# CS-E5875 High-Throughput Bioinformatics
# Introduction, hypothesis testing, multiple testing

Harri Lähdesmäki

Department of Computer Science
Aalto University

October 24, 2023

# Contents

- Introduction
- Statistical hypothesis testing
- Types of error
- Multiple testing

# This course

- Focuses on **methods to analyze** high-throughput biological data
- Primary data type: sequencing data
- Aims to give an understanding of how, why and when these methods work
- Less focus on applications or implementations of methods

# What is high-throughput biological data?

- High-throughput technologies can be thought of as massively parallel automated methods to carry out a large number of individual experiments/biochemical tests simultaneously
- Examples: a microarray or a sequencing experiment can simultaneously
  - Measure expression (=abundance) of tens of thousands of genes in a biological sample
  - Quantify genetic variants at millions of positions throughout a genome

  → Data are produced at a massive scale

# What is high-throughput biological data?

- High-throughput technologies can be thought of as massively parallel automated methods to carry out a large number of individual experiments/biochemical tests simultaneously
- Examples: a microarray or a sequencing experiment can simultaneously
  - Measure expression (=abundance) of tens of thousands of genes in a biological sample
  - Quantify genetic variants at millions of positions throughout a genome

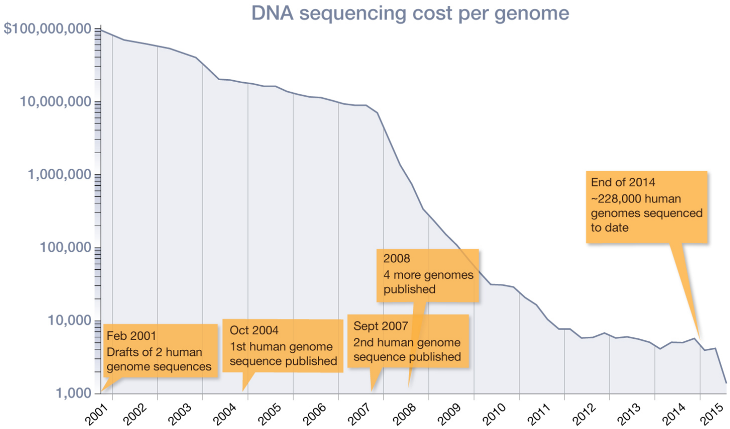
  → Data are produced at a massive scale
- Suitable computational methods are needed to analyze and exploit these data
  - Bioinformatic methods include: algorithmic, computational, mathematical, data mining, statistical, machine learning, and deep learning techniques
  - This course focuses mostly on statistical and machine/deep learning methods (or questions that are naturally answered by these methods)

# What is high-throughput biological data?

- High-throughput technologies can be thought of as massively parallel automated methods to carry out a large number of individual experiments/biochemical tests simultaneously
- Examples: a microarray or a sequencing experiment can simultaneously
  - Measure expression (=abundance) of tens of thousands of genes in a biological sample
  - Quantify genetic variants at millions of positions throughout a genome

  → Data are produced at a massive scale
- Suitable computational methods are needed to analyze and exploit these data
  - Bioinformatic methods include: algorithmic, computational, mathematical, data mining, statistical, machine learning, and deep learning techniques
  - This course focuses mostly on statistical and machine/deep learning methods (or questions that are naturally answered by these methods)
- Bioinformatics provides essential tools for molecular biology, genetics, biomedicine, healthcare, drug development, evolutionary studies, synthetic biology and more

# Data growth and sequencing costs



http://learn.genetics.utah.edu/content/precision/time/

# Beyond genome identification

After having sequenced the genome (e.g. human reference genome):

- Characterize genetic variation between individuals
- Identify the location of genes
- Analyze gene activity, functions, interactions, and regulation
- Quantify and analyze epigenomics
- Characterize dynamic properties of genome and functional genomics
- . . .
- Translate this data / knowledge for health and disease

# Contents

- Introduction
- Statistical hypothesis testing
- Types of error
- Multiple testing

# Statistical hypothesis testing

- Hypothesis testing is the main inferential statistics concept that we will use throughout this course
- We will briefly review the basics of hypothesis testing
    - We follow parts of J. Orloff's and J. Bloom's lecture notes "Null Hypothesis Significance Testing" (Orloff and Bloom, 2014)
    - You may also refer to several / any statistics book

# Statistical hypothesis testing

- Hypothesis testing is the main inferential statistics concept that we will use throughout this course
- We will briefly review the basics of hypothesis testing
    - We follow parts of J. Orloff's and J. Bloom's lecture notes "Null Hypothesis Significance Testing" (Orloff and Bloom, 2014)
    - You may also refer to several / any statistics book
- Conceptually speaking, the hypothesis testing framework asks if the observed data is outside the region where we expect the data to be
- If it is, then we have evidence to reject our initial conservative hypothesis

# Null hypothesis significance testing (NHST)

Key concepts:

- $H_0$: the null hypothesis. This specifies our conservative default assumptions for the model that generates the data
- $H_A$: the alternative hypothesis (also denoted as $H_1$). We are interested in testing the null hypothesis; if null is rejected we accept the alternative hypothesis as the best explanation for the data
- $T$: the test statistic of our choice, computed from the observed data
- Null distribution: the probability density of the test statistic, assuming the null hypothesis holds true

Typically the null hypothesis is chosen to be a simple and conservative hypothesis, which we reject if we have sufficient amount of evidence to reject $H_0$

# Example: coin flipping

We flip a coin $N$ times to test whether the coin is fair or unfair

The rationale is to check whether our coin results in unexpectedly few or many heads/tails

Let $\theta$ denote the probability that the coin flipping results in a head (or tail), then:

- Null hypothesis: $H_0 =$ "the coin is fair", i.e. $\theta = 0.5$
- Alternative hypothesis: $H_A =$ "coin is not fair", i.e. $\theta \neq 0.5$
- Test statistic: $T =$ number of heads in $N$ flips
- Null distribution: assuming the null hypothesis holds, the number of heads follows binomial distribution

  $$T \sim \mathrm{Binomial}(N, 0.5)$$

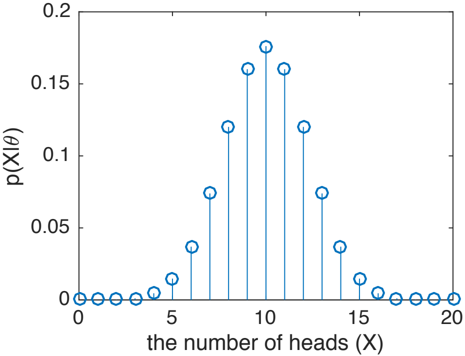
  with the probability density function

  $$P(T = k) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}$$

  for $k = 0, 1, \ldots, N$ and $\theta = 0.5$

# Example: coin flipping

- $N = 20$ coin flipping experiments
- The probabilities of obtaining any number of heads between 0 and 20 with a fair coin are shown on right (here $X$ is used to denote the test statistic, instead of $T$)
- So, is it “too unlikely” to observe e.g. as many as 15 heads? What about observing as few as 5 heads?



# $p$-value

- For a given realization $T = t$, the $p$-value is the probability of seeing test statistic value that is at least as extreme as the observed value $t$

$$p = P(\text{``test statistic at least as extreme as } t\text{''}),$$

  where the probability is computed using the null distribution, i.e., by assuming the null hypothesis is true

# $p$-value

- For a given realization $T = t$, the $p$-value is the probability of seeing test statistic value that is at least as extreme as the observed value $t$

  $$p = P(\text{``test statistic at least as extreme as } t\text{''}),$$

  where the probability is computed using the null distribution, i.e., by assuming the null hypothesis is true
- "At least as extreme as" depends on the application (i.e., hypothesis test, test statistic, experimental design)
- Standard hypothesis tests are either one-sided or two-sided:
  - One-sided: the test statistic can have significantly low values or high values (but not both)
    - One-sided test has directionality
  - Two-sided: the test statistic can have both significantly low values and high values
    - E.g. the coin flipping test is two-sided

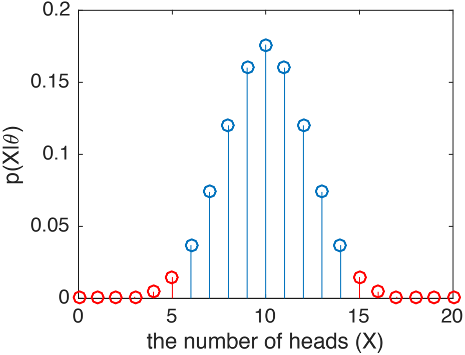
# Example: coin flipping cont'd

- The probability of observing $T$ smaller than 6 or larger than 14 is

$$P(T \leq 5 \text{ or } T \geq 15) \approx 0.0414$$

- $p$-value of smaller than 0.05 is a commonly used threshold
- By choosing a $p$-value (here 0.05) we get the rejection region formed by the extreme values (red)
- If the test statistic falls in the rejection region, then we consider to have enough evidence to reject the null hypothesis and accept the alternative hypothesis
- The typical values (blue) form the "acceptance" region

- In the "acceptance" region we do not have enough evidence to reject $H_0$
- In the "acceptance" region we do not make any decision based on data



# Types of null hypothesis

- Simple hypothesis: a null hypothesis that specifies the null distribution exactly
  - E.g. data is sampled from a given normal distribution with known mean and variance
- Composite hypothesis: a null hypothesis that does not specify the null distribution completely
  - E.g. data is sampled from a given normal distribution with known mean but unknown variance

# Types of null hypothesis

- Simple hypothesis: a null hypothesis that specifies the null distribution exactly
  - E.g. data is sampled from a given normal distribution with known mean and variance
- Composite hypothesis: a null hypothesis that does not specify the null distribution completely
  - E.g. data is sampled from a given normal distribution with known mean but unknown variance

- Exact hypothesis: a null hypothesis that specifies an exact parameter value, e.g., $\mathrm{mean} = 0$
- Inexact hypothesis: a null hypothesis that specifies a range of parameter values, e.g., $\mathrm{mean} \leq 0$

- Our coin flipping example has a null hypothesis that is simple and exact

# $t$-test

- In many applications data is assumed to be normally distributed
- Two-sample $t$-test can be applied to test the means of two samples which are assumed to be drawn from two normal distributions (we assume the same variance here)

$$\begin{aligned} x_1, \ldots, x_n &\sim N(\mu_1, \sigma^2) \\ y_1, \ldots, y_m &\sim N(\mu_2, \sigma^2) \end{aligned}$$

- Unknowns: $\mu_1$, $\mu_2$, and $\sigma^2$
  - This is a composite null hypothesis
- The null hypothesis $H_0$: $\mu_1 = \mu_2$
- The alternative hypothesis $H_A$: $\mu_1 \neq \mu_2$

# $t$-test

- Notation: $T$ is a random variable, $t$ is a particular realization of $T$
- The test statistic $T$ for the $t$-test:

  $$t = \frac{\overline{x} - \overline{y}}{s},$$

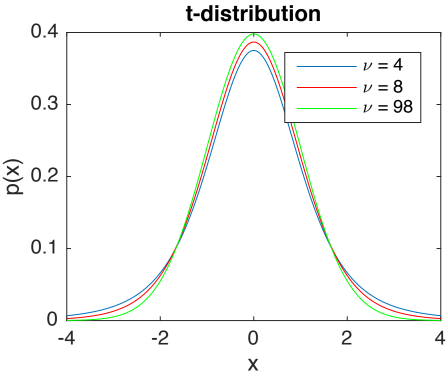
  where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ are the sample means, and $s^2$ is the pooled variance

  $$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right) \quad \text{and} \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- The null distribution: $p(T|H_0)$ can be shown to be the $t$-distribution with $n+m-2$ degrees of freedom

# *t*-test

- *t*-distribution for different degrees of freedom



# $t$-test

- One-sided $p$-value (right side): $p = P(T \geq t \mid H_0)$
- One-sided $p$-value (left side): $p = P(T \leq t \mid H_0)$
- Two-sided $p$-value: $p = P(|T| \geq |t| \mid H_0)$

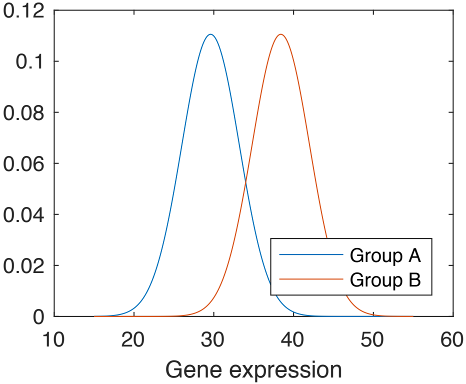
# $t$-test: example

- An example: let us assume that we are interested in quantifying whether a gene of interest is differentially expressed between two groups $A$ and $B$ (say, between healthy and diseased individuals)
- Measured gene expression values are

$$\begin{aligned} \text{Group } A: &\qquad 32, 25, 36, 27, 28 \\ \text{Group } B: &\qquad 29, 48, 39, 37, 39 \end{aligned}$$

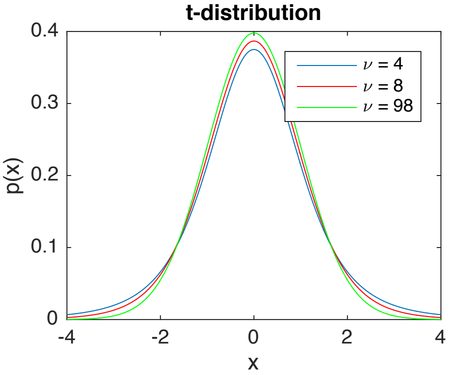
## $t$-test: example

- We can explore the data by plotting estimated normal densities for both groups: $\mathcal{N}(\overline{\mu}_A, s^2)$ and $\mathcal{N}(\overline{\mu}_B, s^2)$



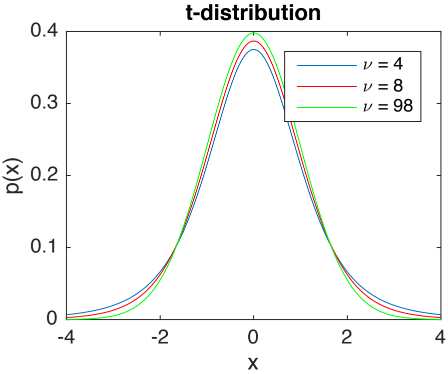
# *t*-test: example

- For quantitative inference, we can use the *t*-test
- The value of the t-statistic for our data is $-2.4388$



# *t*-test: example

- For quantitative inference, we can use the *t*-test
- The value of the t-statistic for our data is $-2.4388$



- In general, we may not know whether our gene can be up- or down-regulated and we need to apply two-sided test, which results in a *p*-value of 0.0406
- If we know that the gene expression value in group B can only be higher we can apply one-sided test (left side), which results in a *p*-value of 0.0203

# Contents

- Introduction
- Statistical hypothesis testing
- Types of error
- Multiple testing

# Types of error

Two types of errors can be made in a hypothesis testing

Type I error:

- Null hypothesis $H_0$ is true but we reject that in favour of $H_1$
- This incorrect decision results in a false positive

Type II error:

- Null hypothesis $H_0$ is false but we do not reject $H_0$
- This incorrect decision results in a false negative

# Types of error

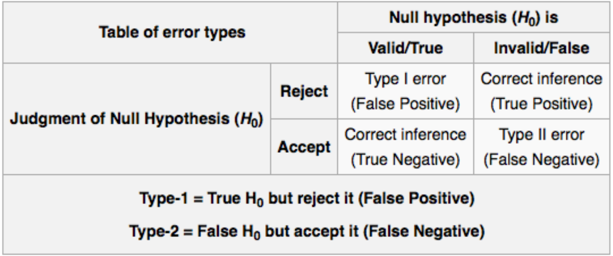
Two types of errors can be made in a hypothesis testing

Type I error:

- Null hypothesis $H_0$ is true but we reject that in favour of $H_1$
- This incorrect decision results in a false positive

Type II error:

- Null hypothesis $H_0$ is false but we do not reject $H_0$
- This incorrect decision results in a false negative



| Table of error types | | Null hypothesis ($H_0$) is | |
|---|---|---|---|
| | | **Valid/True** | **Invalid/False** |
| **Judgment of Null Hypothesis ($H_0$)** | **Reject** | Type I error<br>(False Positive) | Correct inference<br>(True Positive) |
| | **Accept** | Correct inference<br>(True Negative) | Type II error<br>(False Negative) |
| **Type-1 = True $H_0$ but reject it (False Positive)**<br>**Type-2 = False $H_0$ but accept it (False Negative)** | | | |

Figure from (Wikipedia)

# Significance of a test

- Significance level of a test (often called $\alpha$) is defined to be the probability that we incorrectly reject $H_0$

$$\text{Significance level} = P(\text{reject } H_0 | H_0) = P(\text{type I error})$$

- Significance level of $\alpha = 0.05$ is commonly used in practise
- In other words, if the computed $p$-value is smaller than $\alpha$, then we reject the null hypothesis
- When we reject the null hypothesis, we say the result is statistically significant at level $\alpha$
- Note: rejecting the null hypothesis with level $\alpha$ does not mean that the alternative hypothesis is correct with probability of 0.95

# Power of a test

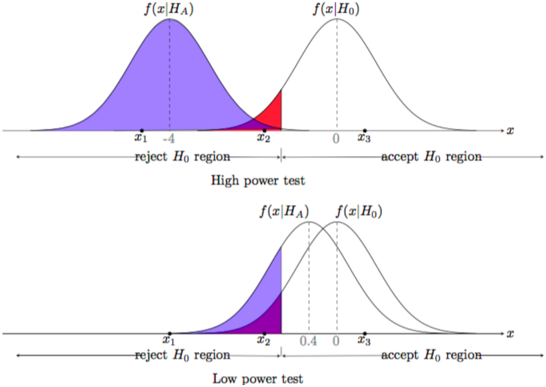
- Power of a test is defined to be the probability that we correctly reject $H_0$

$$\begin{aligned}\text{Power} &= P(\text{reject } H_0|H_A) \\ &= 1 - P(\text{do not reject } H_0|H_A) \\ &= 1 - P(\text{type II error})\end{aligned}$$

# Illustration of the significance and power of a test

Figure from (Orloff and Bloom, 2014) illustrates the concepts of significance and power

- Red shaded area below $f(x|H_0)$ represents the significance
- Violet shaded area below $f(x|H_A)$ represents the power: the probability that the test statistic is in the rejection region of $H_0$ when $H_A$ is true
- Note that the null hypothesis significance testing works without caring about $f(x|H_A)$



# NHST steps

- Choose a null hypothesis $H_0$
- Choose a test statistic
- Decide if your alternative hypothesis is one-sided or two-sided
- Choose a significance level
- Perform the hypothesis test

# Contents

- Introduction
- Statistical hypothesis testing
- Types of error
- Multiple testing

# Multiple testing

- Multiple testing problem occurs when a statistical analysis and decision making involves multiple simultaneous statistical hypothesis tests
- The *p*-values (i.e., confidence levels) described above are valid for a single test
- Consider the previous example of comparing gene expression (for gene $x_1$) between Groups A and B
  - If 5% confidence level is used for a single test, then there is only 0.05 probability that null hypothesis is rejected incorrectly
  - If the test is applied to 100 genes $(x_i, i \in \{1, \ldots 100\})$ for which the null hypothesis holds (i.e., they are not differentially expressed) independently, then the expected number of genes for which the null hypothesis is rejected incorrectly is 5

# Multiple testing

- Multiple testing problem occurs when a statistical analysis and decision making involves multiple simultaneous statistical hypothesis tests
- The *p*-values (i.e., confidence levels) described above are valid for a single test
- Consider the previous example of comparing gene expression (for gene $x_1$) between Groups A and B
  - If 5% confidence level is used for a single test, then there is only 0.05 probability that null hypothesis is rejected incorrectly
  - If the test is applied to 100 genes ($x_i, i \in \{1, \ldots 100\}$) for which the null hypothesis holds (i.e., they are not differentially expressed) independently, then the expected number of genes for which the null hypothesis is rejected incorrectly is 5

→ Hypothesis testing will lead to many false positives if the *p*-values are not corrected for multiple testing

- Multiple testing is a real challenge in most bioinformatics applications
  - Differential gene expression analysis
  - Detecting disease associated genomic variant
  - Detection of protein binding sites along whole genome from ChIP-seq
  - ...

# Multiple testing problem[1]

- Lets assume we have $m$ independent hypothesis $H_0^{(1)}, \ldots, H_0^{(m)}$ and lets assume we know already beforehand that the null hypothesis holds for every one of them (that's a boring assumption to start with, but lets continue with that assumption anyways)
- If we make $m$ independent tests with significance level $\alpha$, then each of the $m$ tests will be significant with probability $\alpha$
- Now the total number of false positives $X$ will have a distribution

  $$X \sim \text{Binomial}(m, \alpha)$$

  (recall the coin flipping, now with a biased coin)
- The expectation of a binomial distribution is $E(X) = m\alpha$
- Once again, if we want to carry out a test e.g. for all approx. 20000 human genes, then the expected number of false positives (assuming we know that null hypothesis holds for all) is $20000 \cdot 0.05 = 1000$

[1] From here onwards, parts of the slides follow Sections 7.2.2–7.2.4 from (Wilkinson, 2017). You can also check Section 18.7 from (Hastie et al., 2017)

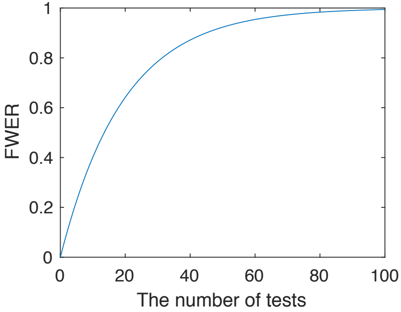
# Family-wise error rate

- Recall the type I error
  - Null hypothesis $H_0$ is true but it is rejected in favour of $H_1$
- Assuming again $m$ independent tests for which we know that the null hypothesis is true, then the probability that any of the hypothesis will be rejected with significance level $\alpha$ is

$$\overline{\alpha} = 1 - (1 - \alpha)^m$$

  i.e., the probability of making one or more type I errors
- This is also called the family-wise error rate $(\mathrm{FWER})$

- FWER for $m \in \{0, \ldots, 100\}$ tests with $\alpha = 0.05$



- Note: for $m = 1$, $\mathrm{FWR} = \alpha$
- FWER is independent of the type of a test or tests

# Bonferroni correction

- Let $H_0^{(1)}, \ldots, H_0^{(m)}$ be a collection of hypotheses and $p_1, \ldots, p_m$ the corresponding $p$-values
- Let $I_0 \subseteq \{1, \ldots, m\}$ be the (unknown) subset of the true null hypotheses, $m_0 = |I_0| \leq m$
- Bonferroni correction is defined as follows:
    - Given the original significance level $\alpha$ and the number of statistical tests $m$, then Bonferroni correction will reject only those null hypothesis $i$ for which $p_i \leq \alpha/m$
    - Equivalently, the multiple testing corrected $p$-value for the $i^{\text{th}}$ test is $\min\{mp_i, 1\}$

# Bonferroni correction

- Let $H_0^{(1)}, \ldots, H_0^{(m)}$ be a collection of hypotheses and $p_1, \ldots, p_m$ the corresponding $p$-values
- Let $I_0 \subseteq \{1, \ldots, m\}$ be the (unknown) subset of the true null hypotheses, $m_0 = |I_0| \leq m$
- Bonferroni correction is defined as follows:
  - Given the original significance level $\alpha$ and the number of statistical tests $m$, then Bonferroni correction will reject only those null hypothesis $i$ for which $p_i \leq \alpha/m$
  - Equivalently, the multiple testing corrected $p$-value for the $i^{\text{th}}$ test is $\min\{mp_i, 1\}$
- For the Bonferroni correction method, $\mathrm{FWER} \leq \alpha$ because

$$\mathrm{FWER} = P\left(\bigcup_{i \in I_0} \left\{p_i \leq \frac{\alpha}{m}\right\}\right) \leq \sum_{i \in I_0} P\left(\left\{p_i \leq \frac{\alpha}{m}\right\}\right) = \sum_{i \in I_0} \frac{\alpha}{m} = m_0 \frac{\alpha}{m} \leq= \alpha$$

  (Note: each $\left\{p_i \leq \frac{\alpha}{m}\right\}$ is considered as an event, and the inequality follows from the union bound)
- The Bonferroni correction is conservative

# False discovery rate

- False discovery rate (FDR) is the proportion of false positives among all positives

$$\mathrm{FDR} = \frac{\#\text{false positives}}{\#\text{false positives} + \#\text{true positives}} \in [0, 1]$$

- Formally FDR is defined as the expectation of the above quantity
- FDR of 0.05 means that 5% of the rejected null hypothesis are false
- However, on the other hand, FDR of 0.05 means that 95% of the rejected hypothesis are true findings (i.e., tests for which $H_A$ holds)
- A small fraction of false positives are often accepted as long as majority of the results are true
- In bioinformatics applications, FDR is typically more useful than FWER

# False discovery rate

- Lets again assume that we have $m$ tests with $p$-values $p_1, \ldots, p_m$
- We can order the $p$-values in increasing order $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
- The choice of significance level $\alpha$ is equivalent to deciding how many of the smallest $p$-values are considered significant
  - Lets denote that number (a positive integer) by $\ell$
- Because a significance level $\alpha$ corresponds to a particular cutoff $\ell$, we can denote that by explicitly writing $\ell(\alpha)$ (although generally we do not that mapping)
- Thus, $\alpha$ gives a list of significant $p$-values, $p_{(1)}, p_{(2)}, \ldots, p_{(\ell(\alpha))}$
  - A small $\alpha$ results in a short list (small $\ell$)
  - A larger $\alpha$ results in a longer list (larger $\ell$)
  - $\ell(\alpha)$ is monotonically increasing in $\alpha$
  - As noted above, we do not know this mapping

# False discovery rate

- Lets assume that the number of true positives (for which the null hypothesis does not hold) is small compared to the total number of tests $m$
- Thus, similarly as above, the number of false positives is still approximatively binomially distributed as $X \sim \text{Binomial}(m, \alpha)$
- Thus, the FDR is (assuming $\ell(\alpha) \geq X$)

$$\text{FDR} \approx \frac{X}{\ell(\alpha)} \quad \text{and} \quad E(\text{FDR}) \approx \frac{E(X)}{\ell(\alpha)} = \frac{m\alpha}{\ell(\alpha)}$$

# False discovery rate

- Lets assume that the number of true positives (for which the null hypothesis does not hold) is small compared to the total number of tests $m$
- Thus, similarly as above, the number of false positives is still approximatively binomially distributed as $X \sim \text{Binomial}(m, \alpha)$
- Thus, the FDR is (assuming $\ell(\alpha) \geq X$)

$$\text{FDR} \approx \frac{X}{\ell(\alpha)} \quad \text{and} \quad E(\text{FDR}) \approx \frac{E(X)}{\ell(\alpha)} = \frac{m\alpha}{\ell(\alpha)}$$

- Generally we want to limit the fraction of false positive findings (i.e., FDR) by a value $q$, thus

$$\frac{m\alpha}{\ell(\alpha)} \leq q \quad \Leftrightarrow \quad \alpha \leq \frac{q\ell(\alpha)}{m}$$

- One needs to choose a small enough $\alpha$ so that the above inequality holds
  - This is little tricky because $\ell(\alpha)$ depends on $\alpha$ too

# False discovery rate

- To solve the inequality on the previous page, *hypothetically* assume we have inverted the function $\ell(\cdot) : [0,1] \to \{1,\ldots,m\}$ as $\alpha(\cdot) : \{1,\ldots,m\} \to [0,1]$
- We can write

$$\alpha(\ell) \leq \frac{q\ell}{m}$$

- Notice that the significance level (or the $p$-value threshold) that gives a list of length $\ell$ is $p_{(\ell)}$, thus we have

$$p_{(\ell)} \leq \frac{q\ell}{m}$$

- Thus, to guarantee FDR $\leq q$, we just need to run through all possible values of $\ell$, from 0 to $m$, in order to find the largest value of $\ell$ that satisfies $p_{(\ell)} \leq \frac{q\ell}{m}$ and to find the corresponding $p_{(\ell)}$

→ The null hypothesis is then rejected for those tests that give the $\ell$ smallest $p$-values

# Benjamini-Hochberg correction

- The Benjamini-Hochberg (BH) step-up procedure is commonly used in bioinformatics applications
- Let $q \in [0,1]$ be given and $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered list of the $m$ $p$-values, then the BH procedure works as follows
  1. Find the largest $k$ such that $p_{(k)} \leq \frac{k}{m} q$
  2. Then reject all $H_{(i)}$ for $i = 1, \ldots, k$
- For BH, the probability of expected proportion of false positives $\leq q$
- The FDR value $q_k$ for each test $k$ can be obtained from mapping

$$\min\left\{\frac{m}{k} p_{(k)}, 1\right\}$$

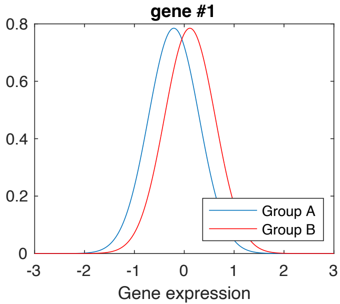
(and by guaranteeing that FDR values do not decrease as $k$ increases)

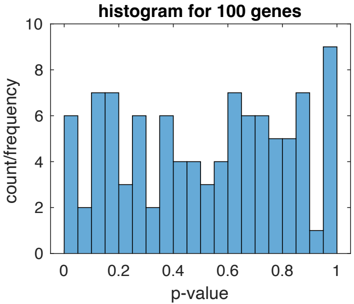
# False discovery rate

- An example: Following the above example with one gene, let us now assume that we measure the expression of 100 genes for two groups, $A$ and $B$. We assume to have five replicate measurements from both groups (for each of the 100 genes).
- For each gene, expression values are normally distributed with means $\mu_A$ and $\mu_B$ and standard deviations $\sigma_A = \sigma_B$.

# False discovery rate

- If $\mu_A = \mu_B = 0$ (and $\sigma_A = \sigma_B = 1$), the null hypothesis holds for all genes and in the ideal case we should not detect any differentially expressed genes
- However, the histogram of the obtained $p$-values look as follows (histogram on right)

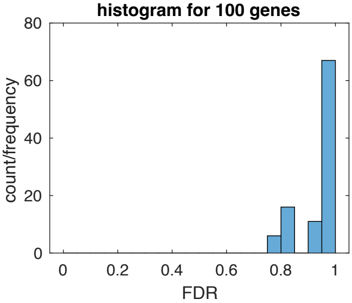




- We detect 6 genes with the significance level of 0.05 (all false positives)

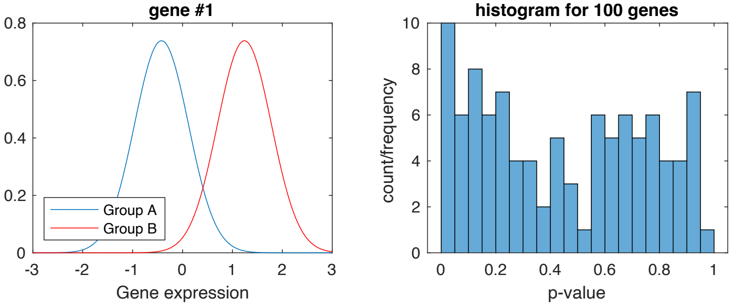
# False discovery rate

- If we correct the p-values for multiple testing using the Benjamini-Hochberg methods described above, we detect no genes that are statistically significantly differentially expressed.



# False discovery rate

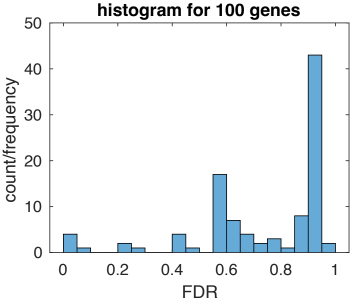
- Let us then see how FDR correction works if we have 90 non-differentially expressed genes and 10 truely differentially expressed genes with $\mu_A = 0$ and $\mu_B = 2$ (and $\sigma_A = \sigma_B = 1$) for the differentially expressed genes.



- We would now detect 10 genes with the significance level of 0.05: 7 true positives and 3 false positives

# False discovery rate

- If we correct the p-values for multiple testing using the Benjamini-Hochberg methods described above, we detect 4 genes that are statistically significantly differentially expressed (all true positives)

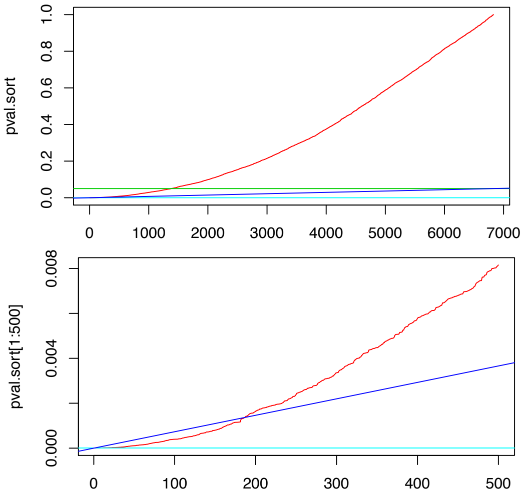


# False discovery rate

- Consider an example from (Wilkinson, 2017): use $t$-test to identify genes differentially expressed in melanoma compared to healthy skin cells
- 6830 genes, i.e., $m = 6830$
- If we assumed that the null hypothesis holds for all genes, then the expected number of false positives would be $6830 \cdot 0.05 = 341.5$
- Using the nominal (non-corrected) $p$-values results in 1377 significantly differentially expressed genes, indicating that the data may contain a considerable number of truly differential genes
- The use of Bonferroni correction would give us only six genes that meet the stringent criterion of $p \leq 0.05/6830 \approx 0.0000073$
- BH correction method would give us 186 differentially expressed genes with a FDR threshold of 0.05

# False discovery rate

- The figures below show
  - Ordered *p*-values (red)
  - The 0.05 uncorrected *p*-value cutoff (green)
  - The Bonferroni-corrected threshold (cyan)
  - The FDR threshold (dark blue)



Figures from (Wilkinson, 2017)

# References

- Hastie T, Tibshirani R, Friedman J, The Elements of Statistical Learning, Springer, 2009.
- Jeremy Orloff and Jonathan Bloom. “Null Hypothesis Significance Testing” I Class 17, 18.05, Spring 2014 (http://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading17b.pdf)
- Wilkinson DJ, Statistics for Big data Part 2: Multivariate Data Analysis using R (Lecture notes) available at https://www.staff.ncl.ac.uk/d.j.wilkinson/teaching/mas8381/notes14.pdf, November 19, 2017