VU VRIJE UNIVERSITEIT AMSTERDAM

Preprocessing of data

Tricks for RNA-seq and other large data sets

Douwe Molenaar
Systems Biology Lab
January 2023

Outline of the lectures

Lecture 1

- Terminology
- Statistical hypothesis testing
- Count data

Lecture 2

- Variance stabilization
- RNA-seq data analysis
- Multiple hypothesis testing

Lecture 1

Examples of random variables

- Side of a coin after throwing it
- Value of a dice after rolling it
- Values of two dice after rolling them
- Value of a dice after rolling is even or odd

Examples of random variables

- Side of a coin after throwing it
- Value of a dice after rolling it
- Values of two dice after rolling them
- Value of a dice after rolling is even or odd

- Number of births in a year time
- Concentration of glucose in blood
- Concentration of glucose is above or below 6 mM
- Fluorescence of a probe that binds ATP
- Ratio of concentrations of an mRNA in two samples

Random variables: notions and notations

- **Random variable**: a variable X whose value is determined by random processes
  - Discrete:
    - `head, tail`
    - `0, 1, 2, 3, ...`
  - Continuous: Real numbers, positive, negative

Random variables: notions and notations

- **Random variable**: a variable X whose value is determined by random processes
  - Discrete:
    - `head, tail`
    - `0, 1, 2, 3, ...`
  - Continuous: Real numbers, positive, negative
- **Sample space** Ω: Set of all possible outcomes for an experiment
  - Throwing a die: $\Omega = \{1, 2, \ldots, 6\}$
  - Number k of births in a year time: $\Omega = \{0, 1, 2, \ldots \infty\}$
  - Concentration or fluorescence x: $\Omega = \mathbb{R}^+$
  - log-concentration ratio of mRNA x: $\Omega = \mathbb{R}$

Random variables: notions and notations

- **Random variable**: a variable X whose value is determined by random processes
  - Discrete:
    - `head, tail`
    - `0, 1, 2, 3, ...`
  - Continuous: Real numbers, positive, negative
- **Sample space** Ω: Set of all possible outcomes for an experiment
  - Throwing a die: $\Omega = \{1, 2, \ldots, 6\}$
  - Number k of births in a year time: $\Omega = \{0, 1, 2, \ldots \infty\}$
  - Concentration or fluorescence x: $\Omega = \mathbb{R}^+$
  - log-concentration ratio of mRNA x: $\Omega = \mathbb{R}$
- **Event**: "Drawing" or "realizing" a random variable X
  - $X \in \{\text{head}\}$ (abbreviated as $X = \text{head}$)
  - $X \in \{2, 4, 6\}$
  - $X \in [0, 6]$

Distribution functions allow calculation of probabilities

Distribution function: a function that yields the probability of realizing an event

Distribution functions allow calculation of probabilities

Distribution function: a function that yields the probability of realizing an event

Examples

- *Bernoulli distribution* with parameter p: Throwing a coin once:

$$\Omega = \{\text{head}, \text{tail}\} \qquad P(X \in \{\text{head}\}) = p,\ P(X \in \{\text{tail}\}) = 1 - p$$

Distribution functions allow calculation of probabilities

Distribution function: a function that yields the probability of realizing an event

Examples

- *Bernoulli distribution* with parameter p: Throwing a coin once:

$$\Omega = \{\text{head}, \text{tail}\} \qquad P(X \in \{\text{head}\}) = p,\ P(X \in \{\text{tail}\}) = 1 - p$$

- *Binomial distribution* with parameters n and p: Number of heads when throwing a coin n times:

$$\Omega = \{0, 1, 2, \dots n\} \qquad P(X \in \{k\}) = \frac{n!}{x!(n-k)!} p^k p^{n-k}$$

Distribution functions allow calculation of probabilities

Distribution function: a function that yields the probability of realizing an event

Examples

- *Bernoulli distribution* with parameter p: Throwing a coin once:

$$\Omega = \{\text{head}, \text{tail}\} \qquad P(X \in \{\text{head}\}) = p,\ P(X \in \{\text{tail}\}) = 1 - p$$

- *Binomial distribution* with parameters n and p: Number of heads when throwing a coin n times:

$$\Omega = \{0, 1, 2, \dots n\} \qquad P(X \in \{k\}) = \frac{n!}{x!(n-k)!} p^k p^{n-k}$$

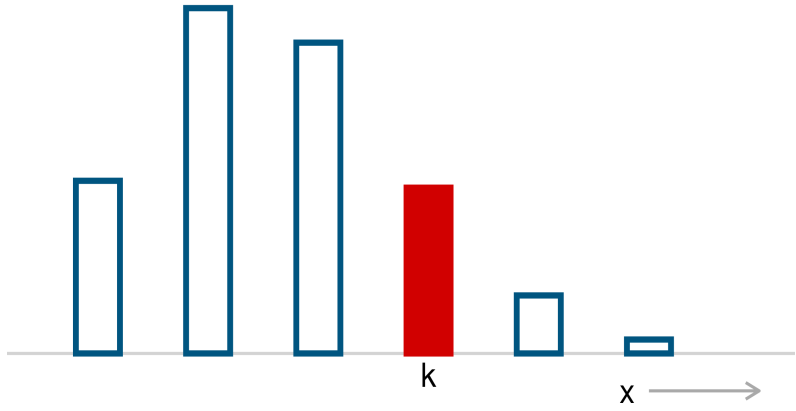
- *Normal distribution* with parameters μ and σ: Concentration or fluorescence:

$$\Omega = \mathbb{R} \qquad P(X \in [a, b]) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\left(\frac{t-\mu}{2\sigma}\right)^2} \mathrm{d}t$$

Distribution functions for discrete and continuous variables

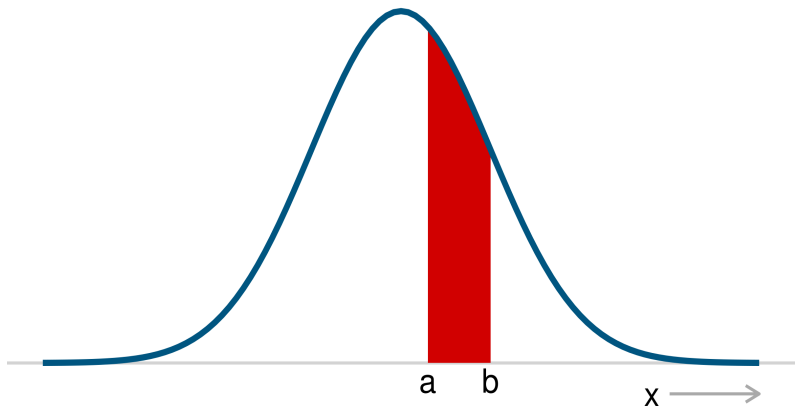
Discrete

- Probability Mass Function (PMF)
- The height is the **probability** $P(X \in \{k\})$



Continuous

- Probability Density Function (PDF)
- The height is the **probability density**
- The probability $P(X \in [a, b])$ is the area under the curve



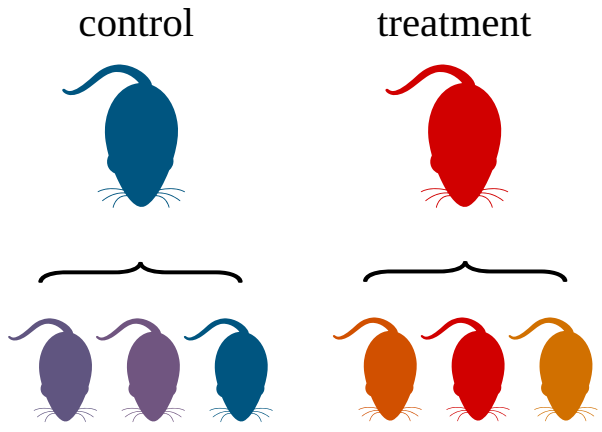
Statistics

Making the best of our lack of understanding the causes of variation

Classifying variation in terms of causes

Biological causes

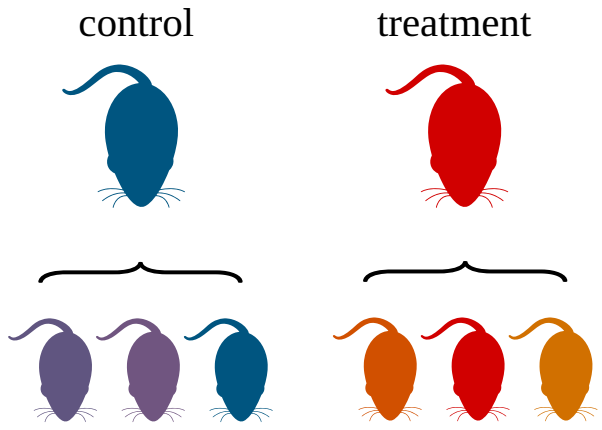
- Variation of interest
  - Response to conditions, treatments
- Variation between similar objects (mice, humans, bacterial cultures)
  - Noteworthy, but usually not the target of the experiment



Classifying variation in terms of causes

Biological causes

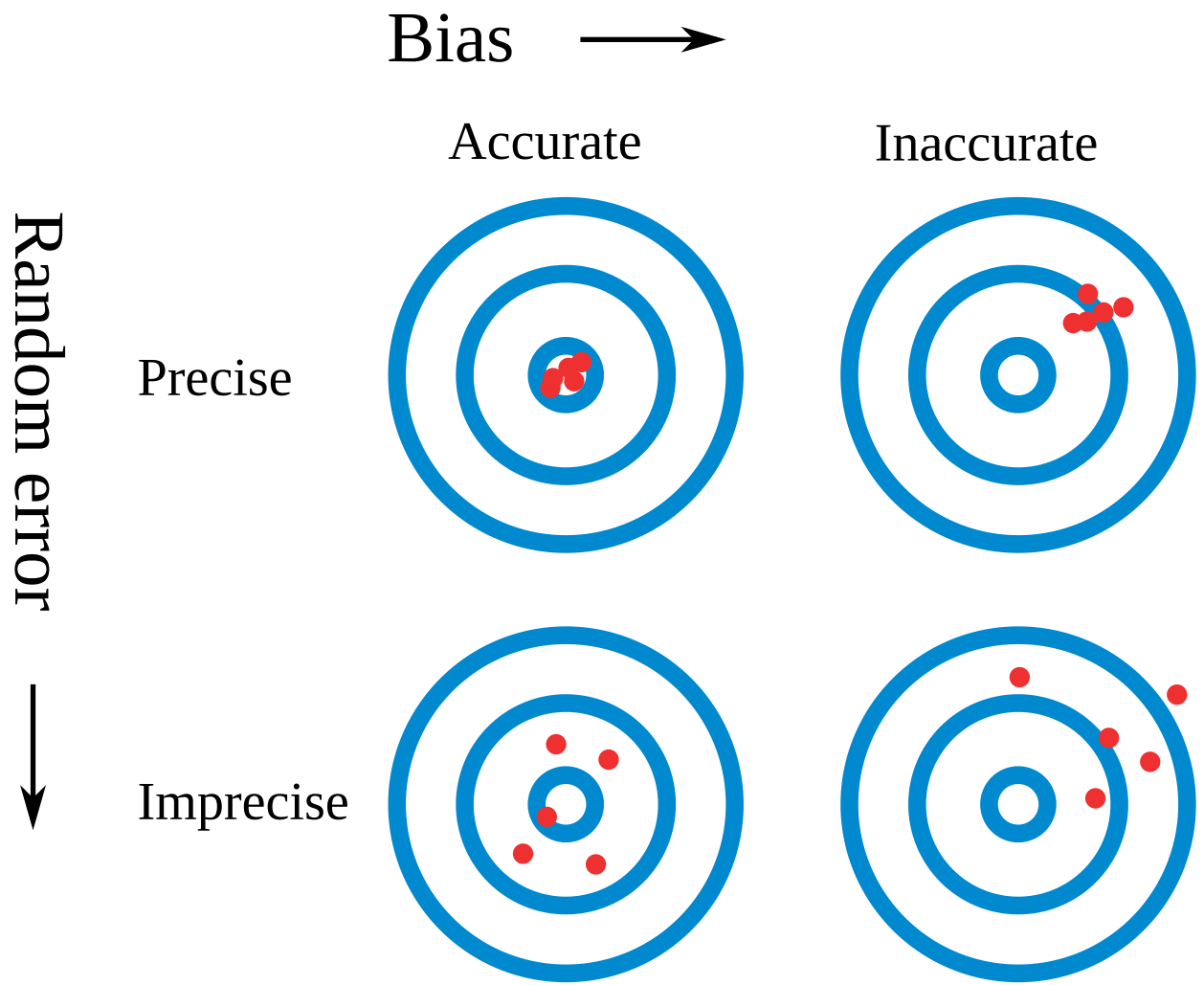
- Variation of interest
  - Response to conditions, treatments
- Variation between similar objects (mice, humans, bacterial cultures)
  - Noteworthy, but usually not the target of the experiment



Technical causes ("errors")

- Variation in sample preparation
- Differences in sample quality
- Differences in amount of sample
- Instrumental or methodical variation

Classifying variation in terms of **accuracy** and **precision**



Handling types of technical variation

Bias / systematic error

1. Improve experimental set-up, instrument settings, sample selection, *etc.* for better **accuracy**
2. Correct the biases (often called **normalization**) *or*
3. Account for the biases in the statistical model

Handling types of technical variation

Bias / systematic error

1. Improve experimental set-up, instrument settings, sample selection, *etc.* for better **accuracy**
2. Correct the biases (often called **normalization**) *or*
3. Account for the biases in the statistical model

Random error / noise

1. Improve experimental set-up, instrument settings, sample selection, *etc.* for better **precision**
2. Investigate the probability distribution of the remaining error
3. Use methods that can handle these probability distributions, like statistical tests

Goal of statistical hypothesis testing

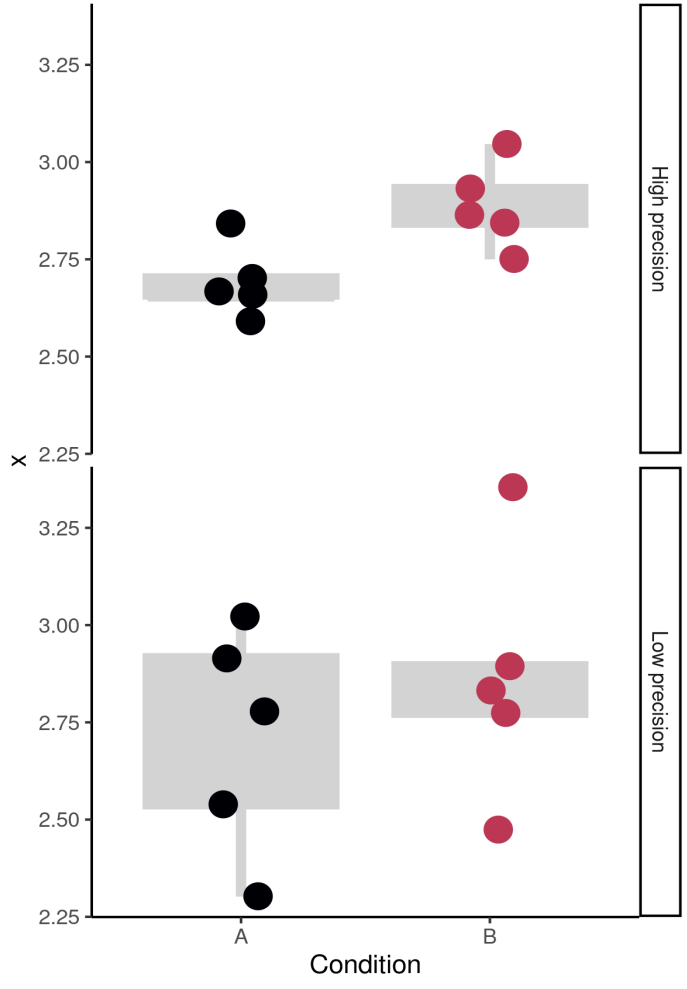
Does a random variable X have different distributions under different conditions?

- The precision or **dispersion** will play a role in the answer
- The answer will be given in terms of probabilities

Goal of statistical hypothesis testing

Does a random variable X have different distributions under different conditions?

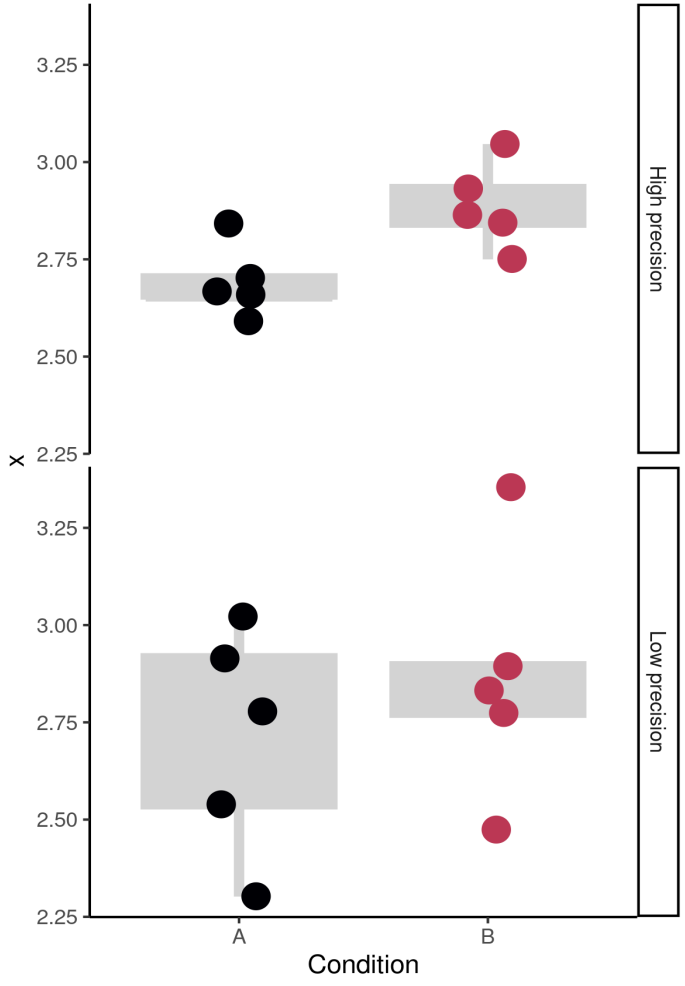
- The precision or **dispersion** will play a role in the answer
- The answer will be given in terms of probabilities



Goal of statistical hypothesis testing

Does a random variable X have different distributions under different conditions?

- The precision or **dispersion** will play a role in the answer
- The answer will be given in terms of probabilities
- Performing T-tests on the examples to the right*:
  - **Low precision**: p-value=0.445
  - **High precision**: p-value=0.016



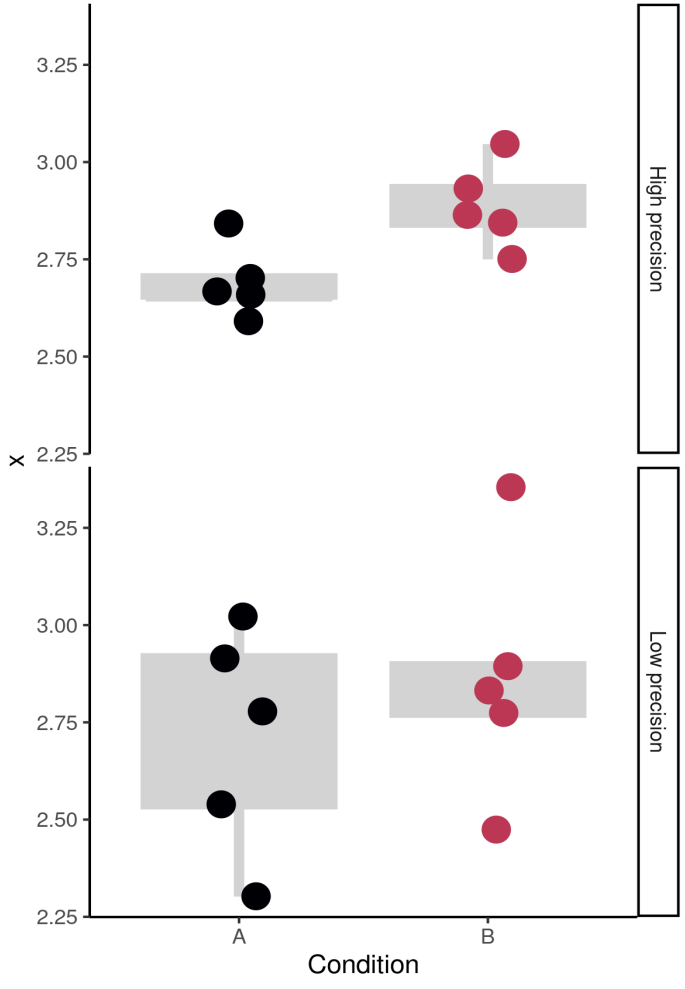
[*] The difference between conditions A and B is 0.3 for both high and low precision

Goal of statistical hypothesis testing

Does a random variable X have different distributions under different conditions?

- The precision or **dispersion** will play a role in the answer
- The answer will be given in terms of probabilities
- Performing T-tests on the examples to the right*:
  - **Low precision**: p-value=0.445
  - **High precision**: p-value=0.016

But what is the meaning of these p-values?



[*] The difference between conditions A and B is 0.3 for both high and low precision

A case of statistical hypothesis testing

We measure concentrations a and b of a metabolite in person under conditions A and B. As the "boring/nothing special" hypothesis we propose the following **Null Hypothesis** or H_0:

There is no difference between concentrations a and b

A case of statistical hypothesis testing

We measure concentrations a and b of a metabolite in person under conditions A and B. As the "boring/nothing special" hypothesis we propose the following **Null Hypothesis** or H_0:

There is no difference between concentrations a and b

More precisely

H_0: a and b are drawn from distributions with the same mean

A case of statistical hypothesis testing

We measure concentrations a and b of a metabolite in person under conditions A and B. As the "boring/nothing special" hypothesis we propose the following **Null Hypothesis** or H_0:

There is no difference between concentrations a and b

More precisely

H_0: a and b are drawn from distributions with the same mean

To be useful in statistical testing we need the following information:

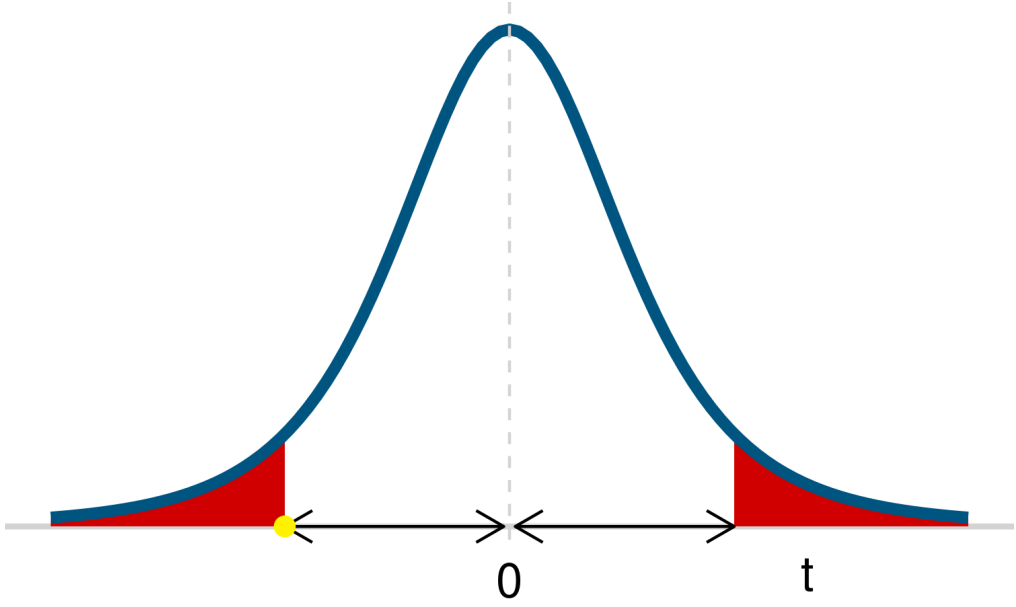
- The type of distribution from which a and b were drawn (*e.g.* a Normal distribution)
- Other parameters of the distribution are known or can be estimated (*e.g.* σ)

Definition of the p-value

The p-value is the probability of measuring the observed value or a more extreme value, if H0 is true

Definition of the p-value

The p-value is the probability of measuring the observed value or a more extreme value, if H0 is true



- If H_0 is true then the probability density of the T-statistic $t = \frac{\bar{a}-\bar{b}}{s_{ab}\sqrt{1/n}}$ is given by the dark-blue curve
- The value obtained for t is indicated by the yellow dot (value is, say, $-d$)
- The p-value of that observation equals the sum of the red areas:

$$\text{p-value} = P(t \in [-\infty, -d] \cup [d, \infty] \mid H_0)$$

Statistical tests *vs.* the truth

H_0	Accept	Reject
True	True negative	False positive
False	False negative	True positive

Statistical tests *vs.* the truth

H_0	Accept	Reject
True	True negative	False positive
False	False negative	True positive

- The rejected cases are the interesting ones* that the statistical test yields, the **positive calls**

[*] Remember: the Null Hypothesis was the "boring" hypothesis

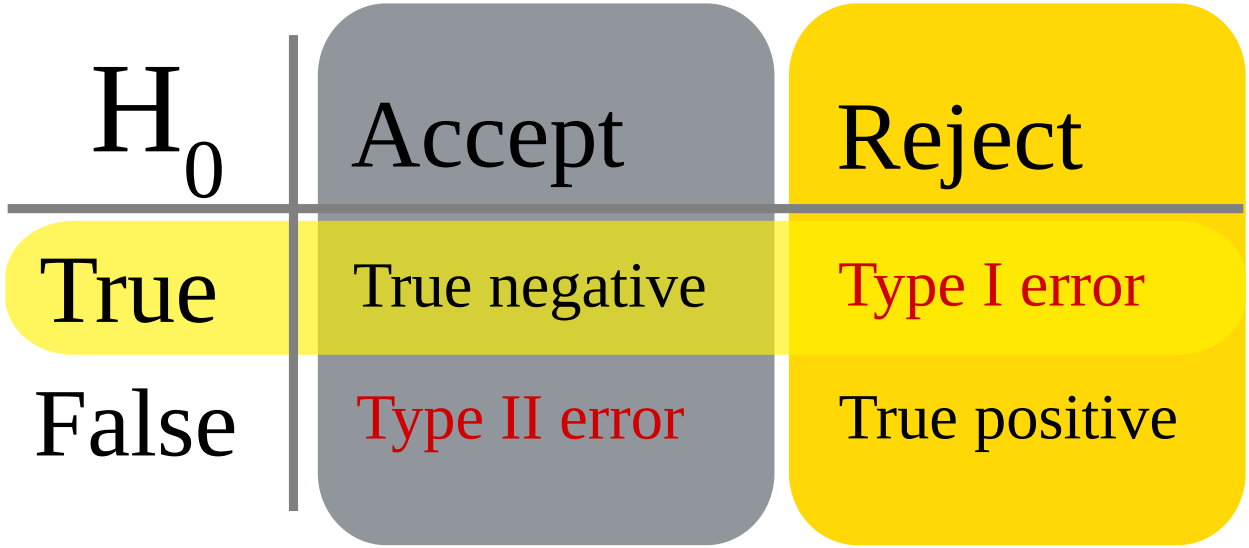
Statistical tests *vs.* the truth

H_0	Accept	Reject
True	True negative	Type I error
False	Type II error	True positive

- The rejected cases are the interesting ones* that the statistical test yields, the **positive calls**
- We also call these False positives and negatives **Type I** and **Type II errors**

[*] Remember: the Null Hypothesis was the "boring" hypothesis

Statistical tests *vs.* the truth



- The rejected cases are the interesting ones* that the statistical test yields, the **positive calls**
- We also call these False positives and negatives **Type I** and **Type II errors**
- The p-value only tells us something about the cases in the first row where H_0 is True

[*] Remember: the Null Hypothesis was the "boring" hypothesis

Recapitulation: What is the meaning of p-value and α?

Definition: The p-value of a statistical test is the probability that the observed or more extreme events happen if the Null Hypothesis is true.

Definition: The α level, or a "p-value cut-off" is the accepted level of a false positive conclusions or **Type I errors**, *i.e.* the probability of rejecting H_0 when it is actually true.

- We call this procedure, using an α level for p-values, **Type I error control**
  - We know that for the true H_0 cases a fraction α will be a false positive

Recapitulation: What is the meaning of p-value and α?

Definition: The p-value of a statistical test is the probability that the observed or more extreme events happen if the Null Hypothesis is true.

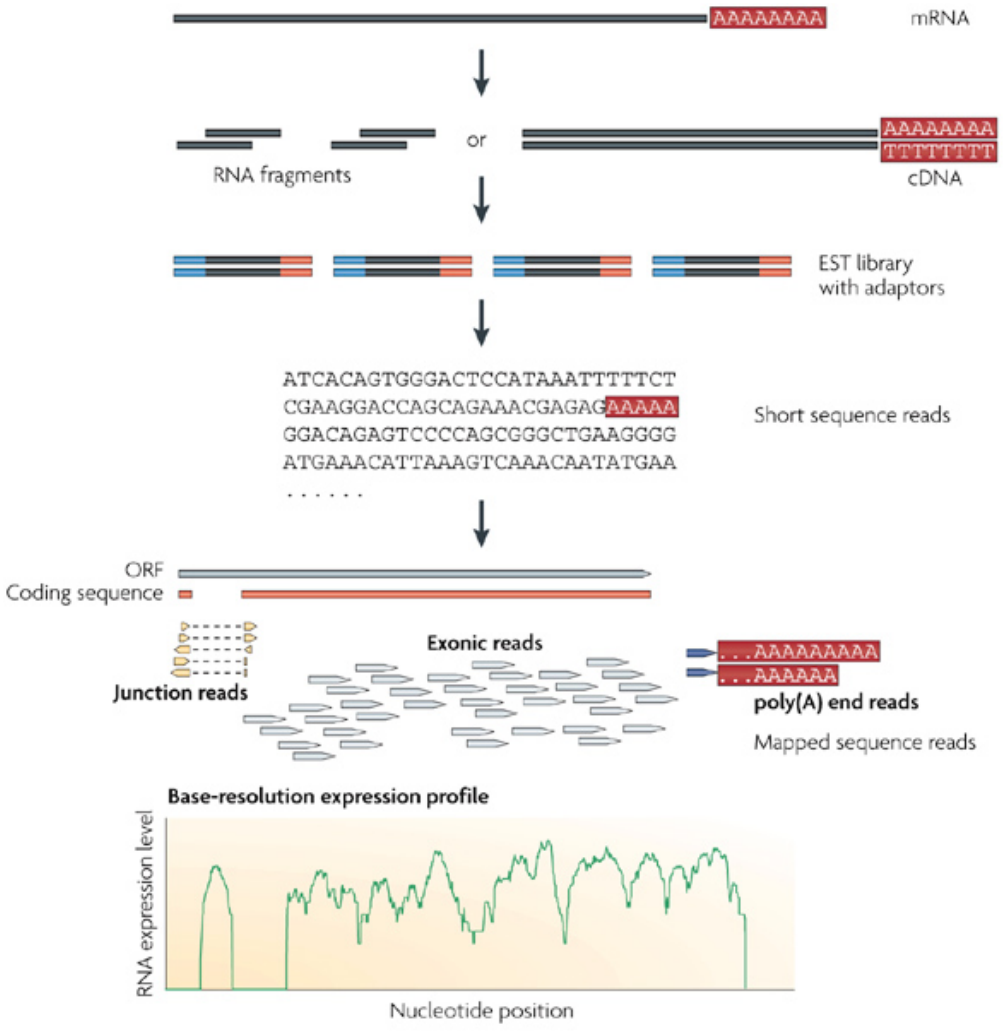
Definition: The α level, or a "p-value cut-off" is the accepted level of a false positive conclusions or **Type I errors**, *i.e.* the probability of rejecting H_0 when it is actually true.

- We call this procedure, using an α level for p-values, **Type I error control**
  - We know that for the true H_0 cases a fraction α will be a false positive

What does the p-value tell about the probability that we have true positive?

Count data

RNA sequencing and mapping sequences: RNA-seq

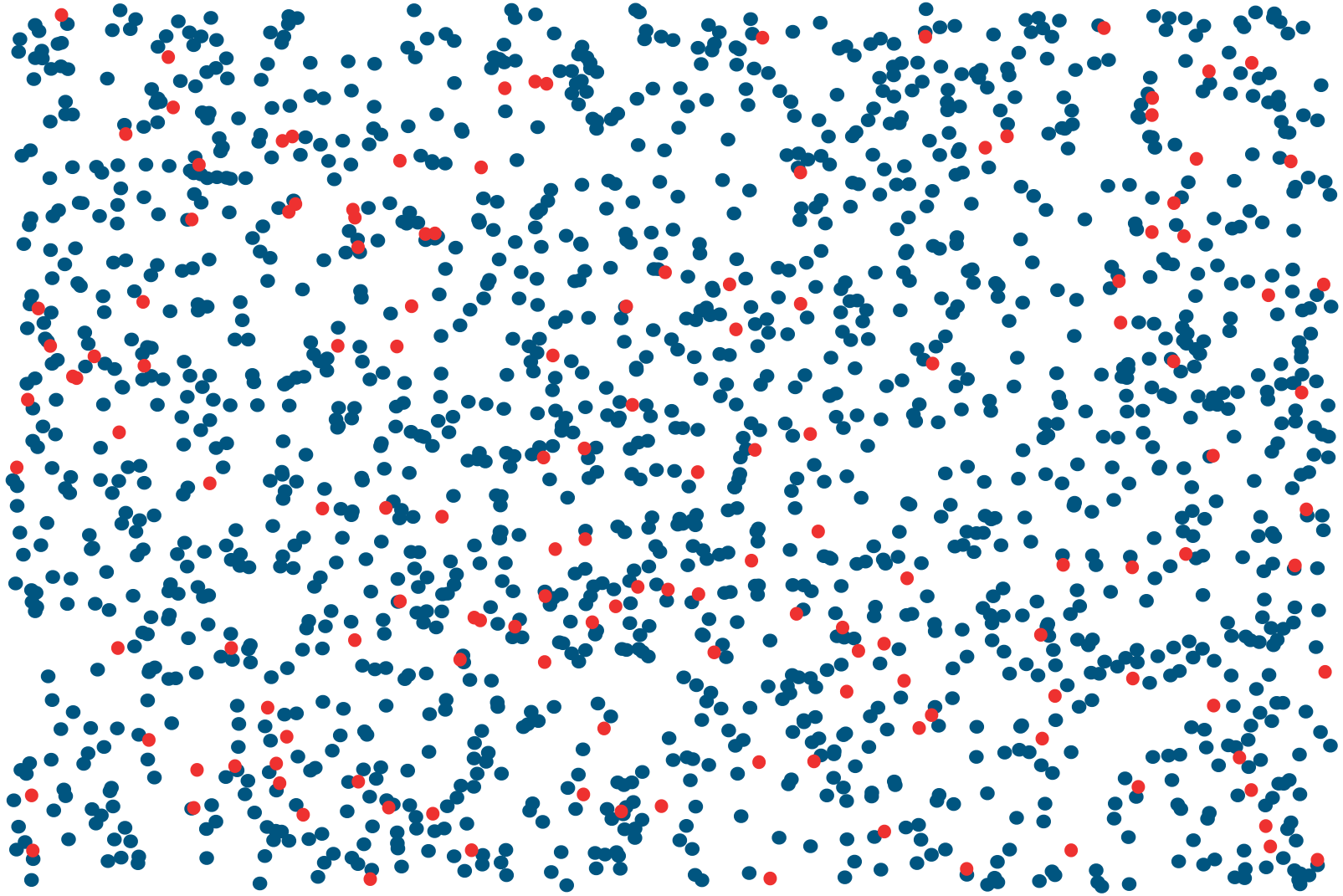


From Wang *et al.* 2009

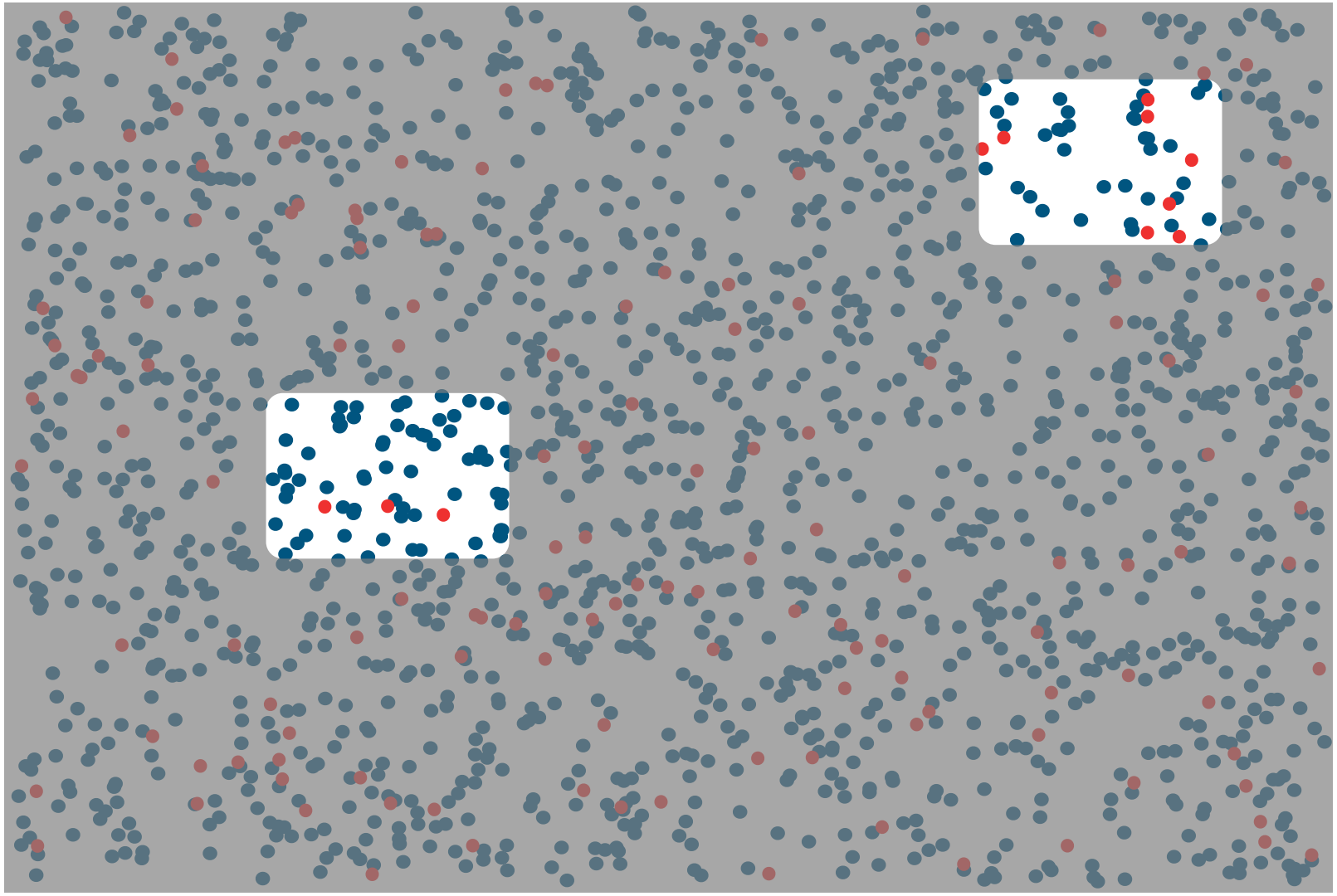
Nature Reviews | Genetics

- Isolate messenger RNA (mRNA)
- Make DNA copies of the RNA
- Add adaptors for sequencing
- Read sequences of fragments
- Map sequences to genes (ORF's)

mRNA molecules: those of the *red* gene are indicated



Random variation of *red* counts in sequenced samples



Example of a table of RNA-seq counts

Five rows from the *Pasilla* data set*

gene_id	untreated1	untreated2	untreated3	untreated4	treated1	treated2	treated3
...	...	...	...	...	...	...	...
FBgn0020369	3387	4295	1315	1853	4884	2133	2165
FBgn0020370	3186	4305	1824	2094	3525	1973	2120
FBgn0020371	1	0	1	1	1	0	0
FBgn0020372	38	84	29	28	63	28	27
...	...	...	...	...	...	...	...
TOTAL COUNT	13972512	21911438	8358426	9841335	18670279	9571826	10343856

Notice:

- Replicate samples of treated and untreated conditions
- The variation in total count between replicates
- The variation in counts per gene

[*] See Huber & Holmes, Ch. 8

Bias in RNA-seq counts: obvious effects

Variation in sequence fragments counted due to **technical variation**:

- Variation in amount of isolated mRNA
- Variation in quality of isolated mRNA
- Variation in cDNA synthesis efficiency
- Variation in sequencing efficiency

These all lead to variation in **sequencing depth** or **total count** between samples.

Normalization of total counts

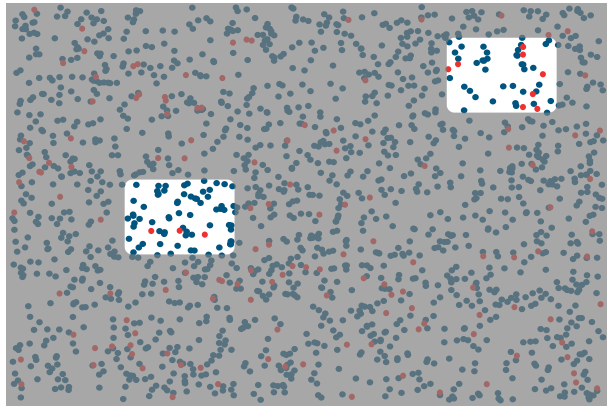
- Simple method
  - Assuming that variation is due to technical reasons and not because concentrations of all mRNA actually differ:
  - **Solution**: divide sequence count for each gene by the total count in the sample
  - **Yields**: RPM, **Rate Per Million** reads

Normalization of total counts

- Simple method
  - Assuming that variation is due to technical reasons and not because concentrations of all mRNA actually differ:
  - **Solution**: divide sequence count for each gene by the total count in the sample
  - **Yields**: RPM, **Rate Per Million** reads
- Advanced method
  - Assume that only a limited number of mRNA differ between samples due to technical or biological reasons
  - **Solution**: perform robust linear regression*
  - **Yields**: **Size factor** for each sample

[*] See Huber & Holmes, Ch. 8

Expected properties of random count variable X for *red*



1. It has a discrete distribution, because we **count** reads per target gene
2. The probability p of sequencing a fragment of gene *red* by randomly picking one from all fragments equals its proportion among all fragments
3. The probability p remains constant after the first, second, *etc.* fragment of *red* has been sequenced, because the number of sequenced fragments $\ll$ available DNA fragments

Expected distribution for RNA-seq counts

- We "draw" a very large sample n (millions) from RNA fragments
- The fraction of fragments p from gene *red* remains constant (does not decrease by earlier draws)

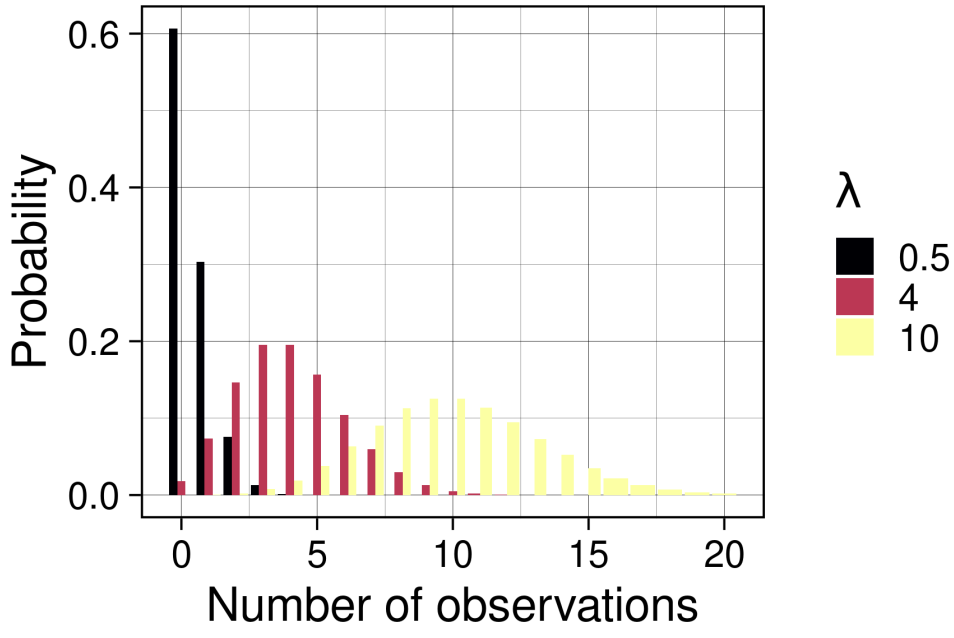
The distribution of X can be approximated by the **Poisson distribution** with parameter λ:

$$P(X = k) \approx \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda = p \times n$ is the mean expected number of counts.

Characteristics of the Poisson distribution

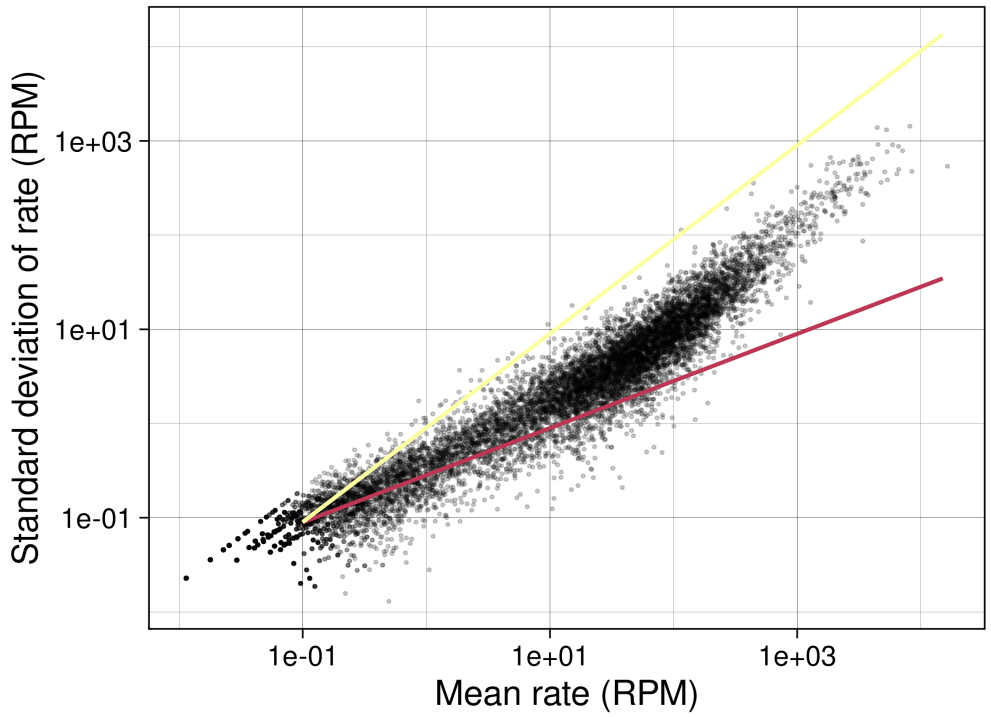
Shape for different values of λ



Mean equals variance

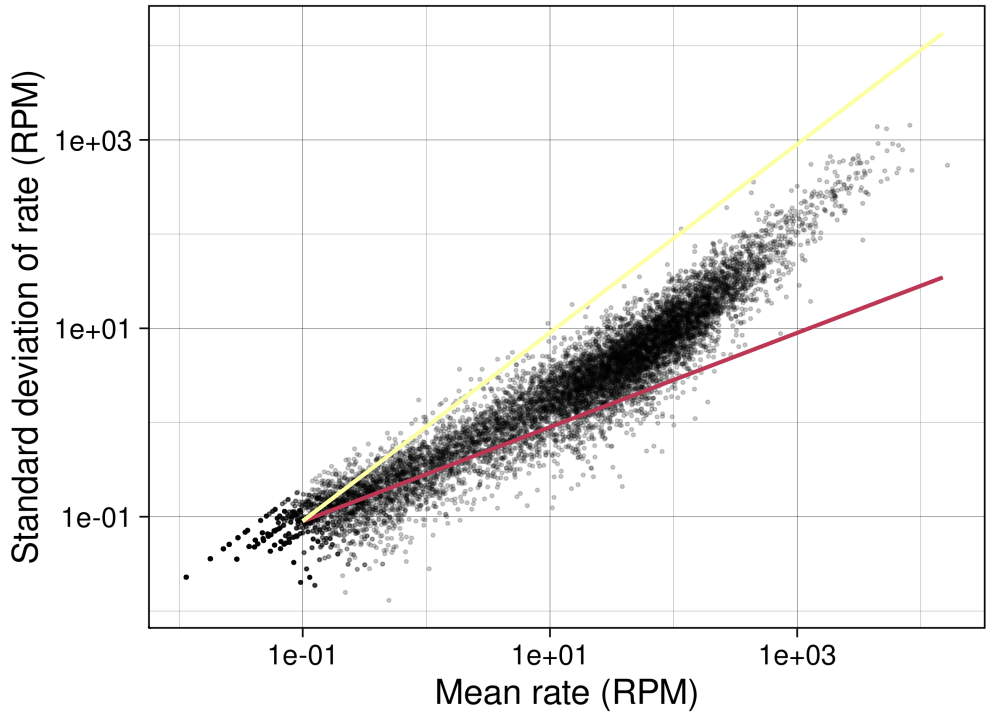
$$\lambda = \mu = \sigma^2 \qquad \rightarrow \sigma = \sqrt{\mu}$$

However: RNA-seq counts are not Poisson-distributed



- The 4 Pasilla *untreated* RNA-seq replicates were used
- The axes have logarithmic scales

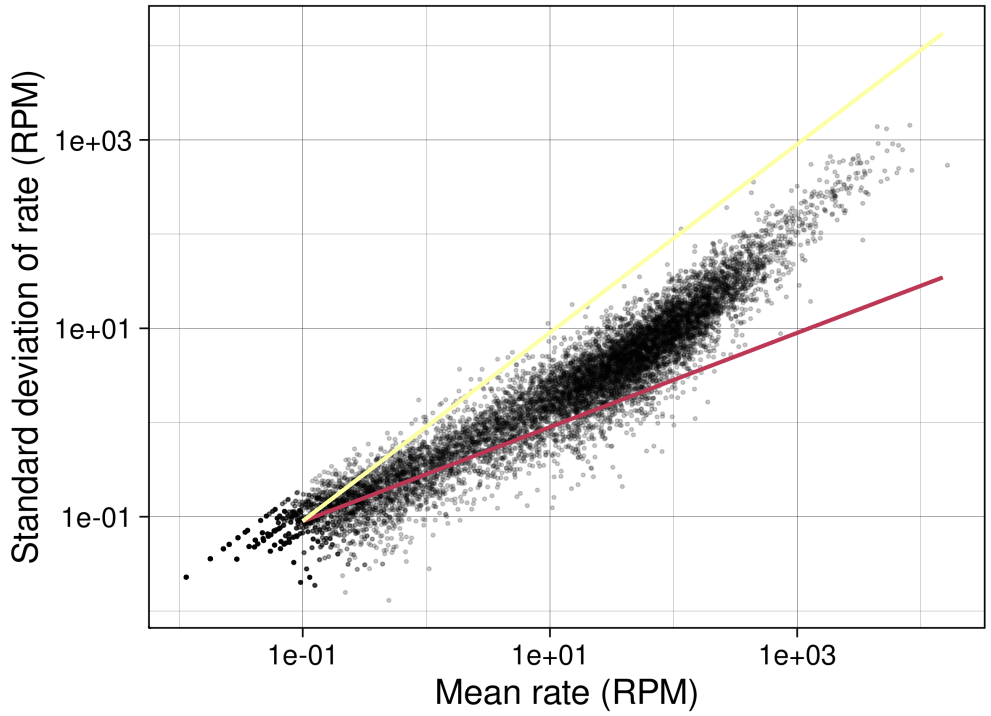
However: RNA-seq counts are not Poisson-distributed



- The 4 Pasilla *untreated* RNA-seq replicates were used
- The axes have logarithmic scales

- The yellow line: $sd(X) \propto (\overline{X})^{1}$, the red line: $sd(X) \propto (\overline{X})^{\frac{1}{2}}$

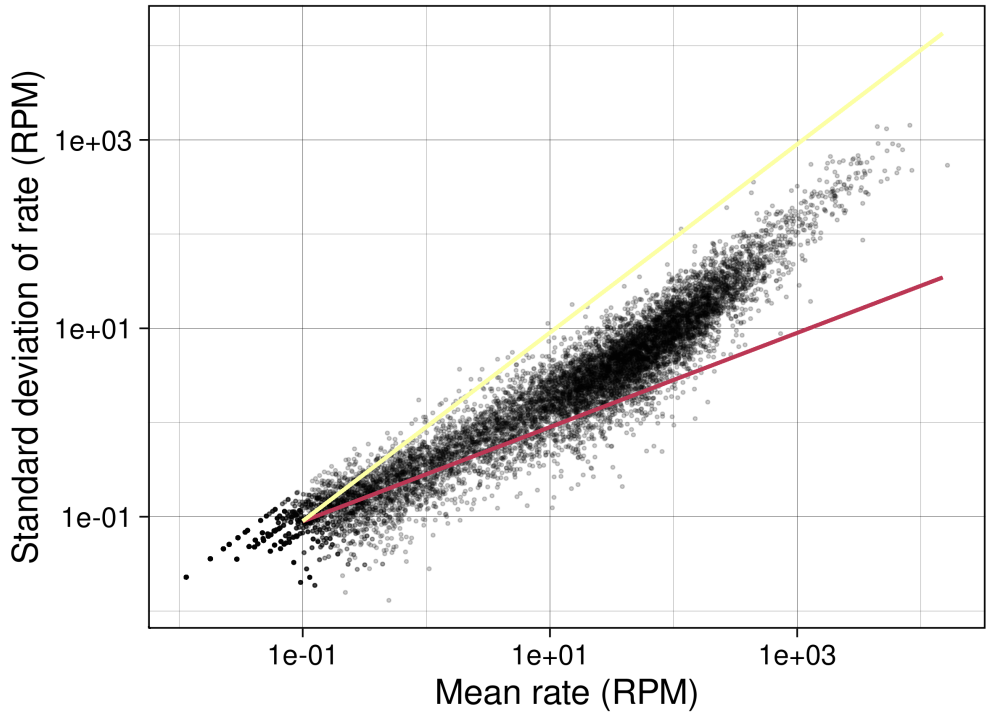
However: RNA-seq counts are not Poisson-distributed



- The 4 Pasilla *untreated* RNA-seq replicates were used
- The axes have logarithmic scales

- The yellow line: $sd(X) \propto (\overline{X})^1$, the red line: $sd(X) \propto (\overline{X})^{\frac{1}{2}}$
- For a Poisson-distributed random variable $\sigma = \sqrt{\mu}$, hence $sd(X) \propto (\overline{X})^{\frac{1}{2}}$

However: RNA-seq counts are not Poisson-distributed



- The 4 Pasilla *untreated* RNA-seq replicates were used
- The axes have logarithmic scales

- The yellow line: $sd(X) \propto (\overline{X})^1$, the red line: $sd(X) \propto (\overline{X})^{\frac{1}{2}}$
- For a Poisson-distributed random variable $\sigma = \sqrt{\mu}$, hence $sd(X) \propto (\overline{X})^{\frac{1}{2}}$
- The observed slope is larger than $\frac{1}{2}$, *i.e.* $sd(X)$ increases more rapidly than expected (**over-dispersion**)

Lecture 2

Variance stabilization

What is variance?

Empirical

- Measure replicates of the value of a variable X, *i.e.* x_i
- The variance is the average squared deviation from the mean; a measure of the spread or dispersion of the measurements:

$$var(X) = \frac{1}{n-1}\sum_{x=1}^{n}\left(x_i - \overline{x}\right)^2$$

- The standard deviation is the square root of the variance $sd(X) = \sqrt{var(X)}$

What is variance?

Empirical

- Measure replicates of the value of a variable X, *i.e.* x_i
- The variance is the average squared deviation from the mean; a measure of the spread or dispersion of the measurements:

$$var(X) = \frac{1}{n-1}\sum_{x=1}^{n}\left(x_i - \overline{x}\right)^2$$

- The standard deviation is the square root of the variance $sd(X) = \sqrt{var(X)}$

Theoretical

- The variance of the distribution of X, *e.g.* if $X \sim N(\mu, \sigma)$ the variance equals σ^2
- The variance is a measure of the width of the distribution

Variable variance

Definitions

- **homoscedasticity**: distributions of variables X and Y have the same variance
- **heteroscedasticity**: distributions of variables X and Y have different variance

Variable variance

Definitions

- **homoscedasticity**: distributions of variables X and Y have the same variance
- **heteroscedasticity**: distributions of variables X and Y have different variance

The problem of heteroscedasticity

- Many statistical techniques assume homoscedasticity
  - *E.g.* all techniques that use least squares fitting (ANOVA)
- Many machine learning algorithms will be affected by heteroscedasticity
  - *E.g.* clustering, dimension reduction, supervised learning
- Graphs can often be better interpreted with homoscedastic data

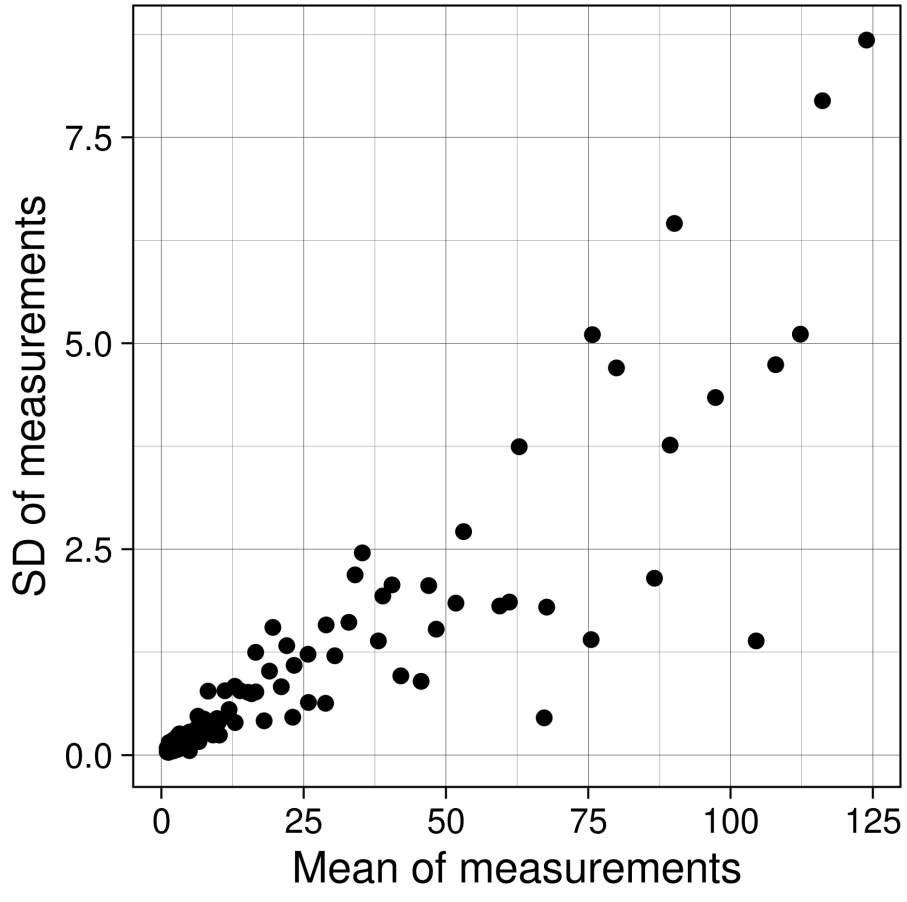
Investigate the variance before doing statistical tests!

Two options:

1. Use a distribution and test that are tailored for the data
    - We will demonstrate this for RNA-seq data
2. Transform the data so that the result is homoscedastic: a **variance stabilizing transformation**
    - Examples follow below

Many experimental variables have proportional error

- Proportional error is the most common instrumental error: $\sigma = a \cdot \mu$
- The standard deviation is a fixed percentage of the measured value
- The data are heteroscedastic

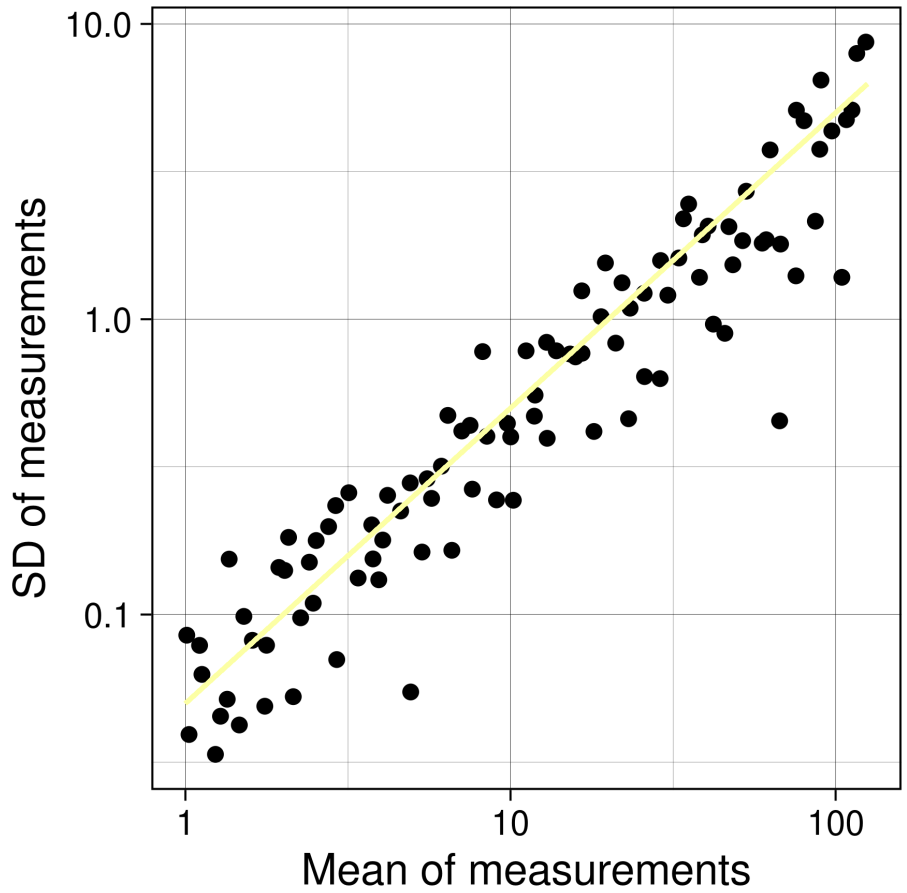


Many experimental variables have proportional error

- Proportional error is the most common instrumental error: $\sigma = a \cdot \mu$
- The standard deviation is a fixed percentage of the measured value
- The data are heteroscedastic

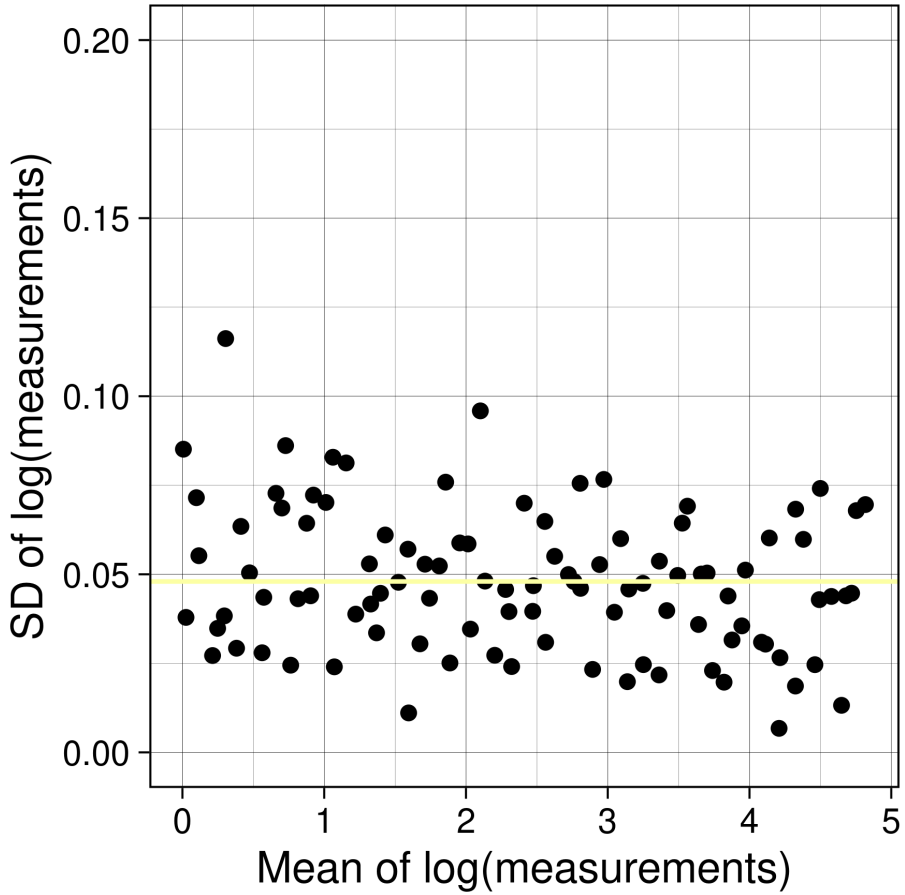
- On a logarithmic scale the slope equals 1:

$$\log(\sigma) = \log(a) + 1 \cdot \log(\mu)$$



A logarithmic transformation stabilizes the variance

- Logarithmic transform is the most common data transformation used: $y = \log(x)$
- The transformed data is homoscedastic

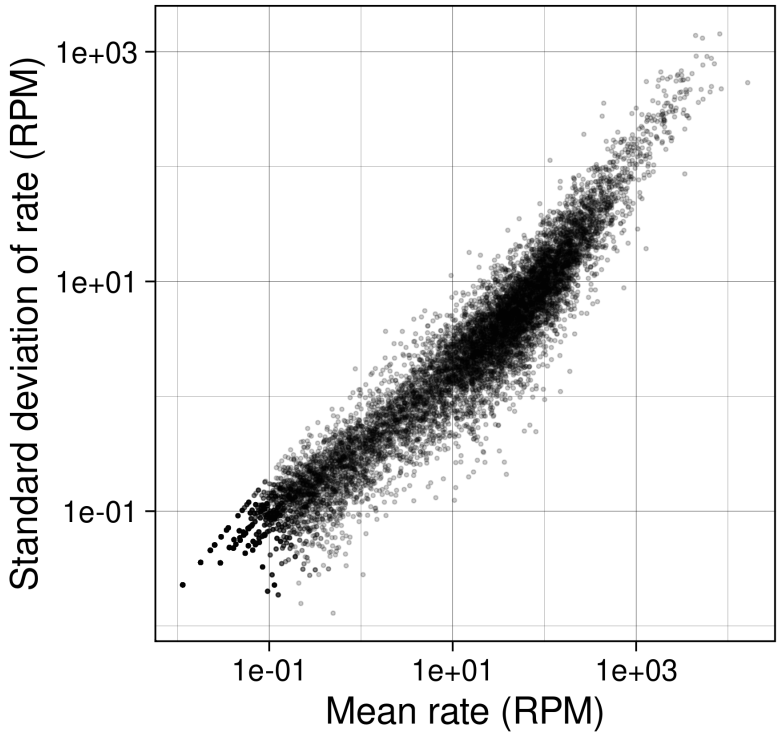


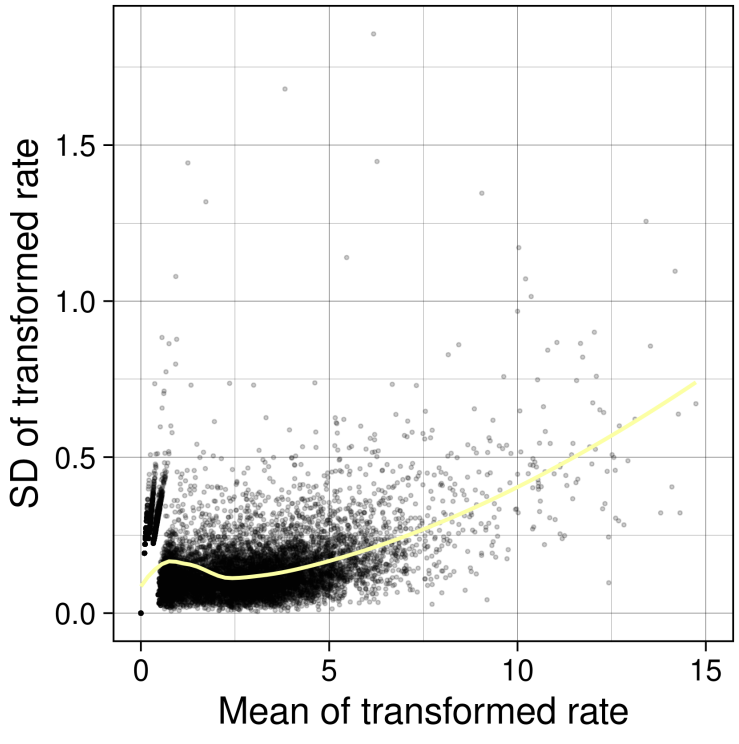
Stabilization of variance

- Transform the data in such a way that the variance becomes independent of the mean
  - Logarithmic transform when error is a fixed percentage of the measured value
  - Square root transform for Poisson-distributed count data
  - In general: if $\sigma \propto \mu^{\beta}$ then the transformation $y = x^{1-\beta}$ will stabilize the variance*
- If transformed data also have normal-distributed error, you can apply t-tests, ANOVA *etc.*

[*] For an even more general treatment with arbitrary but smooth relation between variance and mean, see Section 4.4.4 in Huber & Holmes

Applying this to the Pasilla RNA-seq data

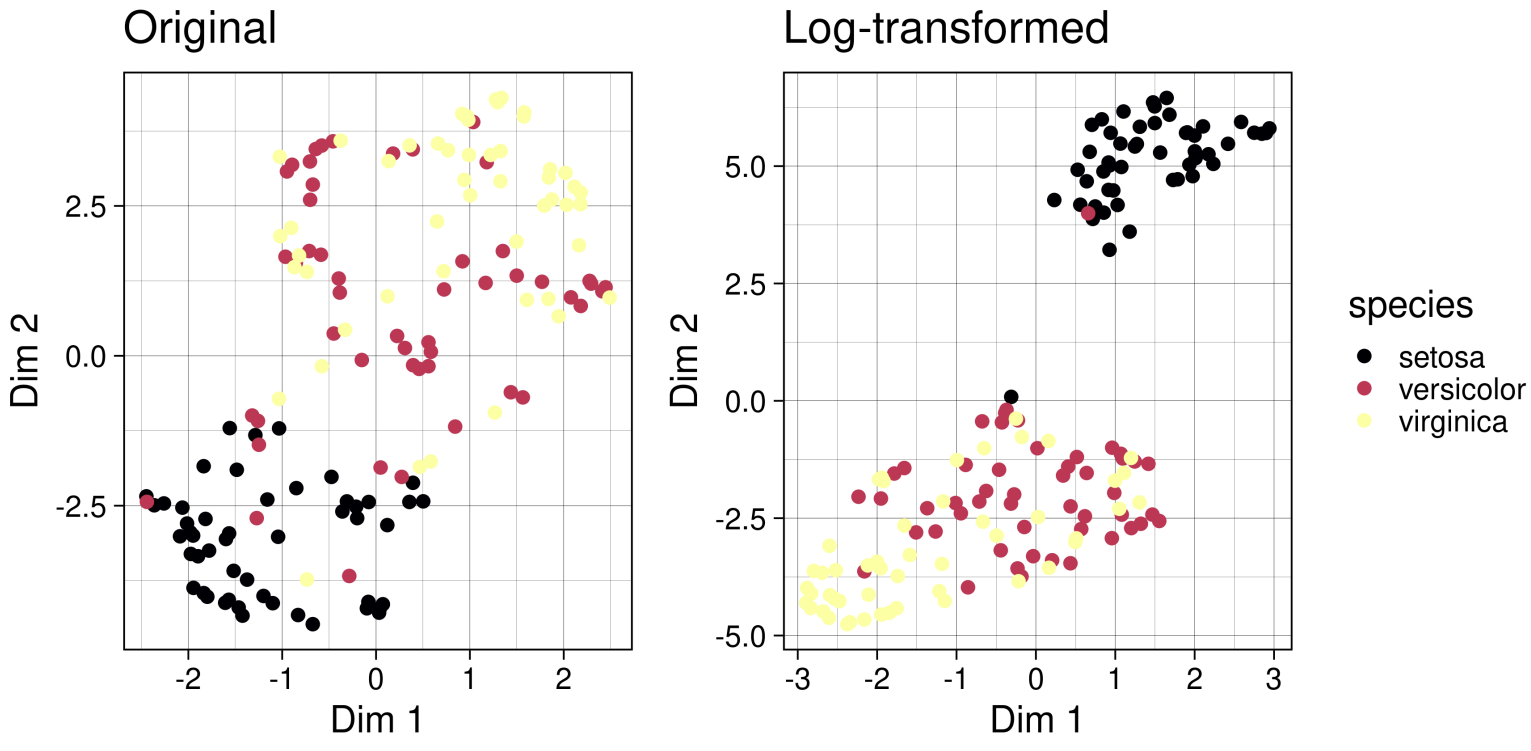




$\text{SD} \propto \text{mean}^{0.7} \longrightarrow$ transformation: $y = x^{0.3}$

Effect of variance stabilization on dimension reduction

UMAP* on original data and on log-transformed data



[*] Uniform Manifold Approximation and Projection

Modeling RNA-seq with the Negative Binomial distribution

Modeling over-dispersed count data

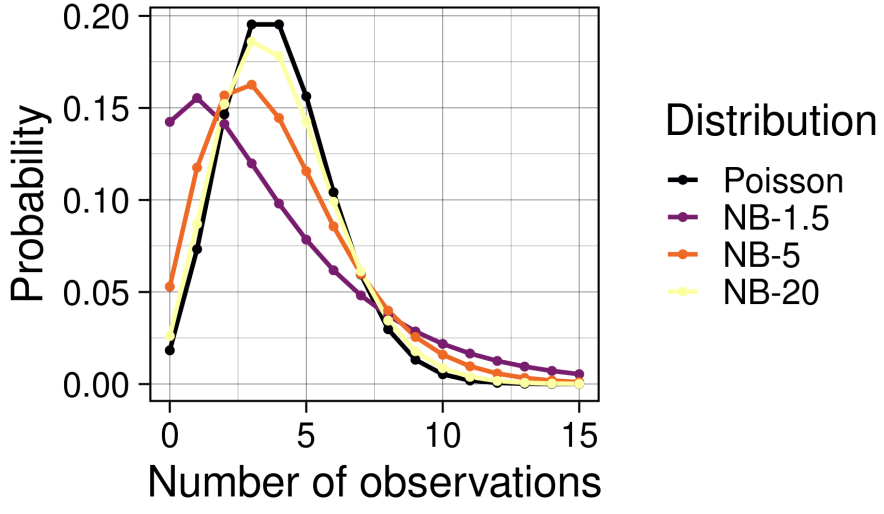
- The Poisson distribution with parameter did not fit the RNA-seq counts
  - It has a single parameter: $\lambda = \mu$
- An alternative distribution is the **Negative Binomial** or **Gamma-Poisson** distribution
  - It is a discrete distribution
  - It has two parameters: mean μ and dispersion parameter r

Modeling over-dispersed count data

- The Poisson distribution with parameter did not fit the RNA-seq counts
  - It has a single parameter: $\lambda = \mu$
- An alternative distribution is the **Negative Binomial** or **Gamma-Poisson** distribution
  - It is a discrete distribution
  - It has two parameters: mean μ and dispersion parameter r

Comparing Poisson- and NB distributions

- Using $\mu = 4$ for all examples
- NB distribution with $r = 1.5$, 5, and 20
- The larger r the smaller the standard deviation
- As $r \to \infty$, NB approaches Poisson



Fitting a model to RNA-seq data

For every gene i

- The log fold-difference of the concentrations $\log\left(\frac{q_i^B}{q_i^A}\right) = \log\left(q_i^B\right) - \log\left(q_i^A\right)$ is a constant β_i^{AB}

Fitting a model to RNA-seq data

For every gene i

- The log fold-difference of the concentrations $\log\left(\frac{q_i^B}{q_i^A}\right) = \log\left(q_i^B\right) - \log\left(q_i^A\right)$ is a constant β_i^{AB}
- **Normalization**: The mean counts μ_i^A and μ_i^B are proportional, with a size factor s^A or s^B, to the concentrations q_i^A and q_i^B:

$$\mu_i^A = s^A q_i^A \qquad \mu_i^B = s^B q_i^B$$

Fitting a model to RNA-seq data

For every gene i

- The log fold-difference of the concentrations $\log\left(\frac{q_i^B}{q_i^A}\right) = \log\left(q_i^B\right) - \log\left(q_i^A\right)$ is a constant β_i^{AB}
- **Normalization**: The mean counts μ_i^A and μ_i^B are proportional, with a size factor s^A or s^B, to the concentrations q_i^A and q_i^B:

$$\mu_i^A = s^A q_i^A \qquad \mu_i^B = s^B q_i^B$$

- The counts K_i^A, K_i^B of gene i are fitted to a Negative-Binomial distribution:

$$K_i^A \sim \mathrm{NB}(\mu_i^A, r_i^A) \qquad K_i^B \sim \mathrm{NB}(\mu_i^B, r_i^B)$$

Fitting a model to RNA-seq data

For every gene i

- The log fold-difference of the concentrations $\log\left(\frac{q_i^B}{q_i^A}\right) = \log\left(q_i^B\right) - \log\left(q_i^A\right)$ is a constant β_i^{AB}
- **Normalization**: The mean counts μ_i^A and μ_i^B are proportional, with a size factor s^A or s^B, to the concentrations q_i^A and q_i^B:

$$\mu_i^A = s^A q_i^A \qquad \mu_i^B = s^B q_i^B$$

- The counts K_i^A, K_i^B of gene i are fitted to a Negative-Binomial distribution:

$$K_i^A \sim \mathrm{NB}(\mu_i^A, r_i^A) \qquad K_i^B \sim \mathrm{NB}(\mu_i^B, r_i^B)$$

- Using the NB distribution, calculate the p-value under H_0 that $\mu_i^A = \mu_i^B$

Estimating the dispersion parameter

Additional constraint

The parameter r is modeled as a smooth function of μ:

$$r = v(\mu)$$

Variance-stabilizing transformations of RNA-seq data

Variance-stabilizing transformations for NB-distributed data exist, see Huber & Holmes

It should be used for applications in which the counts are used, for example to assess the quality of the data:

- Hierarchical clustering & heat maps
- Dimension reduction, like PCA

Multiple hypothesis testing

In a data set with multiple genes we test multiple null hypotheses

Comparing gene expression under conditions A and B

Gene	H_0
G_1	$\mu_1^A = \mu_1^B$
G_2	$\mu_2^A = \mu_2^B$
G_3	$\mu_3^A = \mu_3^B$
$\vdots$	$\vdots$
G_n	$\mu_n^A = \mu_n^B$

In a data set with multiple genes we test multiple null hypotheses

Comparing gene expression under conditions A and B

Gene	H_0	p-value	test
G_1	$\mu_1^A = \mu_1^B$	0.47	accept
G_2	$\mu_2^A = \mu_2^B$	0.0001	reject
G_3	$\mu_3^A = \mu_3^B$	0.04	reject
$\vdots$	$\vdots$	$\vdots$	$\vdots$
G_n	$\mu_n^A = \mu_n^B$	0.74	accept

In a data set with multiple genes we test multiple null hypotheses

Comparing gene expression under conditions A and B

Gene	H_0	p-value	test
G_1	$\mu_1^A = \mu_1^B$	0.47	accept
G_2	$\mu_2^A = \mu_2^B$	0.0001	reject
G_3	$\mu_3^A = \mu_3^B$	0.04	reject
$\vdots$	$\vdots$	$\vdots$	$\vdots$
G_n	$\mu_n^A = \mu_n^B$	0.74	accept

- The genes with rejected null hypotheses, $p < 0.05$, are the interesting ones; the *positive calls* or *discoveries*

The trouble with Type I error control

Suppose, we measure the expression of 5050 genes

The trouble with Type I error control

Suppose, we measure the expression of 5050 genes

- Suppose also that 50 genes are truly differentially expressed, and their H_0 is rejected

The trouble with Type I error control

Suppose, we measure the expression of 5050 genes

- Suppose also that 50 genes are truly differentially expressed, and their H_0 is rejected
- Then the remaining ≈5000 genes are not differentially expressed: their H_0 is true

The trouble with Type I error control

Suppose, we measure the expression of 5050 genes

- Suppose also that 50 genes are truly differentially expressed, and their H_0 is rejected
- Then the remaining ≈5000 genes are not differentially expressed: their H_0 is true
- Using an α level (cut-off p-value) of 0.05 for the test, 5% of these 5000 genes (250) will be false positive

The trouble with Type I error control

Suppose, we measure the expression of 5050 genes

- Suppose also that 50 genes are truly differentially expressed, and their H_0 is rejected
- Then the remaining ≈5000 genes are not differentially expressed: their H_0 is true
- Using an α level (cut-off p-value) of 0.05 for the test, 5% of these 5000 genes (250) will be false positive

Conclusion

The fraction of false positives in the list of all positive calls equals $\frac{250}{50+250} = 83\%$!

The trouble with Type I error control

Suppose, we measure the expression of 5050 genes

- Suppose also that 50 genes are truly differentially expressed, and their H_0 is rejected
- Then the remaining ≈5000 genes are not differentially expressed: their H_0 is true
- Using an α level (cut-off p-value) of 0.05 for the test, 5% of these 5000 genes (250) will be false positive

Conclusion

The fraction of false positives in the list of all positive calls equals $\frac{250}{50+250} = 83\%$!

What does the p-value tell about the probability that we have true positive?

What do we want to control?

H_0	Accept	Reject
True	True negative	False positive
False	False negative	True positive

Type I error

- With Type I error control or α-level we control

$$\alpha = \frac{\text{false positives}}{\text{true negatives} + \text{false positives}}$$

What do we want to control?

H_0	Accept	Reject
True	True negative	False positive
False	False negative	True positive

Type I error

- With Type I error control or α-level we control

$$\alpha = \frac{\text{false positives}}{\text{true negatives} + \text{false positives}}$$

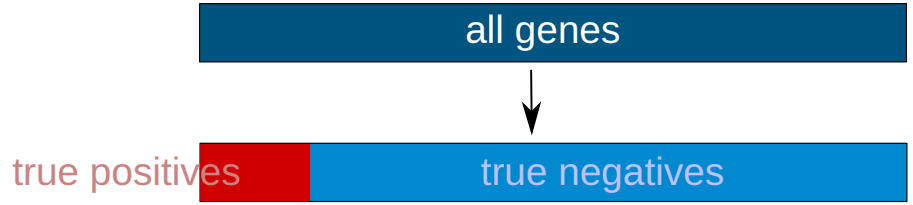
- We would like to control the fraction of false discoveries, or **false discovery rate**, *i.e.* the fraction of Type I errors in all *positives* (all *discoveries*):

$$\text{FDR} = \frac{\text{false positives}}{\text{true positives} + \text{false positives}}$$

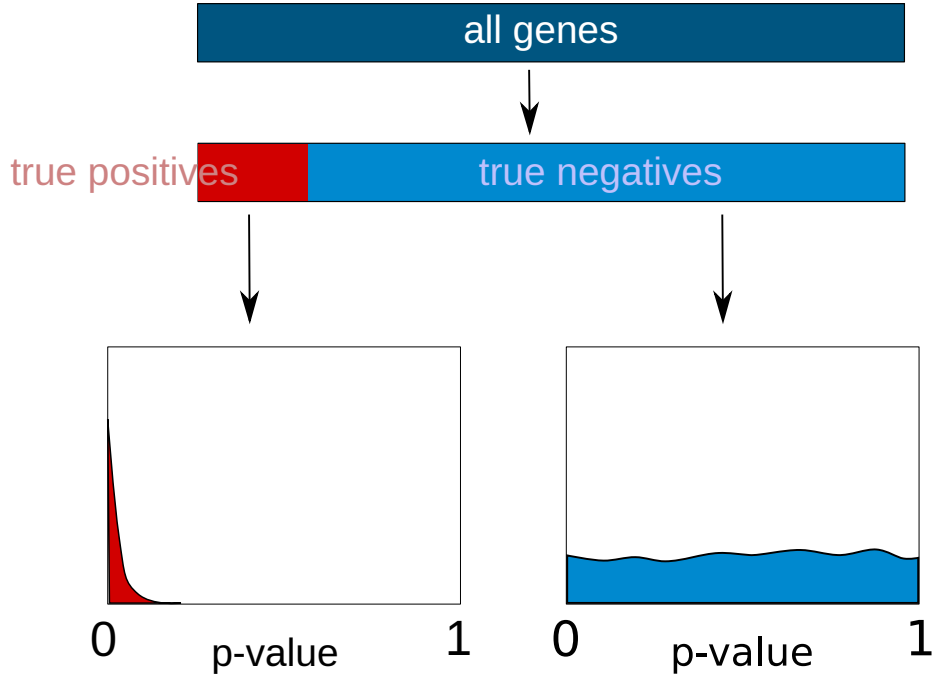
How can we estimate this fraction?

All genes that we test

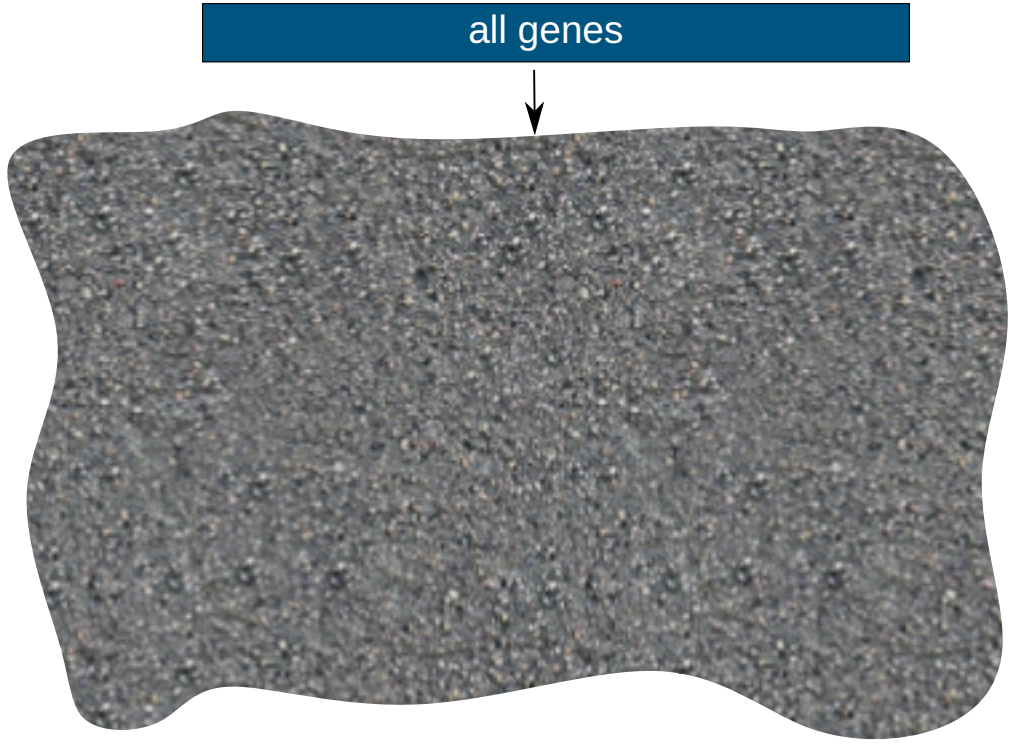
Some positives, mostly negatives



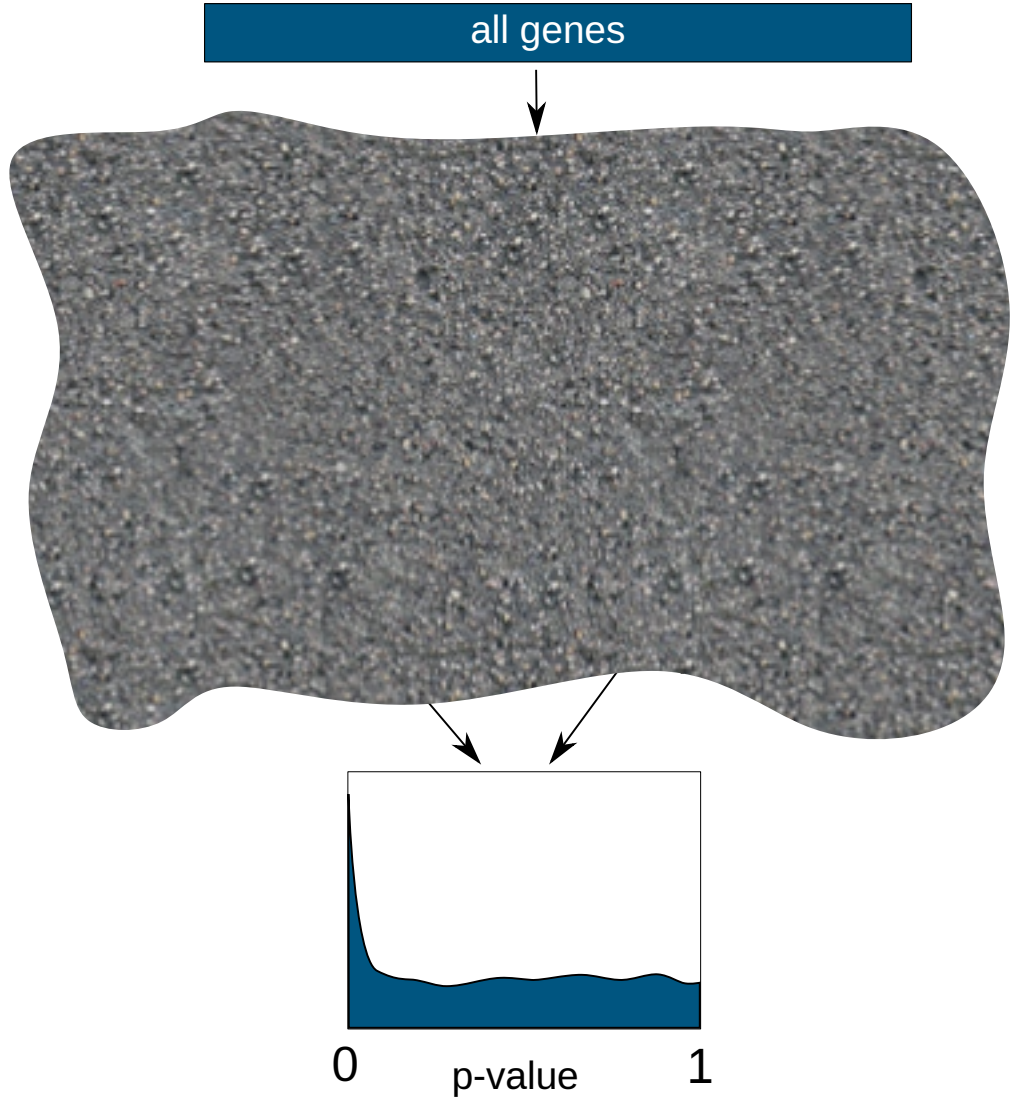
Distributions of their p-values



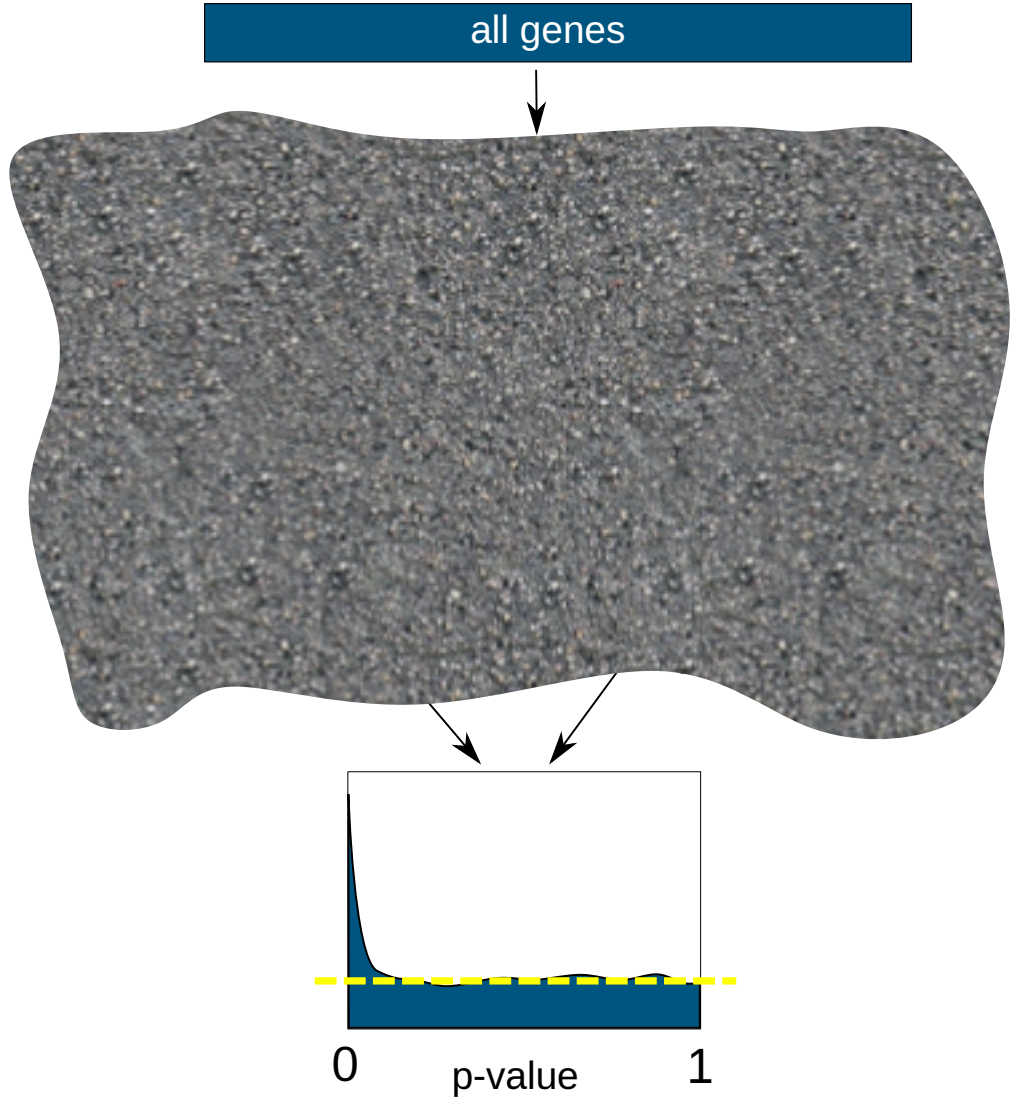
Unfortunately, the truth is hidden



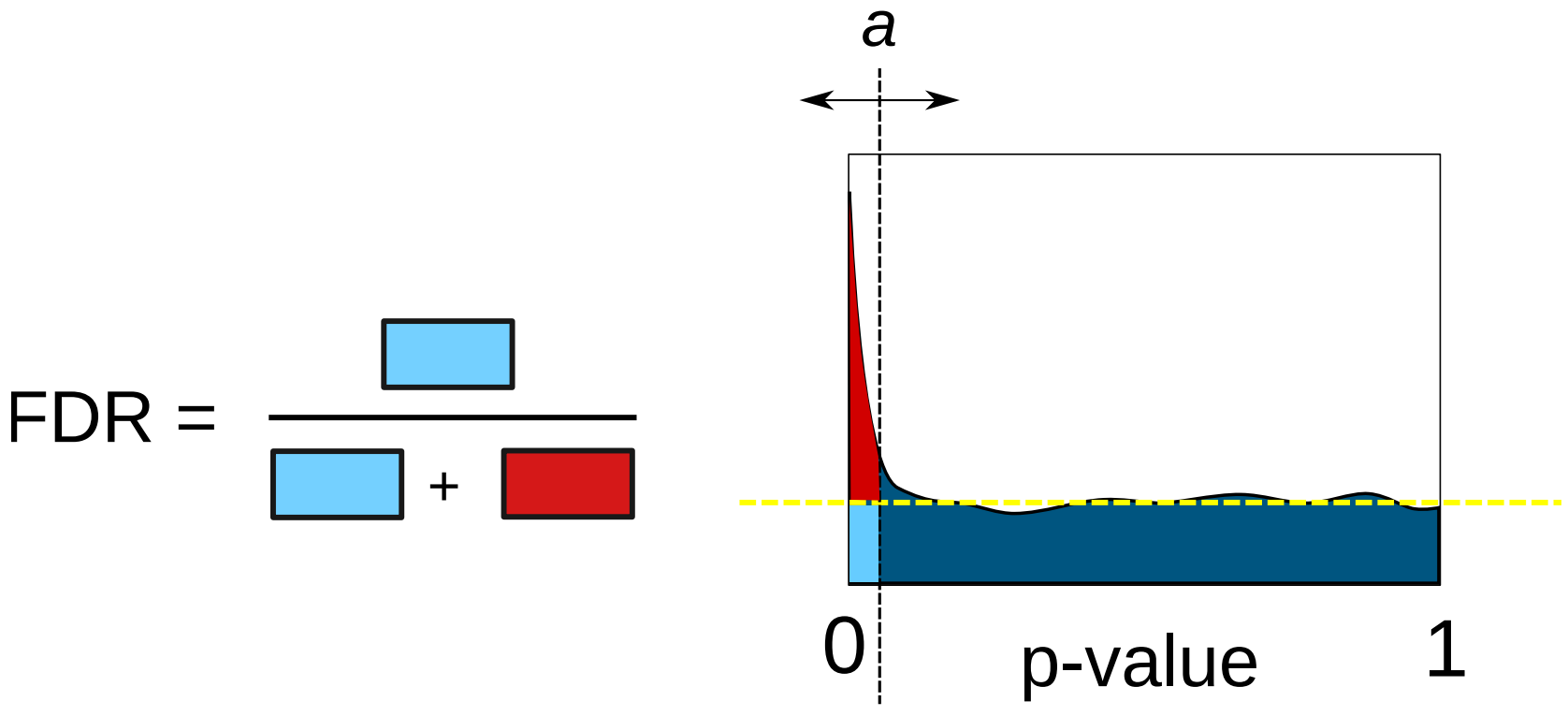
We obtain a combined distribution of p-values



We can reconstruct part of the truth



Controlling the false discovery rate



By moving the dashed line *a* we can change the FDR to our desire*

[*] Almost: the lower limit of the FDR is determined by the data

Difference between FDR and Type I error control

- With α-level control we have no knowledge of the fraction of false positives
- With FDR control we get (determine) an estimate of the fraction of false positives in a list of selected genes
  - This is **essential information for follow-up experiments**
- The p-value is a property of a single hypothesis test, the FDR is a property of a list of positive calls

Using the FDR in follow-up experiments

Based on a list of significant (positive) genes you might:

- Make knock-outs of each gene.

  Very time-consuming and expensive; use a very low FDR
- Do a sequence motif analysis.

  False positives may blur motifs; use a low FDR
- Do a gene-class enrichment analysis.

  False positives as well as false negatives may decrease statistical power; use an intermediate FDR
- Cluster genes.

  False positives may influence the clustering result; Test a range of FDR

Other method: controlling the Family-Wise Error Rate

- The *family* is a collection of hypotheses for which it is reasonable to take into account a combined measure of error
- The **FWER** is the probability of obtaining **at least one Type I error in the family**
- A well-known, simple procedure to control FWER is the **Bonferroni correction**:
  - We have a family of N hypotheses (*e.g.* N genes)
  - We have a FWER acceptance level of α (probability of having at least one false positive)
  - Only reject H_{0i} if the p-value $p_i \leq \frac{\alpha}{N}$ (*e.g.* if $N = 1000$ and $\alpha = 0.05$ then $p_i \leq 0.00005$)
  - Usually much more stringent than FDR (more false negatives)

Literature and remarks

Further reading and background

Holmes & Huber: "Modern Statistics for Modern Biology": online book, free access.

- Chapter 4, Gamma-Poisson or Negative Binomial distribution, Variance stabilization, zero-inflated data (4.4)
- Chapter 8, High-throughput count data (8.1-8.4, 8.8, 8.10)
- Chapter 6, Multiple Testing (6.7-6.10)

Please, browse through the book: other chapters may be relevant for this course and for future reference

Remark about normalization

Note that the term *normalization* has different meanings in statistics. You will likely encounter both!

- In this slide it is used to indicate the correction of bias
- It is also used to indicate the operation of scaling random variables

Remarks about the NB distribution

The Negative Binomial distribution $\mathrm{NB}(X = k)$ with parameters r and p can be obtained in different ways. One is the following:

- k is the number of times that you get a tail if you repeat a Bernoulli trial (coin flip) until you have observed r heads. In a single Bernoulli trial the probability to throw head equals p, and to throw tail $1 - p$
- The mean, μ of tails in such an experiment equals $\frac{rp}{1-p}$

Another way to generate the NB distribution is as a infinite mixture of Poisson distributions, whose parameters λ are themselves drawn from a gamma-distribution, see Ch4 in Huber & Holmes. It is the reason why the NB distribution is also known as gamma-Poisson distribution.

Remarks about FDR

- You may encounter the concept *local FDR* (often indicated by lower case **fdr**), *e.g.* in Holmes and Huber.
  In this slide about the FDR the *local fdr* is the ratio of blue line segment to total (blue + red) line segment on the line a, whereas the *total FDR* is the ratio of blue area to total area left of the line a. (The areas are the integral of the line segments). It estimates the likelihood of a positive call being a false positive.

- You may encounter the concept *q-value*.
  In a list of hypothesis tests ordered by increasing p-values, the q-value of a test is the FDR (expected fraction of false positives) of the collection of hypothesis tests up to and including that test.

- You will often encounter the terms *"corrected p-value"* or *"adjusted p-value"* in the literature when the false discovery rate is meant.
  Do not use these terms: they are confusing and imprecise because the meaning of an FDR is very different from that of a p-value.