Exploring the replication crisis: causes, consequences, possibilities

Sacha Dedeken

Supervised by Charlotte Baey

sacha.dedeken.etu@univ-lille.fr

1er avril 2021

Abstract

science goes wrong, $p<0.05$

Key words — p-value, hypothesis testing, bayesian testing, replication crisis

Université de Lille

Master 1 Data Science

Contents

1 Introduction

2 Actual situation in experimental sciences

2.1 The normal process of testing and publication

Let's start by presenting the main approach of research in experimental sciences. The vast majority of articles address an issue, and try to answer it by collecting quantitative data. The idea is to reformulate the issue in the form of a statistical question, to carry out a statistical test to provide an answer to the statistical question. The results of this test are then discussed and interpreted in order to formulate a conclusion to the problem. Let us illustrate this process with an example:

Doctors who want to study the effectiveness of drug X on hypothyroidism (insufficient thyroid hormone production) are faced with a complex question. They can rephrase it as :

« Is the average T3 hormone level measured in patients who have been treated with X for one month higher than a placebo control group? »

and get a statistical question. It is now possible to set up an experimental protocol with 2 groups, and to collect measurements to answer the statistical question. We will obtain 2 sets of measures of size n :

$\mathcal{X}_n : \{X_1, ..., X_n\}$ with X_i the T3 hormone level measured in the ith patient of the test group

$\mathcal{Y}_n : \{Y_1, ..., Y_n\}$ with Y_i the T3 hormone level measured in the ith patient of the placebo group

The challenge of inferential statistics is to consider the individuals in each group as representative of a random draw from the overall population. We will therefore consider X_n and Y_n as sets of random variables, independent and identically distributed (*iid*), following an unknown law of mean μ_x (resp. μ_y) and of variance assumed to be equal to σ^2. We will also measure statistical indicators on our observations: the empirical means $\overline{X_n}$ and $\overline{Y_n}$ and the overall corrected empirical variance S^2. It is now possible to carry out a statistical test on these data, taking into account sampling fluctuations, putting two hypotheses in competition :

- The null hypothesis, considered to be true a priori (here: the treatment has no effect)

$$\mathcal{H}_0 : \mu_x = \mu_y \iff \mu_x - \mu_y = 0$$

- The alternative hypothesis, complementary to H0 (here: the treatment has an effect)

$$\mathcal{H}_1 : \mu_x > \mu_y \iff \mu_x - \mu_y > 0$$

To make a decision between these two hypotheses, the researchers will then choose a test statistic on their samples (here $T = \overline{X_n} - \overline{Y_n}$), and express its law under $\mathcal{H}_0$.

$$\overline{X_n} \overset{H_0}{\sim} \mathcal{N}(\mu_x, \frac{\sigma^2}{n})$$
$$\overline{Y_n} \overset{H_0}{\sim} \mathcal{N}(\mu_y, \frac{\sigma^2}{n})$$
$$\overline{X_n} - \overline{Y_n} \overset{H_0}{\sim} \mathcal{N}(\mu_x - \mu_y, \frac{2\sigma^2}{n})$$

Under the null hypothesis, the test statistic will be close to 0, i.e. the greater the T, the less likely the data will be under $\mathcal{H}_0$. The next step is therefore to find a c threshold separating the acceptance region (the t values for which $\mathcal{H}_0$ is conserved) from the rejection region (the t values for which $\mathcal{H}_0$ is rejected). Whatever the chosen threshold, there are 2 possible errors:

- The type I error, α, which corresponds to the risk of wrongly rejecting the null hypothesis (here consider that treatment X increases the hormone level when in reality it does not).

$$\alpha = P(R\mathcal{H}_0|\mathcal{H}_0) = P(T > c|\mu_x = \mu_y)$$

- The type II error, β, which corresponds to the probability of wrongly conserved the null hypothesis (here consider that treatment X does not increase hormone levels when in fact it does).

$$\beta = P(\cancel{R\mathcal{H}_0}|\mathcal{H}_1) = P(T < c|\mu_x > \mu_y)$$

The approach towards these risks is generally as follows: set the level of the test $1-\alpha$ (generally 0.05), deduct the threshold c equivalent to this level and calculate the power function of the test $\gamma(\mu_x - \mu_y) = 1 - \beta$, which corresponds to the probability of rejecting $\mathcal{H}_0$ if it is false according to the true value of $\mu_x - \mu_y$.