# Hypothesis Testing

Exploratory Data Analysis with **R**

Oedipus ponders the riddle of the Sphinx. Classical (~400 BCE)

Boris Steipe

UNIVERSITY OF TORONTO

Department of Biochemistry
Department of Molecular Genetics

This module includes some material originally developed by Raphael Gottardo, FHCRC and by Sohrab Shah, UBC

## Objectives

Understand the principal idea behind a statistical test;

Know about the concepts true/falsepositives/nergatives, *p*-value, and significance;

Be able to apply simple parametric and non-parametric test to your data;

Know how to interpret the results;

Understand the problem behind *multiple testing*;

Know what to do about it in the context of expression data analysis.



## Hypothesis testing

Once we have a statistical model that describes the distribution of our data, we can explore data points with reference to our model.

In **hypothesis testing** we typically ask questions such as:

Is a particular sample a part of the distribution, or is it an outlier?

Can two sets of samples have been drawn from the same distribution, or did they come from different distributions?



## Hypothesis testing

Hypothesis testing is *confirmatory data analysis*, in contrast to *exploratory data analysis.*

Concepts:

- Null and Alternative Hypothesis
- Region of acceptance / rejection and critical value
- Error types
- *p* - value
- Significance level
- Power of a test (1 - false negative)



## Null hypothesis / Alternative hypothesis

The **null hypothesis** $H_0$ states that nothing of consequence is apparent in the data distribution. The data corresponds to our expectation. We learn nothing new.

The **alternative hypothesis** $H_1$ states that some effect is apparent in the data distribution. The data is different from our expectation. We need to account for something new. Not in all cases will this result in a new model, but a new model always begins with the observation that the old model is inadequate.



## Test types

Just like the large variety of types of hypotheses, the number of test is large. The proper application of tests can be confusing and it is easy to make mistakes.

Common types of tests

A **one-sample test** compares a sample with a population.

A **two-sample test** compares samples with each other.

**Paired sample tests** compare matched pairs of observations with each other. Typically we ask whether their difference is significant.

...

## Test types

... common types of tests (as you would find them in a statistics textbook...)

A ***Z*test** compares a sample mean with a normal distribution.

A ***t*test** compares a sample mean with a *t*-distribution and thus relaxes the requirements on normality for the sample.

**Nonparametric tests** can be applied if we have no reasonable model from which to derive a distribution for the null hypothesis.

Chi**squared tests** analyze whether samples are drawn from the same distribution.

**F-tests** analyze the variance of populations (ANOVA).

...



## The Hypothesis Test principle

Think about what hypothesis testing really means.

- You have some observation;
- You have a model of the data;
- You ask about the probability that the model of your data would contain your observation.

...



## Error types

| Truth / Decision | $H_0$ | $H_1$ |
|---|---|---|
| Accept $H_0$ | $1 - \alpha$ | $\beta$ "False negative" "Type II error" |
| Reject $H_0$ | $\alpha$ "False positive" "Type I error" | $1 - \beta$ |



## Introduction

One sample and two sample t-tests are used to test a hypothesis about the mean(s) of a distribution.

Gene expression: Is the mean expression level under condition 1 different from the mean expression level under condition 2?

Assume that the data are from a normal distribution.



## one sample t-test

*t*-tests apply to *n* observations that are **independent** and **normally distributed** with equal variance about a mean $\mu$.

$$H_0 : \mu = \mu^0 \qquad H_1 : \mu \neq \mu^0$$

The 1-sample *t*-statistic is defined as:

$$t = \frac{\bar{y} - \mu^0}{SE_{\bar{y}}} \gg \frac{\bar{y} - \mu^0}{s/\sqrt{n}}$$

i.e. t is the difference in sample mean and $\mu^0$, divided by the *Standard Error of the Mean*, to penalize noisy samples.

If the sample mean is indeed $\mu^0$, *t* follows a *t*-distribution with n-1 degrees of freedom.



## what is a *p*value?

a) A measure of how much evidence we have against the alternative hypothesis.
b) The probability of making an error.
c) Something that biologists want to be below 0.05 .
d) The probability of observing a value as extreme or more extreme by chance alone.
e) All of the above.

## twosample *t*test

Test if the means of two distributions are the same.

The datasets $y^i_1, ..., y^i_n$ are independent and normally distributed with mean $\mu_i$ and variance $^2$, $N(\mu_i, ^2)$, where $i$=1,2.

In addition, we assume that the data in the two groups are independent and that the variance is the same.

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$



## twosample *t*test

2 sample *t*-test

$$t = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Distance between the two sample means

Penalized noisy samples (Stand. error of the mean difference )

If the means are equal, *t* follows a *t*-distribution with $n_1+n_2-2$ degrees of freedom

*p*-value $p = 2 \cdot \Pr(|T_{n_1+n_2-2}| > |t|)$



## ttest assumptions

**Normality:** The data need to be Normal. If not, one can use a transformation or a non-parametric test. If the sample size is large enough (n>30), the t-test will work just fine (CLT).

**Independence:** Usually satisfied. If not independent, more complex modeling is required.

**Independence between groups:** In the two sample t- test, the groups need to be independent. If not, one can use a paired t-test.

**Equal variances:** If the variances are not equal in the two groups, use Welch.s t-test (default in R).



## nonparametric tests

Non-parametric tests constitute a flexible alternative to *t*-tests if you don.t have a model of the distribution.

In cases where a parametric test would be appropriate, non-parametric tests have less power.

Several non parametric alternatives exist e.g. the Wilcoxon and Mann-Whitney tests.

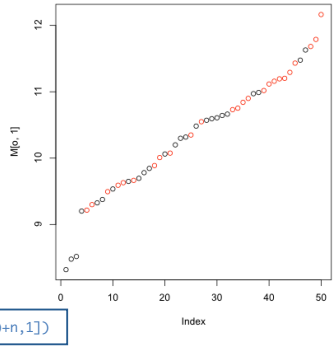


## Wilcoxon test principle

```
o <- order(M[,1])
plot(M[o,1], col=M[o,2])
```

For each observation in **a**, count the number of observations in **b** that have a smaller rank.

The sum of these counts is the test statistic.




```
wilcox.test(M[1:n,1], M[(1:n)+n,1])
```



## permutation test

A *p*-value characterizes where an observation lies with reference to the distribution of our statistics **under the null hypothesis**.

How can we estimate the null distribution?

In the two sample case, to simulate the null distribution, one could simply **randomly permute** the group labels and recompute the statistics.

Repeat this for a (sufficienty large) number of permutations and compute the number of times you randomly observed a value as extreme or more extreme than the observation of interest.

## permutation test

For data that has multiple "categories" associated with each observation:

Select a statistic (e.g. mean difference, *t* statistic)

Compute the statistic for the observation of interest *t*.

For a number of permutations $N_p$

Randomly permute the labels and compute the associated statistic $t_i^0$

Count how often the statistic exceeds the observation

$$p(t) = \frac{\#\left(t_i^0 > t\right)}{N_p}$$

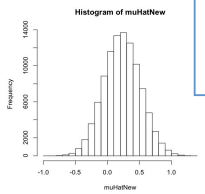


## the Bootstrap

The basic idea is to resample the data we have observed and compute a new value of the statistic/estimator for each resampled data set.

Then one can *assess the estimator* by looking at the empirical distribution across the resampled data sets.

```
set.seed(100)
x <- rnorm(15)
muHat <- mean(x)
sigmaHat <- sd(x)
Nrep <- 100
muHatNew <- rep(0, Nrep)
for(i in 1:Nrep) {
  xNew <- sample(x, replace=TRUE)
  muHatNew[i] <- median(xNew)
}
se <- sd(muHatNew)
muHat
se
```






## statistical "power"

The **power** of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false (i.e. that it will not make a Type II error, or a **false negative** decision). As the power increases, the chances of a Type II error occurring decrease. The probability of a Type II error occurring is referred to as the false negative rate ($\beta$). Therefore power is equal to 1 $\beta$, which is also known as the **sensitivity**.

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size. In addition, the concept of power is used to make comparisons between different statistical testing procedures: for example, between a parametric and a nonparametric test of the same hypothesis.

From Wikipedia Statistical_Power



## One sample *t*-test power calculation

1 sample *t*-test: $t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}$

If the mean is $\mu_0$, *t* follows a *t*-distribution with n-1 degrees of freedom.

If the mean is not $\mu_0$, *t* follows a non central *t*-distribution with n-1 degrees of freedom and noncentrality parameter
$(\mu_1 - \mu_0)$ x $(s/\sqrt{n})$.



## Power, error rates and decision

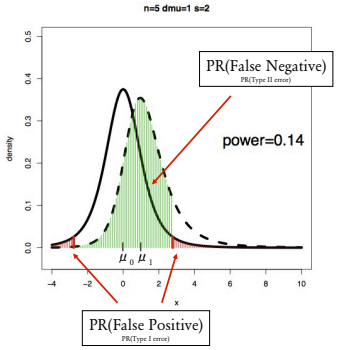
Power calculation in R:

```
> power.t.test(n = 5, delta = 1, sd=2,
               alternative="two.sided", type="one.sample")

     One-sample t test power calculation

              n = 5
          delta = 1
             sd = 2
      sig.level = 0.05
          power = 0.1384528
    alternative = two.sided
```

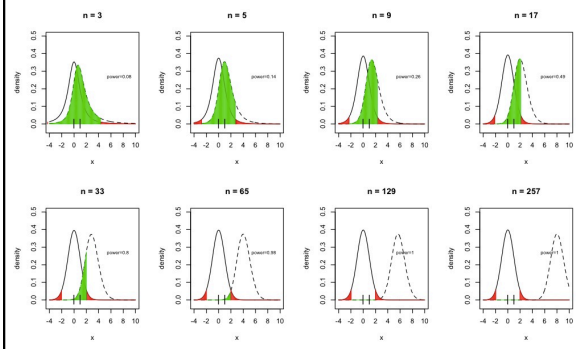
Other tests are available see `??power`.



## Power, error rates and decision



## Power, error rates and decision






## multiple testing

**Single hypothesis testing**

Fix the False Positive error rate (eg. = 0.05).

Minimize the False Negative (maximize sensitivity)

This is what traditional testing does.

What if we perform many tests at once? Does this affect our False Positive rate?



## multiple testing

With high-throughput methods, we usually look at a very large number of decisions for each experiment. For example, we ask for every gene on an array whether it is significantly up- or downregulated.

This creates a *multiple testing paradox*. The more data we collect, the harder it is for every observation to appear significant.

Therefore:

We need ways to assess error probability in multiple testing situations correctly;

We need approaches that address the paradox.



## FWER

The **FamilyWise Error Rate** is the probability of having at least one False Positive (making at least one type I error) in a "family" of observations.

Example: Bonferroni multiple adjustment.

$$p_g = N \times p_g$$

If $p_g$ then FWER

This is simple and conservative, but there are many other (more powerful) FWER procedures.



## False Discovery Rate (FDR)

The FDR is the proportion of False Positives among the genes called *differentially expressed* (DE).

Order the *p*-values for each of *N* observations:

$P_{(1)} \cdots P_{(i)} \cdots P_{(N)}$

Let $k$ be the largest $i$ such that $p_{(i)} \; i / N \times \alpha$

... then the FDR for genes 1 ... $k$ is controlled at $\alpha$.

Hypotheses need to be independent!

FDR: Benjamini and Hochberg (1995)



## SAM

SAM (Significance Analysis of Microarrays) is a statistical technique to find significant expression changes of genes in microaray experiments.

The input is an expression profile. SAM measures the strength of the association of the expression value and the conditions of the expression profile.

SAM employs a **modified *t*-statistic** that is more stable if the number of conditions is small.

False Discovery Rates are estimated through permutations.

```
library(samr)
?samr
?SAM
```

## summary

| Number of tests \ Sample size | n < 30 | n 30 |
|---|---|---|
| p = 1 | non-parametric *t*-test/*F*- test | *t*-test, *F*- test |
| p > 1 | regularized t-test/F-test (e.g. SAM, limma) + multiple testing | *t*-test, *F*- test + multiple testing |

**Multiple testing:**

If hypotheses are independent or weakly dependent use an FDR correction, otherwise use Bonferroni.s FWER.

For more complex hypotheses, try an ANOVA (p=1) or limma (p>1).



## FROM HERE ...

**Get a book.** (e.g. Peter Dalgaard, Introductory Statistics with R is available online through UofT library)

**Simulate your data.** (Don't just use the packaged functions.)

**Have fun.**



boris.steipe@utoronto.ca