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 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 $\mathcal{H}_0$ (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 rejection region $\mathcal{W}$ (the t values for which $\mathcal{H}_0$ is rejected) from the acceptance region $\mathcal{W}^c$ (the t values for which $\mathcal{H}_0$ is conserved). 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).

$$\alpha = P(T \in \mathcal{W}|\mathcal{H}_0) = P(T > c^*|\mu_x = \mu_y)$$

- 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).

$$\beta = P(T \not\in \mathcal{W}|\mathcal{H}_1) = P(T < c^*|\mu_x > \mu_y)$$

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 $\mu_x - \mu_y$.

Another way of making this decision is the following: instead of deriving a decision threshold on T from α, a value comparable to α can be derived from T. This value, called the **p-value**, is the probability knowing the null hypothesis to be true of obtaining the data T or a more extreme value (i.e. more in favour of $\mathcal{H}_1$). The smaller the p-value, the stronger the presumption against $\mathcal{H}_0$. By convention, a p-value of less than 0.05 is considered *significant*.

Once the statistical test has been carried out, the researchers can return to their original question, discuss the results obtained and draw conclusions. They write it all up in a **scientific article** for which the format, structure and information given are partly determined by the scientific journal.

The next step in the scientific process is to make this experiment and its interpretation known to other researchers. This is done by publishing the article in a specialised journal. The **scientific journal** is composed of an editorial committee that guarantees the line of the journal, but also of a reading committee that allows the proposed article to be reread by other experts in the field (peer review), thus guaranteeing the quality and conformity of the proposed article. It is thus possible for an article to be rejected by a journal, or for the reviewers to request modifications or clarifications from the researchers. A period of back and forth of the article then takes place between the researchers and the journal(s), until a journal agrees to publish the article as is. The main quantitative indicator of a journal's visibility is called the **Impact Factor** (IF), and corresponds to the average number of citations garnered by articles published in the journal over the previous two years. Thus the impact factor of the renowned journal *Nature Neuroscience* is 20, that of the journal *Neuron* is 14 and that of *Neuropsychology Review* is 4.8.

Over time, other studies will address the same or very similar scientific question, or even replicate the original experiment. One of the great strengths of the experiments created and of the statistics carried out is that they are *replicable*, it is possible to carry out a published experiment again and in the same way. It was the epistemologist Karl Popper who introduced **replicability** as an important criterion of scientificity. It will then be possible to carry out a quantitative **meta-analysis** of all these studies, i.e. to carry out new statistical tests and measure new indicators by aggregating the data from all the studies in a systematic way. For certain questions, **consensus conferences** will sometimes be held to establish the opinion, at a given time, of experts in the field, in view of the available data.

Once the article has been published, the researchers can prepare a new experiment, and the cycle begins again. As time goes by, this process of creating experiments and accumulating evidence in the form of statistical tests and in particular p-values is supposed to enlarge and solidify the field of scientific knowledge, and this is in short what is taught in many statistics courses within a classical training in medicine, psychology, biology, economics, epidemiology and many other disciplines.

2 The statistical crisis

Unfortunately, this mode of production of scientific knowledge, which we can unhesitatingly call a **paradigm** (in the sense of Thomas Kuhn), is fraught with problems, which have begun to be studied extensively since 2012 under the term **replication crisis**. The notion of crisis here does not refer to a new phenomenon, but rather to the awareness, media coverage and study of these problems in experimental science. In this section, we will briefly discuss the growing awareness of this crisis among scientists before presenting in detail its causes and consequences.

At the time of writing, the notion of a replication crisis is well known to researchers and part of the general public, and an entire body of research (often called **Metascience** or **Research on Research**) is studying these issues. It is difficult to establish the beginning of this crisis, as the scientific articles warning against misunderstandings and misuses of p-value are so old and would even take us back to Ronald Fisher's first works on this subject. However, the most famous recent article dealing with this issue is certainly John P.A. Ioannidis' article published in 2005 entitled *Why Most Published Research Findings Are False*. This bombshell has now been consulted more than 3 million times and constitutes a reference in the analysis of reproducibility. Then several articles in medicine and in particular in cancerology were published between 2007 and 2015 on the assessment of the reproducibility of findings. In psychology, it was Brian Nosek and his *Open Science Collaboration* project created in 2015 that made clear the low reproducibility of his discipline and called for reforms. Finally, a famous Nature survey in 2016 among 1500 scientists stated that 70% of them have already failed to replicate a colleague's experiment and 50% of them have already failed to replicate their own experiment.