Statistics primer: Hypothesis testing

Laurent Gatto

January 20, 2016

Foreword

> *the function of significance tests is to prevent you from making a fool of yourself, and not to make unpublishable results publishable [1]*

at the end of the course, we will understand

> *Why Most Published Research Findings Are False. [3]*

and will pledge to not make fools of ourselves too frequently.

Statistical hypothesis testing

- Testing a working hypothesis
- Contrast an observed result to a *comparable* random distribution (assumptions!). If it is *different enough* from what we would expect by chance, then we might have an interesting result.

The testing process

1. Research hypothesis, as defined by our experimental design
2. State the *null hypothesis* H_0 (no difference with random distribution) and *alternative hypothesis* H_1 (there is a difference with random distribution)
3. Consider statistical assumptions and decide which test in appropriate.
4. Perform test and carefully interpret results (see below)

Choosing a test statistic

- Continuous data (such as micro-array, quantitative proteomic, . . .) and we want to compare means: *t-test*
- Count data (high-throughput sequencing) and we want to compare the number of reads between two conditions: *negative binomial*
- Gene set enrichment (is there an enrichment of genes with a specific funtion in my set of interesting genes): *hypergeometric*

Check what is used in the literature!

Type of tests: (non-)parametric

- ▶ Parametric: assumption is that the data comes from a population that follows a probability distribution
- ▶ Non-parametric: no defined/fixed parameters

When assumptions are correct, a parametric test has more *statistical power*.
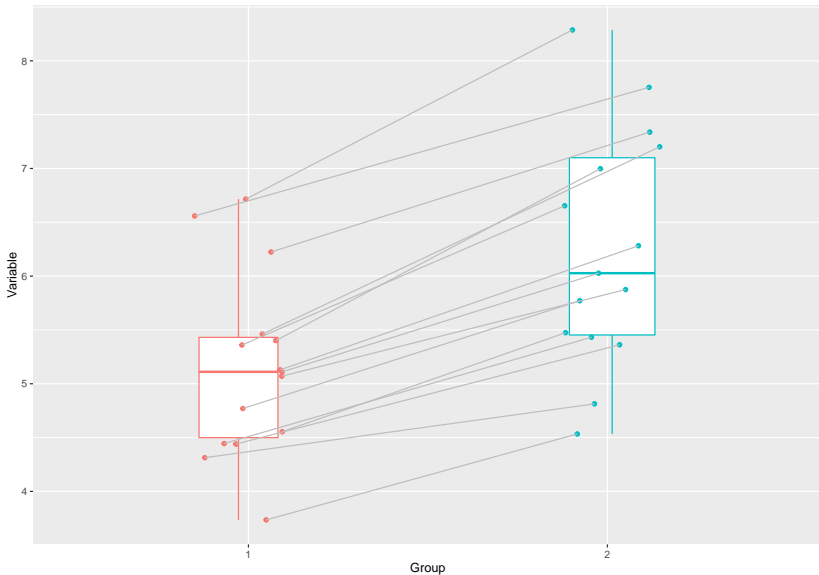
In practice, as most studies are under-powered, non-parametric tests are not an option.

Types of tests: One-sample or two-sample tests
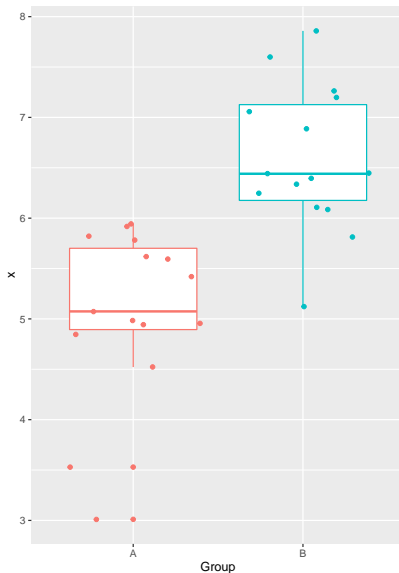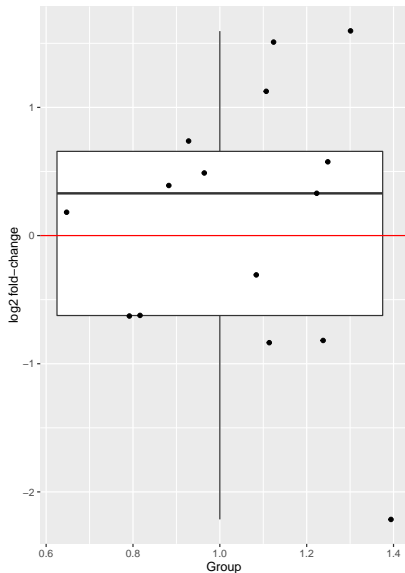
Types of tests: paired tests

Measurements in two conditions are paired.

Working example: Student's t-test

Comparing means from 2 groups, continous data.

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

Assumptions:

- data is normally distributed
- data are independent and identically distributed
- equal or un-equal (Welch test) variance
- t-test is robust to deviations.
- (All models are wrong. Some are useful)
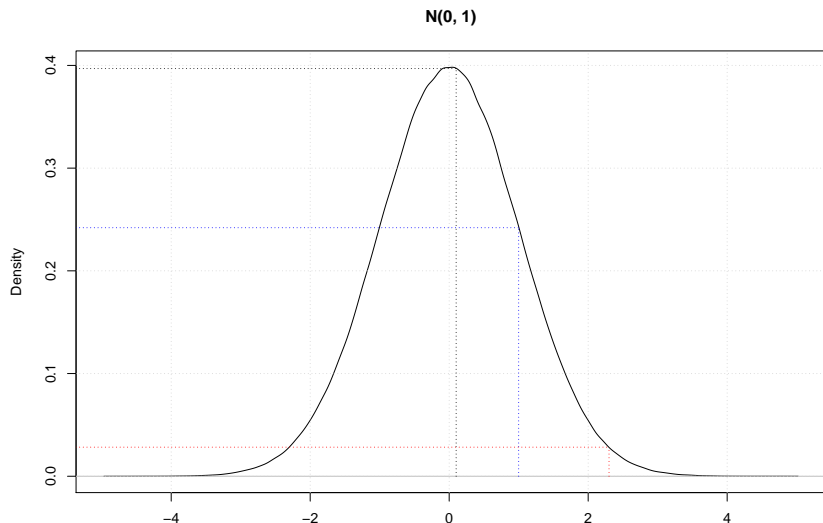
Welch test (t-test with unequal variances)

Two samples of sizes n_1 and n_2

$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}_1 - \bar{x}_2}}$

where

$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_{x1}^2}{n_1} + \frac{s_{x2}^2}{n_2}}$

How to we estimate how different we are from *random*

N(0, 1)

p-value: how (un)likely it would be to observe a value as extreme or more extreme under H_0.

P-value

p-value: how (un)likely it would be to observe a value as extreme or more extreme under H_0.

If p-value $<=$ than (arbitrary) significance level α, then we reject H_0.

What can go wrong

	H0 is true	H0 is false
H0 is rejected	Type I error, FP	correct TP
H0 not rejected	correct TN	Type II error, FN

power of a test $= 1$ - *type II*

(FP: false postitive, TP: true positive, FN: false negative, TN: true positive)

In R

```r
set.seed(1)
x <- rnorm(15, mean = 7, sd = 1)
y <- rnorm(15, mean = 5, sd = 1.2)
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 5.4214, df = 27.998, p-value = 8.778e-06
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  1.259224 2.788685
## sample estimates:
## mean of x mean of y
##  7.100843  5.076888
```

But

Is the p-value really what we want? What about the power of our test?

References

[1] David Colquhoun *An investigation of the false discovery rate and the misinterpretation of p-values* R. Soc. open sci. 2014 1 140216; `doi:10.1098/rsos.140216`.

[2] Regina Nuzzo *Scientific method: Statistical errors* 2014 Nature 506, 150–152 `doi:10.1038/506150a`

[3] Ioannidis JPA Why Most Published Research Findings Are False. 2015 PLoS Med 2(8): e124. `doi:10.1371/journal.pmed.0020124`.