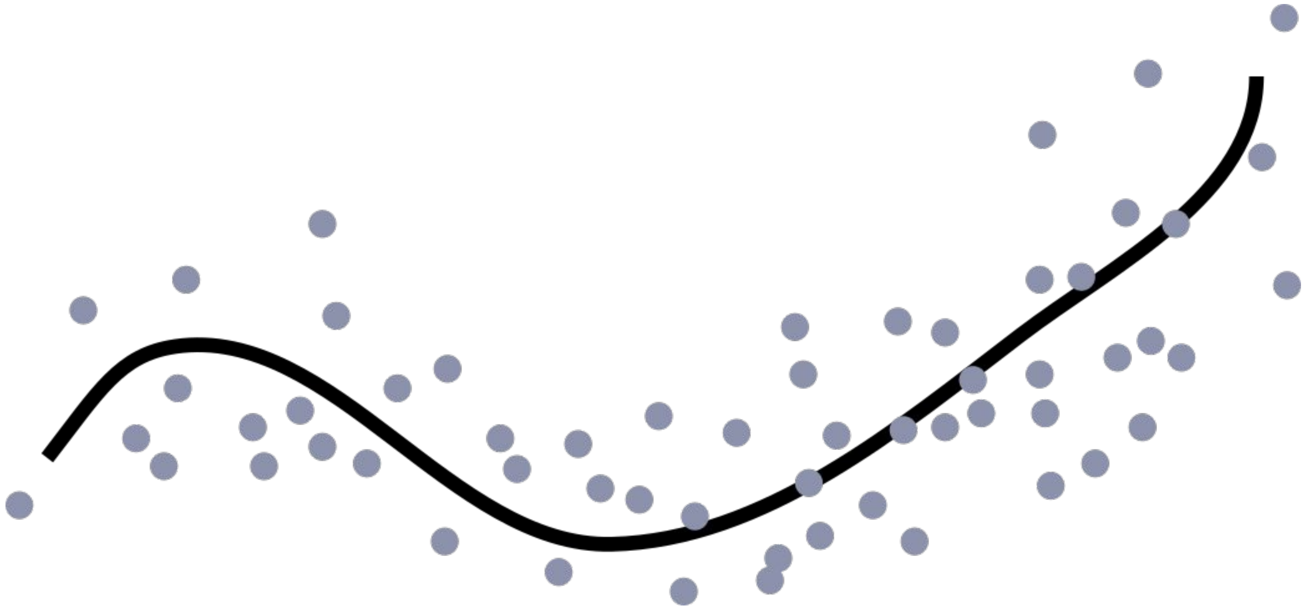


This follows from the central limit theorem and large enough sample sizes.

Fundamental Problem
We usually need covariates (can’t just use two-group means), and need to model the original count data.

Usual Linear Regression

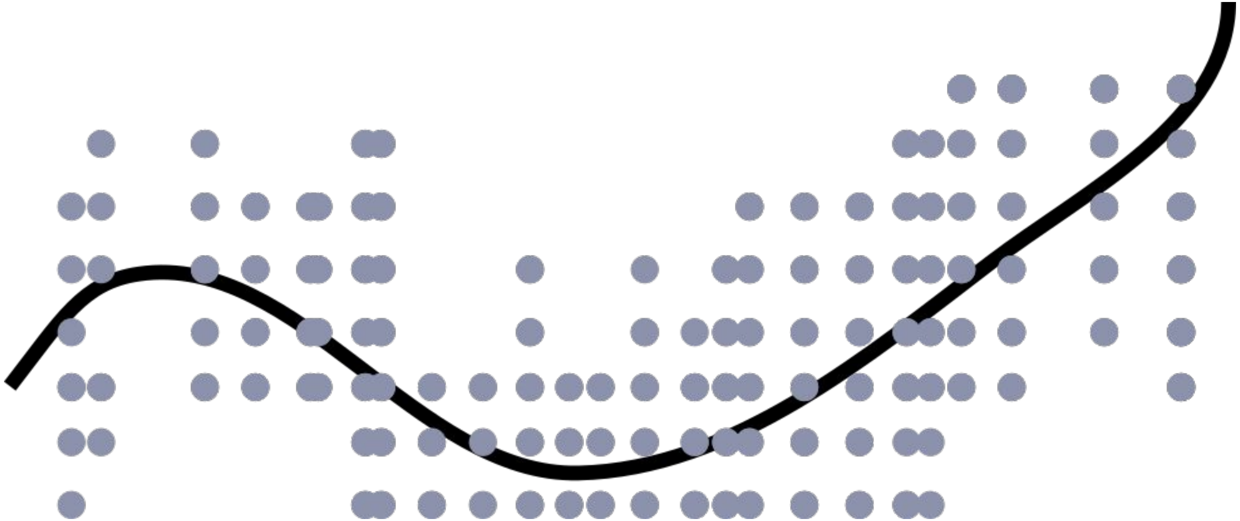
$$y_i \sim N(y_i|\mu(x_i), \sigma^2)$$



Gaussian errors around a regression function.

Usual Linear Regression

$$y_i \sim N(y_i | \mu(x_i), \sigma^2)$$



Gaussian errors around a regression function.

This error structure makes no sense for count data!

Extension: Generalized Linear Models

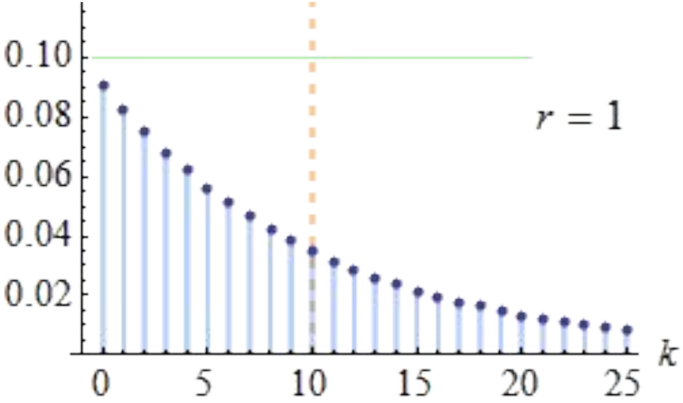
- Generalized linear models extend linear regression to other error structures

$$y_i \sim \mathrm{ExpFam}(y_i | \eta(x_i))$$

- For example, can make the data be Poisson around a regression function
- Inferential theory from linear regression carries over (confidence intervals, prediction intervals, ...)

Overdispersion → Negative Binomial Distn.

- Poisson models tend to underestimate variance (only one parameter)
- A two-parameter alternative is the Negative Binomial distribution
- (also called “Gamma-Poisson”)



Fixed mean, but different amounts of *dispersion*, according to parameter r [from wikipedia]

Aside: Unsupervised Versions

- Active area of research

DESeq2 Dispersion Estimates

Alternative: Transform Counts

- Log transform (with pseudocounts)
- Variance Stabilizing Transformation / Regularized Log
- Advantage: Plug into methods expecting ‘more gaussian’ data
- Disadvantage: Lose probabilistic interpretations

High-Dimensionality

- Lots of RSVs, relatively few samples
- Two very general principles,
    - Share information whenever possible
    - Control the False Discovery Rate

MA Plots

- The usual paradigm

Random Effects Models

- Gelman schools example
- Do similar sharing across genes

Sharing for Dispersions

False Discovery Rate control

- Gene level tests
  - t-tests
  - GLMs
- Correction (picture of sorted p-values)

DESeq2 Summary

F

Conclusion

- Have some powerful tools at your disposal
  - Negative controls
  - Latent factor corrections
  - Count data modeling
  - Information sharing
  - False Discovery Rate Control
- Every new technology needs new normalization and