"What's a p-value?": On Social Science and Statistics

Zachary del Rosario

December 11, 2017

Statistical machinery is critically important to social science. However, many engineering curricula shortchange their students by skipping these important topics. This lecture and handout are meant to address this problem. Here I motivate and elucidate some of the techniques for drawing conclusions from data, some failures of these methods, and provide recommendations for informed reading of literature.

An Analogy

I like playing a game called Dungeons and Dragons.[1] It's a game that means different things to different people, but to me, it's an exercise in *collaborative storytelling*. The players assume the roles of different characters with their own motivations and proclivities, and work together to determine how these individuals interact.

[1] Mike Mearls. *Player's Handbook*. Wizards of the Coast, 2014

A central part of the game is the addition of *randomness*; a player decides upon the action their character would like to take, and then rolls a twenty-sided die to determine whether or not that action succeeds – the higher the number, the more likely is success. Players could decide to attempt actions which are nearly impossible, in which case the target number for success is quite high. However, there is a rule for an automatic success if one rolls the maximum value; if one scores a *natural twenty*.

Coincidentally, the probability of rolling a natural twenty is $Pr = 0.05$, which is precisely the standard significance threshold for publication.

This is a fun bit of game design, which gently encourages players to occasionally try outlandish things. The possibility of trying many different crazy schemes and having one of them succeed leads to great storytelling.

However, this same approach leads to very bad science. [2]

[2] We would call the specific practice of testing many different possibilities and extracting promising ones *multiple hypothesis testing*. This is somewhat similar to rolling a twenty-sided die many times, and only paying attention to the natural twenties. As one might expect, performing science like a tabletop game leads to some unfortunate pathologies.

Motivation

In social science, randomness is omnipresent. Unlike the clean, simple, repeatable behavior of the natural world, human beings are messy, complicated, and variable in their responses. Statistical techniques provide the means to extract signal in the presence of noise, but the use of these techniques, and the interpretation of their results, requires great care. I'm writing these notes to give a very brief introduction to drawing conclusions from data using statistics.

Engineers are my intended audience, as many engineering curric-

ula skip over statistics entirely, despite the relevance of the subject to our personal and professional lives. My reasons for presenting this tutorial are the following:

Statistics and engineering practice

Randomness often arises in engineering applications, and statistics is vital to handling it. *Industrial statistics* is applied especially to manufacturing, and is used both to measure outcomes of manufacturing processes for informed decision making, and proactively to design and conrol processes. The basic concepts from statistics are needed to understand these industrial applications.

Statistics and social science

Statistics is vitally important to understand modern issues in social science, most pressingly the *reproducibity crisis*.[3] This crisis undermines many long-held conclusions from social science, and calls into question what we really know about the human experience. However, nothing about the reproducibility crisis makes sense, except in the context of statistics.

Let's start with a specific example.

[3] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015

Case Study: Power Pose

In 2010, authors Carney, Cuddy, and Yap [4] published a paper claiming that "power posing" – holding an expansive physical pose (Fig. 1) – for a short time caused positive changes in an individual's personal outlook, hormone levels, and risk taking. The 2010 study considered 42 subjects, randomly split between groups that were instructed to adopt low and high power poses. Author Amy Cuddy [5] become quite famous for this work, presenting it in a 2012 TED talk.

However, in 2015 a team of Eva Ranehill et. al [6] attempted to reproduce the results of the original 2010 experiment, using a larger sample size of 200 persons. Their findings confirmed the subjective experience of participants, but failed to find a significant effect on hormone levels or risk taking. What followed was a flurry of additional replication studies, which ultimately resulted in original author Dana Carney [7] publically rescinding her belief in the physiological effects of power posing.

This example illustrates a case of a *failed replication*; the failure to find the same results as an original study. One would hope that results could be replicated, even under differing conditions. The failure to replicate a result calls into question the original study's findings, though is not by itself a conclusive refutation. For power posing, a

[4] Dana R. Carney, Amy J.C. Cuddy, and Andy J. Yap. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science*, 21(10):1363–1368, 2010

[5] Amy Cuddy. Your body language may shape who you are

[6] Eva Ranehill, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5):653–656, 2015

[7] Dana Carney. My position on "power poses". 2015

Fig. 1. The two high-power poses used in the study. Participants in the high-power-pose condition were posed in expansive positions with open limbs.

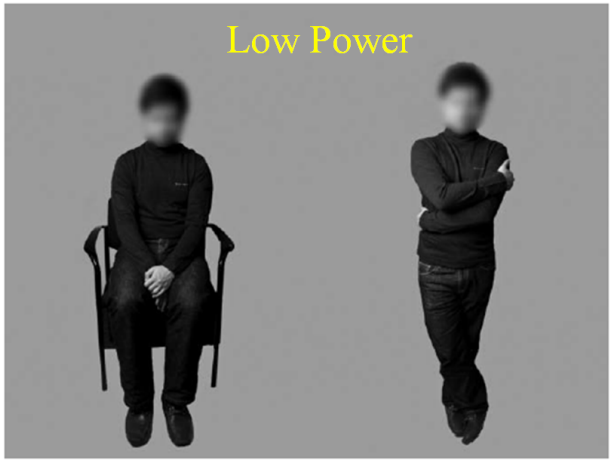


Fig. 2. The two low-power poses used in the study. Participants in the low-power-pose condition were posed in contractive positions with closed limbs.

Figure 1: From Carney, Cuddy, and Yap (2010).

collection of failed replications undermines the original study, and review of the existing literature [8] points to publication bias as a reasonable explanation for the existing successful replications.

It is important to note that, despite the evidence against power posing, it is still impossible to conclusively say that no effect exists. It is possible that the 'true' effect is negative, or may be context-dependent. At best, we can say that no strong evidence exists in favor of the physiological benefits of power posing.

However, this example also raises some natural questions for the statistical neophyte: "What is a significant effect? How is significance determined? What aspects of an experiment matter?" These are the sorts of questions we shall study below.

[8] http://datacolada.org/3

Statistical Inference

Drawing quantitative conclusions from data is an exercise in *statistical inference*. [9] For example, we may be interested in education practice, and may want to know how effective a new form of teaching might be; for example, we might believe that getting students more involved via active participation in the classroom will lead to better educational outcomes.[10] This change in practice is called a *treatment*.

In order to make an informed decision, we would want to know whether or not the treatment is effective. To determine efficacy in reality, we would perform an experiment, separating subjects into a group which receives the treatment or into a *control group* which does not,[11] measure some desired outcome with a quantitative met-

[9] Formally, statistical inference involves determining parameters of an underlying probability distribution. Choosing the *correct* distribution to describe the data is a key step in drawing reasonable conclusions.

[10] For instance: Active learning.

Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415, 2014

[11] The choice of control varies based on application. In medical testing, it would be unethical to deny treatment to those who are ill, so the standard practice in clinical trials is to compare a new treatment against the best existing standard practice.

ric, and perform data analysis to determine significance. Explicitly, we have some measureable outcome Y_i for each studied individual $i = 1, \ldots, n$, and postulate some model for the treatment, such as

$$Y_i = \mu + \beta x_i + \epsilon_i, \tag{1}$$

where μ is the average outcome without treatment, β is the amount the outcome increases (or decreases!) with the treatment, x_i is one only for those individuals i who received the treatment and zero otherwise, and ϵ_i is a noise term.[12] We'll use this model as a running example throughout this document.

In addition to our treatment, there may be some number of other variables at play; for example, the prior knowledge students bring to a class, or socioeconomic factors which help or hinder learning. We would want to control these factors in order to draw a valid inference, usually through careful *experimental design*.

Finally, it is not enough to determine that a treatment works; we would like an idea of how *practically effective* treatment is, so that we can make informed decisions about how to spend limited resources. If a particular treatment works, but causes negligible changes in student learning, we might be better off looking for a different approach, or just staying with the current best practice.

We will elucidate these concepts below.

[12] As engineers, our models tend to arise from fundamental physical principles, such as conservation laws. For statisticians, models often arise from simple compositions of probability distributions. Statisticians do not *literally* believe that effects are additive, but rather make pragmatic choices that balance utility with analytic tractability. This philosophy is embodied in the quote (attributed to George Box) "All models are wrong; some models are useful."

Sidenote: Unlike engineers, statisticians draw a distinction between the terms *variable* and *parameter*. Parameters are properties of a distribution, while variables are (often random) measured quantities. Knowing the difference may save you a headache when reading literature.

Significance testing

Testing for significance usually follows the framework of *null hypothesis significance testing* (NHST). The high-level philosophy of this framework is to define a statistical model under the assumption that the treatment has no effect. We then collect data, and check to see how compatible the data are with this hypothesis. If the data are incompatible (in a precise statistical sense), we reject the original assumption, and conclude that the treatment has some effect.[13]

In this framework, we must define a *null hypothesis*, which defines a statistical model under which the treatment confers no effect. For example, a null hypothesis of zero effect in the model defined by (1) corresponds to $\beta = 0$. The null hypothesis may also include a statement about the distribution of noise; it is common to assume the ϵ_i are drawn from a normal distribution with zero mean and (possibly unknown) variance $\epsilon_i \sim N(0, \sigma^2)$. [14]

Still working with (1), if we assume the null hypothesis, we can manipulate the outcome variable to state

$$\begin{aligned} Y_i &= \mu + \cancel{\beta} x_i + \epsilon_i, \\ &\sim N(\mu, \sigma^2). \end{aligned} \tag{2}$$

[13] This framework may seem a bit backwards at first. Proponents of NHST note that the framework is in line with a *falsificationist* perspective of science, a philosphy of scientific practice advanced by Karl Popper. This is a somewhat controversial viewpoint, as the null hypothesis is rarely advanced as a reasonable state of reality, and NHST is often used to advance an alternative hypothesis rather than to falsify another. However, NHST is certainly normative in social science, and from that perspective alone is worth studying.

Karl Popper. *The logic of scientific discovery*. Routledge, 2005

[14] It is possible in some cases to get away without assuming a particular distribution. This is called a *nonparametric* approach. Assuming a distribution in a *parametric* approach often gives a convenient mathematical form, and depending on the circumstances, can be robust to departures from the underlying assumptions.

This is not a testable statement, though! We need a statement that involves the treatment in some fashion. Let's separate the individuals that received the treatment Y_i^t from those that did not Y_i^0, and compute their averages

$$\begin{aligned} \hat{Y}^t &= \frac{1}{n_t}\sum_{i=1}^{n_t} Y_i^t, \\ \hat{Y}^c &= \frac{1}{n_c}\sum_{i=1}^{n_c} Y_i^c. \end{aligned} \tag{3}$$

Under the null hypothesis, these are still normally distributed, with $\hat{Y}^t \sim N(\mu, \sigma^2/n_t)$ and $\hat{Y}^c \sim N(\mu, \sigma^2/n_c)$.[15] Since these quantities share the same mean, we may write

$$\hat{Y}^t - \hat{Y}^c \sim N(0, \sigma^2(1/n_t + 1/n_c)). \tag{4}$$

Equation (4) provides *testable* structure. If we collect data and find that the observed difference is significantly non-zero, this will lead us to reject the null hypothesis, and conclude that the treatment has some effect.[16] Determining *how large* a difference is significant usually involves computing a *p-value*.

[15] Taking an average reduces the variance; this is part of why averaging is a useful operation.

[16] Note that in Equation (4), while we would probably like to see a positive difference, a negative value is entirely possible!

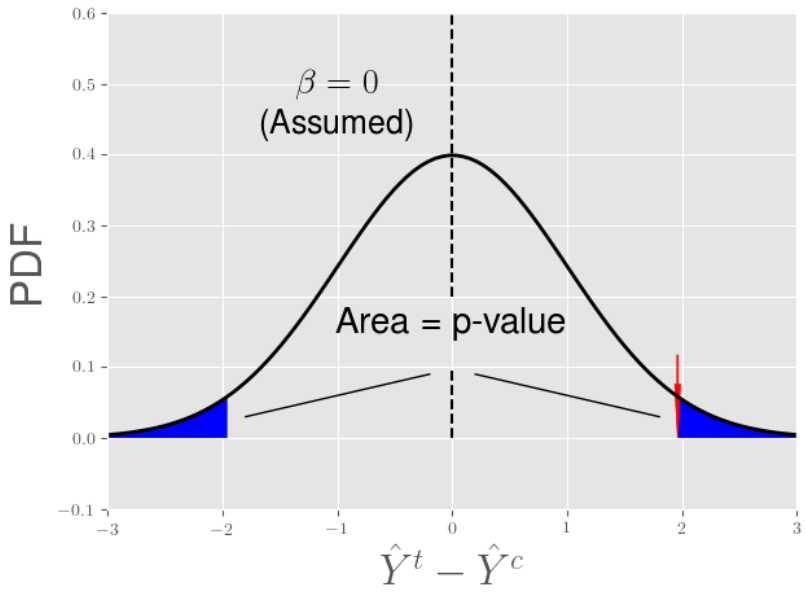


Figure 2: A visual depiction of a p-value. The probability density function (PDF) pictured is arises from our assumed null hypothesis, thus it is centered around zero ($\beta = 0$). The red arrow is the value of the treatment effect estimate we happened to observe. The p-value is defined as the probability of observing a result as or more extreme than what we observed, thus it is the area in blue. Note that we consider the area corresponding to the negative and positive value of the observed effect – this is known as a two-sided p-value.

The p-value

The *p-value* is a measure of how incompatible some data are with a given hypothesis. It is a measure of surprise, and corresponds to

the probability of observing a result at least as extreme as what was observed. Figure 2 visually depicts the p-value arising from a normal distribution.

The p-value is used to determine *statistical significance*; the smaller the p-value, the less likely we were to find the observed results under the null hypothesis. If the p-value arising from the data is smaller than some chosen threshold, then we reject the null hypothesis. A standard *significance threshold* is 5%; if $p < 0.05$, then we are said to reject the null hypothesis at the 5% level.

Figure 3: Example code to simulate our model and run a t-test.

```
### P-value computation example
# Generates synthetic data, performs two-sample t-test
import numpy as np
from scipy.stats import ttest_ind
np.random.seed(0)                      # For reproducibility
# Define the ground truth
mu   = 0.0                             # Mean response
beta = 0.5                             # Treatment effect
sig  = 1.0                             # Noise standard deviation
# Define the sampling parameters
n = 100                                # Sample size
# Draw samples
Y_t = mu + beta + np.random.normal(loc=0,scale=sig,size=n) # Treatment
Y_0 = mu        + np.random.normal(loc=0,scale=sig,size=n) # Control
# Perform t-test
res = ttest_ind(Y_t,Y_0)        # Automates a two-sided t-test
pval = res[1]                   # Extract the p-value from results
# Report
print("pval = {0:}".format(pval))
# Recorded result
pval = 0.0011808425369
```

In our running example, we would compute a p-value from (4). If σ were known exactly, we could do this easily from the normal distribution. In practice, we rarely know σ, so we use a *t-test* instead; this accounts for the additional randomness from having to estimate σ from the data. Figure 3 provides example python code that generates some fake data under our simple model, and performs a two-sample t-test on the resulting data.

Note that, despite the presence of large noise ($\sigma = 1.0$) and a (relatively) smaller effect ($\beta = 0.5$), we obtain a rather small p-value of $p \approx 10^{-3}$. This is largely due to our sample size; we happened to draw enough samples such that we could cleanly discover the

underlying effect. In general, the more samples we draw, the more probable it is that we will find the true effect (if it exists). The probability of rejecting the null hypothesis when there exists a nonzero effect is called the *power* of a test. Power is typically hard to estimate, and is a complicated function of the effect size, noise, and number of samples drawn.

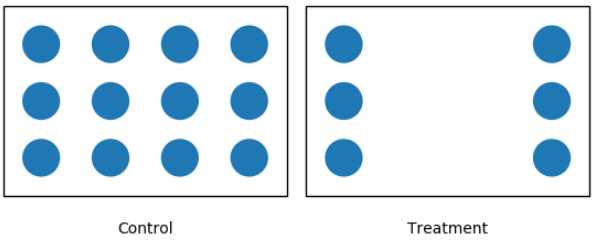
The simple procedure above assumes certain properties of the noise ϵ_i. If there is additional variability arising from other uncontrolled variables, this may interfere with correct inference. For example, if the treatment is a change in instructor approach, but only students of low socioeconomic status took the control version of the class, then we would observe only a combination of the treatment and socioeconomic effects – this is called a *confound*. Standard practices from *experimental design* are meant to address these issues.

Balance and randomization

Balance is a property of an experimental design, and is important to help ensure particular beneficial statistical properties.[17] Suppose we are studying an education treatment, but we have 12 students in the control group and only six in the treatment (Fig. 4). With this sample, we have less information about the treatment group. A *balanced* design would place an equal number of students in each group.

[17] Namely, a balanced design will have higher statistical *power* (probability of detecting a real effect), and will help prevent *heteroskedasticity* (unequal variability among groups).

Figure 4: Example of an un-balanced design. Generally, we want the treatment and control to have the same number of subjects.



Other factors are not easy to measure or control. In an educational setting, prior knowledge of a student can be difficult to measure, and socioeconomic status may be challenging to determine. In practice, we balance the variables that we are aware of, and *randomize* the remaining samples. On average, randomization takes care of any

More accurately, we want to balance the *standard errors*; here, I'm assuming the noise is the same for both the treatment and the effect. If the treatment happened to increase variability, we might actually want to put more samples in the treatment group, in order to deal with this increase.

confounding variables we did not explicitly control (Fig. 5). It is standard practice to randomize a design, and you should be highly suspicious of an experiment which does not randomize! Figure 5 illustrates the importance of randomizing an experiment through a somewhat cartoonish example.

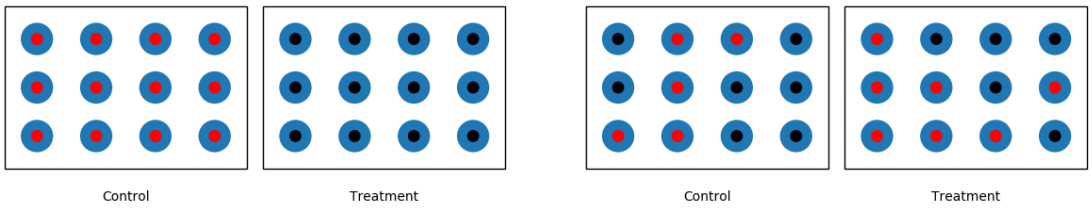


Figure 5: Suppose we have selected students from two different schools to take part in an educational study, and are bringing 12 students from each school to our lab. The schools are identical in terms of demographics, socioeconomic status, and other factors, so for simplicity's sake we place all the students from one school in the control, and the other school in the treatment group (Left image). However, though a bit of miscommunication, we are unaware that one school is a 10 minute walk away from our lab, while the other lies 5 hours away. The students from the distant school rode a bus all night to take place in this exciting study, and are extremely tired. This effect will confound the results of the experiment! Had we been aware of this issue, we could have manually balanced. However, had we randomized the design (Right image), this issue would have been handled of automatically. Randomization helps to deal with issues of which we are not aware.

Practical vs statistical significance

When making a decision, it is not enough to consider statistical significance. Even if the effect size is vanishingly small, it is still possible to find a significant p-value by drawing enough samples. Thus, it is important to consider *practical significance* when deciding on the efficacy of a treatment.

There are various measures of practical significance. One may directly consider the estimated effect size; in our example, this would be $\hat{\beta} = \hat{Y}^t - \hat{Y}^0$ estimated from the data. There is also *Cohen's d*, a normalized version of the effect size. With our model notation, d is defined by

$$d = \frac{\hat{Y}^t - \hat{Y}^0}{s}, \tag{5}$$

where s is the estimated standard deviation. Cohen's d is useful when there are no meaningful units for the measured response, and there exist standardized values for judging the size of d.

And so on

There are many more important aspects of statistical inference, and what I've covered has been at a rather high level. Rather than get too lost in the weeds, let's move on to how learning from people can fail, and talk about the *replication crisis*.

Issues with Sigificance Testing

We've seen some of the machinery of null hypothesis significance testing: How we can draw particular kinds of conclusions from data, and a bit about how to design experiments. If these techniques are well-founded, why is there a reproducibility crisis? In this section, we'll explore some plausible reasons.

P-hacking

A dirty secret of scientific literature is that papers containing statistically significant results are more likely to be published. This is known as publication bias,[18] and leads to a number of unfortunate effects. One of these effects is on research practice itself – the desire to publish incentivizes obtaining statistical significance, and adjusts researcher behavior, whether consciously or not.

P-hacking is the practice of modifying a data analysis [19] in order to obtain an acceptably low p-value. There are many ways to accomplish this. One method is called *multiple hypothesis testing*.

[18] https://en.wikipedia.org/wiki/Publication_bias

[19] https://fivethirtyeight.com/features/science-isnt-broken/#part1

Multiple hypotheses testing

Remember the Dungeons and Dragons analogy above? D&D is played by rolling dice many times – eventually, someone will roll a twenty, and succeed where they would otherwise fail. The *multiple hypothesis testing* pathology works much the same way: A hapless (or unscrupulous!) researcher decides to measure many different responses from the same subjects, and reports only those results which are statistically significant. Even if no effect exists, since we are forced to measure in the presence of noise, there is a small chance (5% for $p < 0.05$) for each hypothesis to come up significant. Falsely rejecting the null hypothesis is called making a *false discovery*, and does not result in real scientific progress.[20]

As an example, suppose a researcher seeks to determine whether there is a link between Skittles and cancer. The researcher finds no effect for Skittles in general, so decides to study the effects of particular *flavors* of Skittles.[21] Finally, after testing all possible flavors, the researcher finds a statistically significant link between Liquorice Aniseed[22] Skittles and cancer, and publishes the results, not noting that multiple hypotheses were tested to arrive at the given conclusion.

One of the big issues here is communication. If all authors noted how many hypotheses they tested to arrive at statistically significant results, we could adjust our expectations accordingly. However, one can give a false impression by selectively reporting analysis and

[20] If we test 100 independent hypotheses at the 5% level, there is a 0.6% chance that we will *not* make a false discovery. If we're trying to avoid false discoveries, that's quite bad!

[21] There are *a lot* of flavors of Skittles.

[22] Apparently a British thing.

results.

Sample sizes

One challenge with studying people is variability. Noise tends to be comparatively tame in experiments involving the physical world; physicists tend to use a 5σ criteria for discovery, which corresponds to $p \lesssim 5 \times 10^{-7}$.[23] Such a stringent significance criterion is only possible through careful control of variability, which is simply not possible in most social experiments. Drawing more samples (e.g. studying more people) helps to lessen the effects of noise, but still probably won't get us to $p < 10^{-7}$ in social experiments!

[23] Luc Demortier. P values: what they are and how to use them. Technical report, 2007

The quantity to control when designing a study is (statistical) *power*[24]: The probability of rejecting the null hypothesis, *given that the null is false*. Power is a function of the significance threshold ($p < p_c$), the effect size (β/σ in the model above), and the sample size.[25] The significance threshold is usually chosen to be $p_c = 0.05$, and an acceptably high power is considered to be $P = 0.80$. This leaves the effect size β/σ as an unknown, and the sample size n as a chooseable quantity.

[24] The power of statistics is completely different from that of physics; statistical power is dimensionless.

[25] Power also depends on a lot of other factors: The analysis used, the particular test statistic chosen, and possibly more stuff.

Determining the required sample size n for the desired power requires making a guess at the effect size! This is challenging to do well. If a researcher is overly optimistic about the effect size, they may unknowingly select a sample size which is too small, leading to a low power study. Low power studies introduce new kinds of issues.

Type S / Type M errors

Classical NHST considers two kinds of error: Type I error is the probability of *falsely rejecting* the null hypothesis,[26] while Type II error is the probability of *failing to reject* the null hypothesis when it is false.[27] Authors Andrew Gelman and John Carlin [28] introduced Type S (Sign) and Type M (Magnitude) errors, which can be rather dramatic in low-power settings (Fig. 6).

[26] There's a related, but different, concept called a *false discovery rate* (FDR). You can think of the false discovery rate as being related to multiple hypothesis testing, and there are procedures which are designed to control the FDR.

[27] Type II error is the complement of the statistical power; that is $P = 1 - Err_{II}$

[28] Andrew Gelman and John Carlin. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014

Type S error is the probability of *incorrectly estimating the sign of an effect*. When studying a treatment, we would probably like to know whether or not it helps or harms. It is possible in noisy settings that we may conclude a treatment is beneficial, *when in reality it is harmful*, or vise versa. Type S error measures how likely we are to make this mistake.

Type M error is also called the *exaggeration ratio*, and is the minimum factor by which we must over-estimate the effect size, in order to reach statistical significance. Figure 6 illustrates this issue well: In the case where noise drowns out the effect, any statistically significant results we may find will *necessarily* over-estimate the true effect.

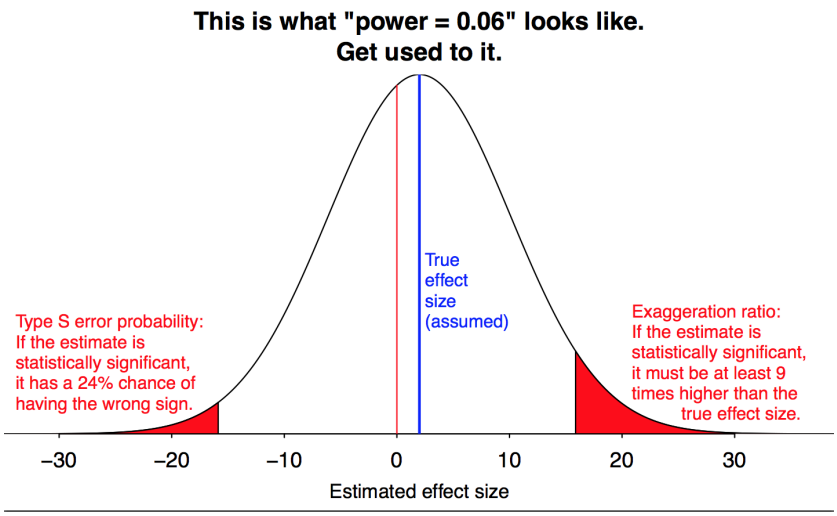


Figure 6: Graphic prepared by Andrew Gelman to illustrate the dangers of low power ($P = 0.06$) studies. In this case,

These sorts of issues chip away at our confidence in published literature. For low-powered studies, the published results may claim that a particular treatment is beneficial, when in reality it actually harms. In the event that a treatment actually is beneficial, it is quite possible the intervention is substantially less effective than the results may suggest. Gelman and Carlin actually recommend framing study design in terms of Type S and M error, rather than the traditional power design.

Recommendations

Given these issues, you may be feeling concerned. For those of us that would like to make practical choices based on studies involving humans (e.g. educators), this is particularly alarming – How can we make informed decisions if most published research findings are false?[29] I've got some practical recommendations below which will (hopefully) help.

[29] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005

Reading literature

1. Practice skeptical reading: When reading about findings regarding human beings, whether from psychology, medicine, or dietary advice, treat findings with skepticism. A great place to practice this skeptical reading is in science journalism. Science journalists are often not trained in statical techniques, and lack the time, incentives, or tools to properly evaluate the work they report on. Therefore, it's a good idea to read press releases and news articles on new fad diets or surprising psychology studies with a healthy bit of skepticism.

This practice is particularly important for *effects you are personally inclined to believe*. For example, much research on *stereotype threat* has been carried out at Stanford, and members of the Stanford community may be inclined to have a favorable outlook towards this research.[30] However, meta-reviews [31] of the literature point to signs of publication bias; given our discussion of Type M errors above, this suggests the effects of stereotype threat may be overestimated.

If that sentence makes you angry, please calm down! I'm not saying stereotype threat doesn't exist. I'm saying that *studying people is inherently difficult, and so our knowledge of the human experience is imperfect*.

There's a relevant quote from the Open Science Collaboration which I really like: "An ideological response would discount the arguments, discredit the sources, and proceed merrily along. The scientific process is not ideological. Science does not always provide comfort for what we wish to be; it confronts us with what is."[32] Our aim in doing science is *not* to confirm what we already believe. It is to discover how the universe operates, in reality. Sometimes that means being wrong.

This leads into my next recommendation:

2. Get comfortable with uncertainty: Studying humans is inherently uncertain and challenging, so it's best to get used to that fact. Expecting to be able to find $p < 10^{-7}$ in a social experiment is patently absurd, but so is pretending to have certainty about the human condition. Rather than thinking or saying "the research shows X", it's better to state "the research *suggests* X". The difference is subtle, but important. Internalizing this difference, and being able to distinguish between degrees of uncertainty, is important for working in high-noise environments.

3. Compare relevant effect sizes: This one is a very practical (though not always applicable) technique: Compare published effect estimates with similar effects. For example, if a treatment purports to add as many years to your life as quitting smoking, that's a *literally incredible* claim! Building a bit of knowledge about relevant effects is useful for orienting yourself in the literature, and is positively necessary if you wish to become well-versed in a subject.

In practice

1. Get solid statistical training: This should come as no surprise, but if you want to do social science in practice, you should get some

[30] Steven J Spencer, Claude M Steele, and Diane M Quinn. Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35(1):4–28, 1999

[31] Paulette C Flore and Jelte M Wicherts. Does stereotype threat influence performance of girls in stereotyped domains? a meta-analysis. *Journal of school psychology*, 53(1):25–44, 2015

[32] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015

serious statistical training. Take classes, and gain some serious statistical background before attempting to study human behavior.

2. Avoid low power designs: Even once you have some statistical background, avoid falling into low power experimental designs. One practice when doing power calculations is to set effect sizes aspirationally high, the idea being that, if the real effect is smaller than guessed or than some practical threshold, the study will simply

   suggest you might find the wrong sign, or grossly overestimate the effect. If you suspect you are studying a small effect, either design the study accounting for that, or seriously question whether studying the effect is worthwhile at all.

3. Preregister studies and establish pre-analysis plans: Preregistration [33] is a practice meant to help prevent p-hacking; it involves detailing and publishing an analysis plan *before the experimenter carries out the study*. One could perform some exploratory data analysis, determine what to test, detail the experimental and analysis plan, and then carry out the plan according to what was publically stated. If the authors stick to the plan, then the reader can be confident that particular decisions were made not in order to p-hack the results, but rather were made before the data were ever seen.

[33] https://www.psychologicalscience.org/observer/research-preregistration-101

Further Reading

The replication crisis (or at least some of the underlying issues) have been around for a very long time. Richard Feynman gave a commencement speech at CalTech back in the day warning about similar issues. More recently, Laura Arnold gave a TEDx talk on the replication crisis – she's a philanthropist, and is personally invested in drawing proper inferences from data to make informed decisions.

If you're serious about learning more about the nuts and bolts of statistics as it pertains to social science and the replication crisis, I have some recommended reading for you. The blogs of Andrew Gelman [34] and Leif Nelson [35] provide lucid, often less-technical perspective on current statistical issues. I've drawn from them for these notes. The website callingbullshit attacks similar issues, but with a somewhat broader perspective.

That being said, the best place to start with statistics is at the very basics; consider taking an introductory statistics course if you've never seen the material before. There's a lot of important concepts I didn't cover in this document.

[34] http://andrewgelman.com/

[35] http://datacolada.org/

Acknowledgements

Lots of folks have helped improve these notes / the talk, and I'd like to thank them here. John Arakaki provided some helpful comments on an early version of these notes. Jessica Hwang and Mike Baiocchi gave some helpful feedback on the accompanying presentation.

References

[1] Dana Carney. My position on "power poses". 2015.

[2] Dana R. Carney, Amy J.C. Cuddy, and Andy J. Yap. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science*, 21(10):1363–1368, 2010.

[3] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[4] Amy Cuddy. Your body language may shape who you are.

[5] Luc Demortier. P values: what they are and how to use them. Technical report, 2007.

[6] Paulette C Flore and Jelte M Wicherts. Does stereotype threat influence performance of girls in stereotyped domains? a meta-analysis. *Journal of school psychology*, 53(1):25–44, 2015.

[7] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415, 2014.

[8] Andrew Gelman and John Carlin. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

[9] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[10] Mike Mearls. *Player's Handbook*. Wizards of the Coast, 2014.

[11] Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

[12] Eva Ranehill, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5):653–656, 2015.

[13] Steven J Spencer, Claude M Steele, and Diane M Quinn. Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35(1):4–28, 1999.