t-tests, one gene

Differential Expression

Kasper D. Hansen

Fall 2023

One gene, two groups

We start by discussing the classic t-test for 1 gene. In doing so, we will touch on many of the important themes of statistical genomics, while keeping complications to a minimum.

We are comparing two groups of samples.

Each group of samples represents a **population** of samples (for example control mice). Each population has a mean (average expression level), which we denote by μ_1, μ_2.

Our question of interest is

$$H_0 : \qquad \mu_1 = \mu_2$$

This is called a **null hypothesis**.

Parameters and estimators

We have two different empirical group averages, which we write as $\bar{X}_1, \bar{X}_2$.

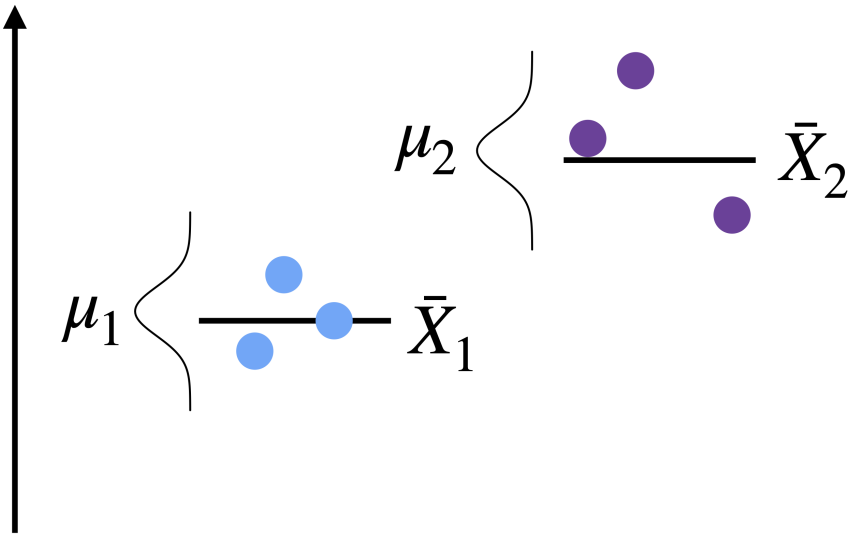
These averages are **estimates** of the population means μ_1, μ_2; these are called **parameters**.

There is also a parameter σ^2 which represents the **variance** of the data. We're not interested in this parameter.

- Parameters: μ_1, μ_2, σ^2
- Estimators of parameters: $\bar{X}_1, \bar{X}_2, s^2$

We are assuming the two groups have the same variance.

Example



Fold-change

It is natural to compute

$$D = \bar{X}_1 - \bar{X}_2$$

which estimates $\mu_1 - \mu_2$. If the data is on the $\log_2$-scale (which is standard in genomics), this is the log2-fold-change.

Usually a log2-fold-change bigger than 1 (or smaller than -1) is considered "large"; this is the same as a fold-change of 2.

Why 2?

- It is suggested by theory (haplo-insufficiency).
- Practice tells us that a fold-change of 2 or greater is "large" in bulk RNA-seq.

Fold-change and bias cancellation

Bias is abundant in genomics. A first approximation is that each gene g has its own bias term b_g which describes a systematic error in measuring the gene g, and let us assume this bias is **not** sample-dependent.

If the bias is additive (corresponding to a multiplicative bias on the original scale) we are thinking of a model where

$$\mu_1 = \theta_1 + b, \qquad \mu_2 = \theta_2 + b$$

in which case, the contrast

$$D = \bar{X}_1 - \bar{X}_2$$

estimates

$$\mu_1 - \mu_2 = (\theta_1 + b) - (\theta_2 + b) = \theta_1 - \theta_2$$

In other words, an additive gene-specific bias – which is not sample-specific – cancels out when forming (log) fold-changes.

This is a very important property of (log) fold-changes which explains their popularity.

Testing the hypothesis

Remember, we're interested in the hypothesis

$$H_0 : \qquad \mu_1 = \mu_2$$

and we have the quantify

$$D = \bar{X}_1 - \bar{X}_2$$

which estimates $\mu_1 - \mu_2$

We will reject the hypothesis if D is large (precisely: far away from zero), but what does that mean?

The t-statistic

The first thing we realize is that the variance of D depends on the data (or more precise on the variance of the data σ^2. This is a problem.

To solve this, we estimate the variance of D (or more precisely the standard deviation) and form

$$t = \frac{D}{\text{sd}(D)}$$

This is called the t-statistic. It is an example of a **test statistic**. A test statistical is a quantity where values far away from zero is

Because we divide with $\text{sd}(D)$, the variance of t is 1.

The null distribution

To precisely answer "what is a large value of t?" we need a null distribution. This is the distribution we use to measure "large".

The null distribution is the distribution of the test statistic under the null hypothesis:

$$t \sim P_{\text{null}} \qquad \text{or} \qquad t \overset{\text{approx}}{\sim} P_{\text{null}}$$

We then compute a p-value

$$p = P_{\text{null}}(t \geq t_{\text{obs}})$$

where t_{obs} is the observed value of the t-statistic. The interpretation is the chance of observing a statistics this extreme (or worse) assuming the null hypothesis is true.

Where do we get the null distribution from

There are different ways to obtain a null distribution:

1. [Theory]. If the observations are Gaussian, the null distribution is a t-distribution.
2. [Theory, large sample]. If the number of observations is large, it is possible to obtain a limiting distribution.
3. [Permutation]. You can use a permutation approach to get a null distribution.

When we refer to a **t-test**, we usually refer to situation 1.

The state of the world

Let us discuss errors:

Decision	True DE (H_A)	True non-DE (H_0)
Significant	correct	error, type I
Not significant	error, type II	correct

Errors of type I: False positives

Errors of type II: False negatives

Error rates

We are interested in controlling the type I error rate (or the chance of a false positive). This means that - assuming our hypothesis is true - we should at most reject the null hypothesis (make an error) in α percentage of tests. Specifically, we pre-specific a *nominal* level α (a p-value cutoff).

When the null distribution is exact, we always control the error rate, in fact we have

$$P_{\text{null}}(\text{type I error}) = \alpha$$

I'll call this a calibrated test if we have equality (or almost equality)

If the null distribution is approximate, at most we can hope to control the error rate

$$P_{\text{null}}(\text{type I error}) \leq \alpha$$

Power

P_{alt} is the distribution of the statistic under the alternative. This is not clearly defined; there are many possible alternatives (how big is $\mu_1 - \mu_2$?). We need to select a specific alternative hypothesis.

$$1 - P_{\text{alt}}(\text{type II error})$$

is called **power**. It reflects our ability to identify true changes. Power is very important for grant writing.

Power depends on the actual effect size, the variation in the data, the sample size and the model.

We want to control the type I error rate while maximizing power. We want to control the type I error rate because that is the definition of a p-value.

Statistical concepts

- Hypothesis
- Test statistic
- Null distribution
- When the null hypothesis is true, and the test is calibrated, the p-value is uniformly distributed
- Trade-off between false positives and false negatives described through power
- In statistics, we aim to control the chance of a false positive

Experimental design

Let us briefly return to our t-statistic

$$t = \frac{D}{\sqrt{D}}$$

which we can also view as an estimate of

$$\frac{\mu_1 - \mu_2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sqrt{\sigma^2}}$$

(n_1, n_2 is the sample size in the two groups of samples). This is important for experimental design. a - $\mu_1 - \mu_2$ (the effect size) is given by the question. - n_1, n_2 are the number of replicates. - σ^2 is the variance of the data.

Typically, we can increase the sample size or try to reduce the variability.