Lecture 4: Statistics For Transcriptomics

BIOINF3005/7160: Transcriptomics Applications

Dr Stephen Pederson

Bioinformatics Hub,
The University of Adelaide

March 30th, 2020

Introduction

Introduction

Today we'll

- Discuss the relationship between our experiment and "truth"
- Revise Hypothesis Testing
- Introduce strategies for managing error rates
- Introduce the moderated T-test

Sampling and the Null Hypothesis

Sampling

Most experiments involve measuring something:

- Continuous values e.g. Ct values, **fluorescence intensity**
  - These values are often *Normally distributed*

Sampling

Most experiments involve measuring something:

- Continuous values e.g. Ct values, **fluorescence intensity**
  - These values are often *Normally distributed*
- Discrete values e.g. **read counts**, number of colonies
  - These values often involve rates, i.e. colonies/cm^2

Sampling

- We are always interested in the **true underlying values** from the **entire population**
- We use our **sample-derived estimates** (i.e. from our data) to make inference about the **true values**

Population Parameters

- Experimentally-obtained values represent an **estimate** of the true effect
  - More formally referred to as *population-level parameters*
- Every experiment is considered a *random sample of the complete population*
- Repeated experiments would give a **different** *(but similar)* estimate

Hypothesis Testing

In biological research we often ask:

"Is something happening?" or "Is nothing happening?"

Hypothesis Testing

In biological research we often ask:

"Is something happening?" or "Is nothing happening?"

We might be comparing:

- Cell proliferation in response to antibiotics in media
- Methylation levels across genomic regions
- Allele frequencies in two populations
- **mRNA abundance in two related cell types**

Hypothesis Testing

In biological research we often ask:

"Is something happening?" or "Is nothing happening?"

How do we decide if our experimental results are "significant"?

- Do our measurements represent normal variability?
- What would the data look like if our *experiment had no effect?*
- What would our data look like if there was *some kind of effect?*

The Null Hypothesis

- The *Null Hypothesis* (H_0) is used to describe the data if **nothing is happening**
- The *Alternate Hypothesis* (H_A) captures **all other possibilities**

The Null Hypothesis

- H_0: we have a test value (e.g. $\mu_0 = 0$) which allows us to define an expected distribution
  - This test value represents our population statistic of interest (e.g. logFC)
- H_A: Values which are unlikely to come from the defined H_0 distribution are assumed to come from H_A
  - H_A is every possibility besides no change $\implies$ *we can't define this statistically*

The Sample Mean

The Sample Mean

For normally distributed data, we usually make inference about a **mean** of some type:

- We have an experiment-specific **estimate** of the *mean logFC* ($\bar{x}$)
- We make inference about the **unknown** *true mean logFC* (μ)
- We use our 'best guess' of the value we care about, e.g. $\mu_0 = 0$

The Sample Mean

For normally distributed data, we usually make inference about a **mean** of some type:

- We have an experiment-specific **estimate** of the *mean logFC* ($\bar{x}$)
- We make inference about the **unknown** *true mean logFC* (μ)
- We use our 'best guess' of the value we care about, e.g. $\mu_0 = 0$
- In regression models, we fit **slope** and **intercept** terms
  - Principles introduced below are analogous: We *estimate* a *true* value

The Sample Mean

For normally distributed data, we usually make inference about a **mean** of some type:

- We have an experiment-specific **estimate** of the *mean logFC* ($\bar{x}$)
- We make inference about the **unknown** *true mean logFC* (μ)
- We use our 'best guess' of the value we care about, e.g. $\mu_0 = 0$
- In regression models, we fit **slope** and **intercept** terms
  - Principles introduced below are analogous: We *estimate* a *true* value

$$\bar{x} \sim \mathcal{N}(\mu, \mathrm{SE}_{\bar{x}})$$

The *standard error* of $\bar{x}$ ($\mathrm{SE}_{\bar{x}}$) represents how variable this value is around μ
e.g. $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where σ is population standard deviation

The Sample Mean

- If we know the population variance (σ^2), and have our sample size (n)
  - We **almost never** know σ and **never** know μ
- We can then use our *value of interest*, e.g. $\mu_0 = 0$
  - This is the value that we expect if H_0 is true

$$\bar{x} \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$$

The Sample Mean

$$\bar{x} \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$$

The Sample Mean

$$\bar{x} \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$$

$$\bar{x} - \mu_0 \sim \mathcal{N}(0, \frac{\sigma}{\sqrt{n}})$$

The Sample Mean

$$\bar{x} \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$$

$$\bar{x} - \mu_0 \sim \mathcal{N}(0, \frac{\sigma}{\sqrt{n}})$$

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

Hypothesis Tests

The Sample Mean

If we know the population variance (σ^2), and have our sample size (n)

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

- We use this as the underlying principle for H_0
- We don't know μ but we have a value (μ_0) of interest (usually $\mu_0 = 0$)

So if H_0 is true, **we know what kind of distribution our data will be drawn from**

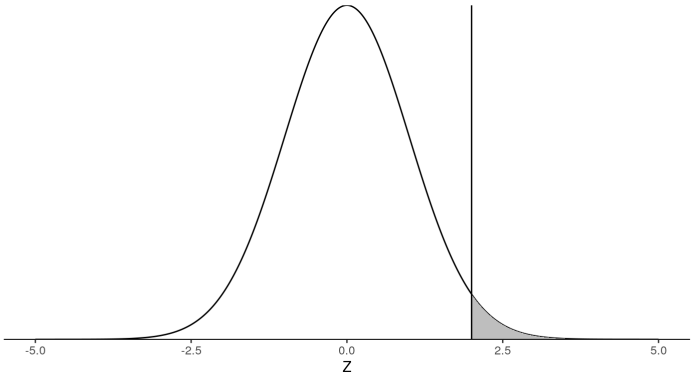
P Values

Once we can calculate a Z-score, we compare this to $\mathcal{N}(0, 1)$ and ask:

How likely are we to see this Z-score if H_0 is true?

P Values

If we obtain $Z = 2$

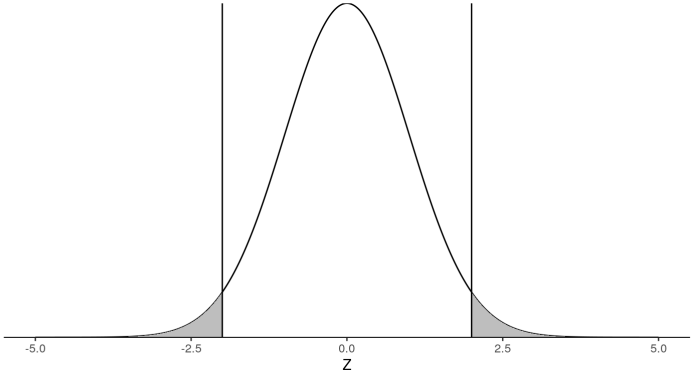


P Values

- The shaded area is the probability of obtaining $Z > 2$, assuming H_0 is true
- Most of the time we are look for $H_A : \mu_0 \neq 0$ so we need to look on *both sides*
- This is known as a **two-sided test**
- Can also be described as $|Z| > 2$

P Values

$$P(|Z| > 2) = 0.0455$$



P Values

$$P(|Z| > 2) = 0.0455$$

- So if H_0 is true, we would see $Z > 2$ about 4.5 times every 100 experimental repeats
- We could then choose to accept H_0 as the most likely truth, or reject H_0 as the most likely truth
- How do we know if we have one of the 4.5 in 100?

P Values

$$P(|Z| > 2) = 0.0455$$

- So if H_0 is true, we would see $Z > 2$ about 4.5 times every 100 experimental repeats
- We could then choose to accept H_0 as the most likely truth, or reject H_0 as the most likely truth
- How do we know if we have one of the 4.5 in 100?
  - We don't know
  - Often set $p < 0.05$ as the rejection value (i.e. $\alpha = 0.05$)

P Values

Definition

A p-value is the probability of obtaining data as extreme, or more extreme than we have, if H_0 is true

Hypothesis Testing

1. We have defined what our data should look like under H_0
2. We have determined how likely we are to see our results
3. We accept or reject H_0 if $p < \alpha$

Hypothesis Testing

In our context

- We are usually comparing μ_1 against $\mu_2 \implies \mu_1 - \mu_2 = 0$
  - This would be the expression level in two groups/conditions/treatments etc
  - $\mu_1 - \mu_2 = 0$ is testing logFC $= 0$

Hypothesis Testing

In our context

- We are usually comparing μ_1 against $\mu_2 \implies \mu_1 - \mu_2 = 0$
  - This would be the expression level in two groups/conditions/treatments etc
  - $\mu_1 - \mu_2 = 0$ is testing logFC $= 0$
- We don't know the population variance (σ)
- We estimate σ using our sample variance $\implies$ T-tests

T-Tests

A T-test is very similar to a Z-test

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim \mathcal{T}_\nu$$

The value ν means 'degrees of freedom'

The Sample Variance

To calculate the sample variance (s^2) for a set of values $x = (x_1, x_2, \ldots, x_n)$

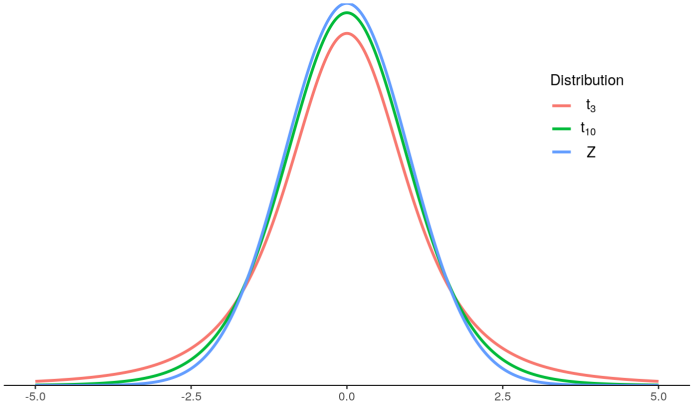
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Degrees of Freedom

- The degrees of freedom (ν) describe how 'fat' the tails of a T-distribution are
  - As $\nu \uparrow$ the tails become 'less fat'
- The more individual samples we have (n), the more degrees of freedom we have
  - The more samples we have, the less likely we are to see extreme values
  - Commonly $\nu = n - 1$

T-Distributions



T-Tests

1. We now compare our T statistic to the appropriate T distribution
2. Find the probability (p) of observing data as (or more) extreme if H_0 is true
3. Accept or reject H_0

Transcriptomics

- For Microarrays (i.e continuous data) we simply perform a T-test for every gene
- Expression estimates are analysed on the $\log_2$ scale

$$H_0 : \mu_1 - \mu_2 = 0 \text{ Vs } H_A : \mu_1 \neq \mu_2$$

- The expression estimates $\bar{x_1}$ and $\bar{x_2}$ estimate μ_1 and μ_2

Transcriptomics

- We will also have a sample variance for each group s_1 and s_2
  - Sample variances are assumed to be equal between groups
  - We pool sample variances ($s_p = \ldots$)
  - $\nu = n_1 + n_2 - 2$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Transcriptomics

- We often have $k > 2$ groups $\implies$ multiple pairwise comparisons
- Or some kind of regression model with discrete predictors (i.e. group-wise)

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

In transcriptomics we usually refer to this as the *residual variance*

Multiple Testing

P Values

- We perform '000's of T-tests in every experiment (one per gene)
- A p-value of 0.05 $\implies$ 1 in 20 times we will see data this (or more) extreme **if** H_0 **is true**
- A p-value of 0.01 $\implies$ 1 in 100 times we will see data this (or more) extreme **if** H_0 **is true**
- So if we have 10,000 genes **for which** H_0 **is true**, how many times will we see:
  - $p < 0.05$
  - $p < 0.01$

P Values

- We perform '000's of T-tests in every experiment (one per gene)
- A p-value of 0.05 $\implies$ 1 in 20 times we will see data this (or more) extreme **if** H_0 **is true**
- A p-value of 0.01 $\implies$ 1 in 100 times we will see data this (or more) extreme **if** H_0 **is true**
- So if we have 10,000 genes **for which** H_0 **is true**, how many times will we see:
  - $p < 0.05 \implies \sim 500$ times
  - $p < 0.01 \implies \sim 100$ times

Error Rates

- If we reject H_0 using $p < 0.05 \implies \sim 500$ errors (false rejections)
- If we reject H_0 using $p < 0.01 \implies \sim 100$ errors (false rejections)

These are known as *Type I* errors

- In biological research, these can waste $$$
- We need to control these errors

Error Rates

	H_0 True	H_0 Not True
Reject H_0	Type I Error	✓
Accept H_0	✓	Type II Error

We need to **minimise both Type I and Type II** errors

Error Rates

Two primary strategies for controlling error rates

1. Bonferroni's Method
   - This sets the bar very high to reject H_0
   - Big increase in Type II errors
2. False Discovery Rate
   - Allows a small number of false discoveries
   - Reduces Type II errors (compared to Bonferroni)

Error Rate

The Bonferroni Adjustment

- If you have $m = 10,000$ tests and $\alpha = 0.05$
- Set $\alpha_{\text{bonf}} = \frac{\alpha}{m} = \frac{0.05}{10000} = 5 \times 10^{-7}$
- Alternatively, adjust each p-value: $p_{\text{bonf}} = \min(1, m * p)$

Error Rate

The Bonferroni Adjustment

- If you have $m = 10,000$ tests and $\alpha = 0.05$
- Set $\alpha_{\text{bonf}} = \frac{\alpha}{m} = \frac{0.05}{10000} = 5 \times 10^{-7}$
- Alternatively, adjust each p-value: $p_{\text{bonf}} = \min(1, m * p)$

The Family-Wise Error Rate (FWER)

- The effect is P(one Type I Error) ≤ 0.05
- This is strict control of the *family-wise error rate*
- The family is the complete set of m tests

The False Discovery Rate

- *False Discovery Rate* strategies are very common in transcriptomics
- We allow a small amount of noise into our results *implies* signal still swamps noise
  - An $FDR = 0.05 \implies \leq 5\%$ of our results are 'false discoveries' (Type I Errors)
- The most common method is the Benjamini-Hochberg method[1]
- Other methods include Storey's q-value[2]
- These methods **do not control** the *FWER* but **do control** the *FDR*

[1] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 00359246. URL: http://www.jstor.org/stable/2346101.

[2] John D. Storey and Robert Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445. ISSN: 0027-8424. DOI: 10.1073/pnas.1530509100. eprint: https://www.pnas.org/content/100/16/9440.full.pdf. URL: https://www.pnas.org/content/100/16/9440.

Moderated T-tests

Moderated T-tests

- When conducting our m simultaneous T-tests, we use an estimate of the population variance s_p
  - Some of these are going to be larger than the true population value
  - Others are going to be smaller than the true population value

Moderated T-tests

- When conducting our m simultaneous T-tests, we use an estimate of the population variance s_p
  - Some of these are going to be larger than the true population value
  - Others are going to be smaller than the true population value
- What impact will this have?

Moderated T-tests

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- If $s_p <<< \sigma \implies T \uparrow$
- We may get significant results with small logFC, due to small variances

Moderated T-tests

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- If $s_p <<< \sigma \implies T \uparrow$
- We may get significant results with small logFC, due to small variances
- If $s_p >>> \sigma \implies T \downarrow$
- We will miss truly DE genes due to large variances

Moderated T-tests

- This situation exists in **every** T-test
- In transcriptomics, we perform '000's in parallel
- We can take advantage of this $\implies$ *Empirical Bayes* model
- This gives us a moderated value of s_p $\implies$ *Moderated T-test*[3]

[3] G. K. Smyth. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". In: *Stat Appl Genet Mol Biol* 3 (2004), Article3.

Moderated T-tests

- Variances are usually drawn from a Scaled Inverse χ^2 distribution
- Given that we have '000's of genes, we can estimate the hyperparameters for a Bayesian Model
- We end up with a *posterior estimate* known as the moderated variance ($\tilde{s}_p^2 = E[\sigma^2 | s_p^2]$)
- Overestimates/Underestimates are shrunk towards the mean
- Increases Power ($\downarrow$ Type II Errors) and Decreases Type I Errors