Statistics for Neurodummies I

Oct 2020

On the meaning of the p-value in t-test and ANOVA: sampling, bias, and estimation

Amanda Sierra

Glial Cell Biology Lab, Achucarro Basque Center for Neuroscience, Spain

Department of Neuroscience, University of the Basque Country, Spain

Ikerbasque Foundation, Spain

Many Neurobiologists dislike Statistics, and many more are disheartened by the underlying Mathematics. Too few properly apply the most commonly used tests in Neurobiology, the t-test and ANOVA. And too few understand the meaning of the p-value, the probability we calculate to determine if an experiment "worked". Truth be told, most statisticians speak what seems to us biologist an obscure, arcane language that we do not know how to translate into our own research. As a Neurobiologist who loves Mathematics and Statistics, here I will try to reach my fellow researchers and put in lay terms the fundamental basis of the most common statistical analysis used in modern Neurobiology. Of course, analogies and simple explanations do not convey the full complexity of Statistics. My hope is that this text can serve as an introduction to understand basic concepts and prompt researchers to read more specialized texts. But before we dig deeper into the meaning of the p-value in t-test and ANOVA, let us examine what it means to perform an experiment.

Of experiments and marbles

Performing an experiment is akin to putting our hands into a bag of marbles and picking up a few of them to figure out how large the marbles in the bag are. We can calculate the average marble size of our handful and assume that the rest of the marbles in the bag will have the same size. In more proper terms what we do is to estimate the variable of interest (marble size) in the population (the bag) from the calculation performed in our sample (the handful) (**Figure 1**). It is obvious that this estimation will be better if we have a large handful than a small one. Similarly, the estimation will be better if the original population is rather homogeneous, i.e., if most marbles have a similar size. Therefore, both sample size and variability influence the fitness of our estimation (**Figure 1**).

In Neurobiology, each experiment is like a handful of marbles collected from an infinite, unfathomable bag. Let´s say we want to estimate the number of neurons in a particular strain of mice, or the expression of a certain gene after our astrocytes are treated with a novel drug. We

Contact information:
Amanda Sierra. Achucarro Basque Center for Neuroscience**.** Parque Científico UPV/EHU, edificio sede, planta 3. Barrio Sarriena, s/n. Leioa, 48940, Bizkaia, Spain.
amanda.sierra@achucarro.org
https://www.achucarro.org/en/research/group/laboratory-of-glial-cell-biology

Please feel free to contact for questions, corrections, and improving suggestions!

would put our hands into the infinite-sized bags of mice or cells and collect a sample of n-size. We would then analyze this sample to calculate our variable of choice (number of neurons, gene expression). However, at this point most of us do not realize that we are using this calculation in our sample to estimate the value of the variable in the original population. In addition, most of us do not realize that our estimation may not be very good if the sample size is too small, particularly if the biological variability of our parameter is high. What is even worse, we never get to analyze the infinite number of mice of a particular strain, nor the infinite number of cells treated with our drug of choice to compare with our sample and, therefore, we can never be sure how good our estimation is. Despite this intrinsic uncertainty, these three concepts, sampling, estimation, and variability are the fundamental core of most statistical tests.

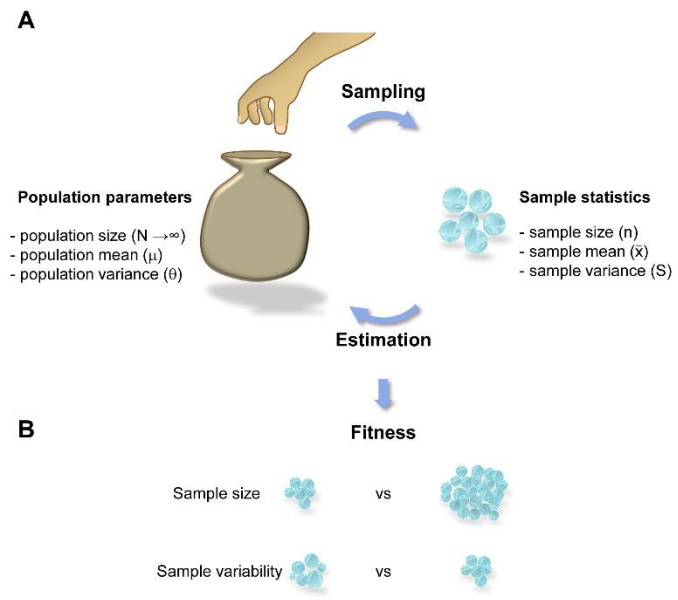

Figure 1. Population parameters are estimated from sample parameters. (A) To determine the size (population mean ± variance) of infinite marbles contained in a bag, we 1, pick up a handful of marbles (a sample); 2, calculate the marble size in the handful (sample mean ± variance); and 3, use these values to estimate the marble size in the population. **(B)** The estimation fitness depends on two main parameters: sample size and variability. Large sample sizes and small sample variability allow good estimations of population parameters.

The most commonly used test to compare our variable of interest between two groups is the Student´s t-test (**Note 1**). In the marble analog, you can imagine a blinded researcher pulling out her hand to pick up two consecutive handfuls of marbles: one handful for the control group (wild type mice, untreated cells, etc); and another handful for our treatment of interest. The question asked then is: were the two handfuls collected from the same bag or from two different bags? This is precisely what the t-test tries to answer: whether the samples collected belong to the same population, as is stated in the null hypothesis (H_0); or, in contrast, whether they belong to different populations, as is stated in the alternative hypothesis (H_1) (**Figure 2**). To determine the likelihood of the samples from each group are the same (A = B) or are different (A ≠ B), the t-test compares

Note 1. ANOVA (ANalysis Of VAriance) is used to compare more than two groups, but the principles behind its testing are similar to those of the t-test: comparing variability between groups and within groups.

Note 2. The variability of a variable is determined by the standard deviation, the standard error of the mean, or the confidence interval of the mean estimation and, in the end, is just a measurement of how spread the values of our variable are.

the variability within each group with the variability between groups (**Note 2**). When there is a clear-cut separation between the spread of each group it seems evident that they belong to different populations, that is, the groups are different.

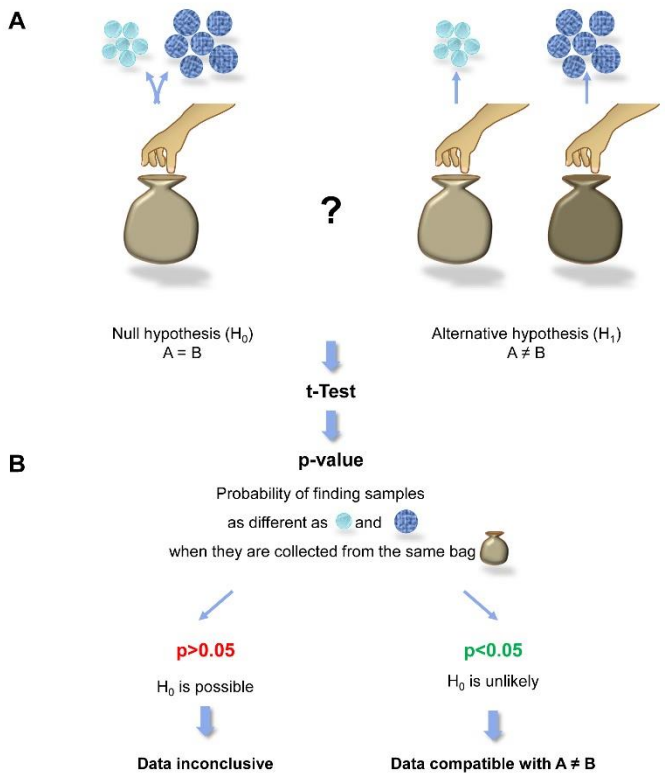

Figure 2. T-test are used to compare whether two samples belong to the same population. (A) Formally, the t-test serves to disregard the null hypothesis (H_0): whether two handfuls (samples) of marbles (light and dark blue, A and B) come from the same population (A=B), that is, whether they are picked from the same marble bag. In the alternative hypothesis (H_1), the samples come from different populations (A≠B), that is, they are picked from different marble bags (light brown and dark brown). **(B)** The t-test calculates the probability of picking up samples as different as A and B if in fact A=B. When this p-value is below the established significance level (usually set at α=0.05), it means that it is very unlikely that A=B and therefore we conclude that our data is compatible with A≠B. In contrast, when p>0.05 we cannot reach a conclusion and we can just report that no significant differences were found.

A sigh of relief: when your p value is less than 5%

The t-test transforms the comparison of the within-group and between-groups variability into a mathematical function called t distribution, which assigns a probability to each t-value (**Note 3**). This probability is what we call the p-value: the smaller the p-value, the less likely it is that our group samples come from the same population, or that our handfuls of marbles come from the same bag. And, what is the threshold to determine if the p-value is small enough or not? This threshold is called alpha level and is conventionally set at a maximum value of α=0.05 (5%). If our p-value is less than alpha=0.05 we sigh of relief and determine that our experimental manipulation (mouse strain, cell treatment) had a significant effect. In our marbles experiment, what the p-value then tells us is how likely it is that we obtained one handful of cherry-sized marbles and another handful of peach-sized marbles if they came from the same bag.

Note 3. Both the t-test and ANOVA are called "parametric tests" because they require compliance with two assumptions on the two main parameters that affect the distribution of the data (normality and equal variances), and this is discussed in more detail in *Statistics for Neurodummies II: Statistical tongue twisters: on normality and homoscedasticity and why they are important in t-test and ANOVA.*

However, it is important to recall at this time that the calculation of the p-value depends on our estimations of the variable of interest in the population based on our samples. As said above, our estimation depends on both the size of the sample and the variability of the populations. In addition, when comparing two populations, the likelihood of finding statistically significant effects also depends on the magnitude of the effect: if one bag contains cherry-sized marbles and the other one peach-sized marbles, regardless of how small our handfuls are (or how poor our sampling is), it is very likely that we can detect significant differences. Finally, another major factor is whether the marbles in each bag are picked randomly, without any bias. Imagine for instance that the first handful is picked by a child and the second one by an adult: it is evident that the child handful will likely contain smaller marbles than the adult handful. Importantly, we may not be aware of all the sources of bias in our experiments. In summary, all these factors – sample size, variability, effect size and lack of bias affect our chances of getting a significant p-value (**Figure 3**).

If we were to repeat the experiment and grab another handful of marbles, it is unlikely that we would get an identical marble size estimation and therefore we would get a different p-value to assess whether our marbles were of different size. In fact, if our two groups of marbles were identical (i.e., came from the same bag) and we repeated the experiment 100 times and collected 100 different pairs of handfuls, we would be satisfied if by chance we obtained a significant difference only 5% of the times (**Note 4**). Therefore, one should treat experiments with significant differences ($p<0.05$) with care, because we may have gotten it by chance, or due to poor sampling or some experimental bias. For this reason, many Statisticians have asked us to change our perspective of the p-value and replace “certainty” with “compatibility” (Amrhein et al., 2019): $p<0.05$ does not mean our hypothesis is true and the two populations where the samples come from are surely different. Rather, it means that the evidence we collected in our samples is compatible with the populations being different (**Figure 2**), provided that our sampling is good enough and we do not have uncontrolled biases (**Figure 3**).

Beware of your p>0.05

A mirroring problem is assuming that when $p>0.05$, the two groups compared are equal, or that the two handful of marbles necessarily come from the same bag. This interpretation of $p>0.05$ is a common problem in Neurobiology (Nieuwenhuis et al., 2011) and papers abound in which the lack of significant effect of an experimental manipulation (a drug, a genotype, etc) is immediately interpreted as this manipulation having no effect. Imagine an experimental setup in which you want to test a library of compounds to determine those that affect the expression of your gene of interest. You would be initially interested in those compounds that result in a significant change

---

Note 4. This is the reason why the p-value needs to be corrected when multiple tests are run (some classical examples are ANOVA post-hoc tests and, in general, all statistics used to analyze -omics data): the more tests, the higher the chances that some will result in a significant p-value even when in fact there are no differences (Jafari and Ansari-Pour, 2019).

in expression, but would you be so sure that the rest of the compounds did not affect your target gene? Is it possible that parameters like compound concentration or incubation time were not optimal for all compounds? Can pipetting errors or other experimental issues be completely disregarded? The answer is obviously no, they cannot.

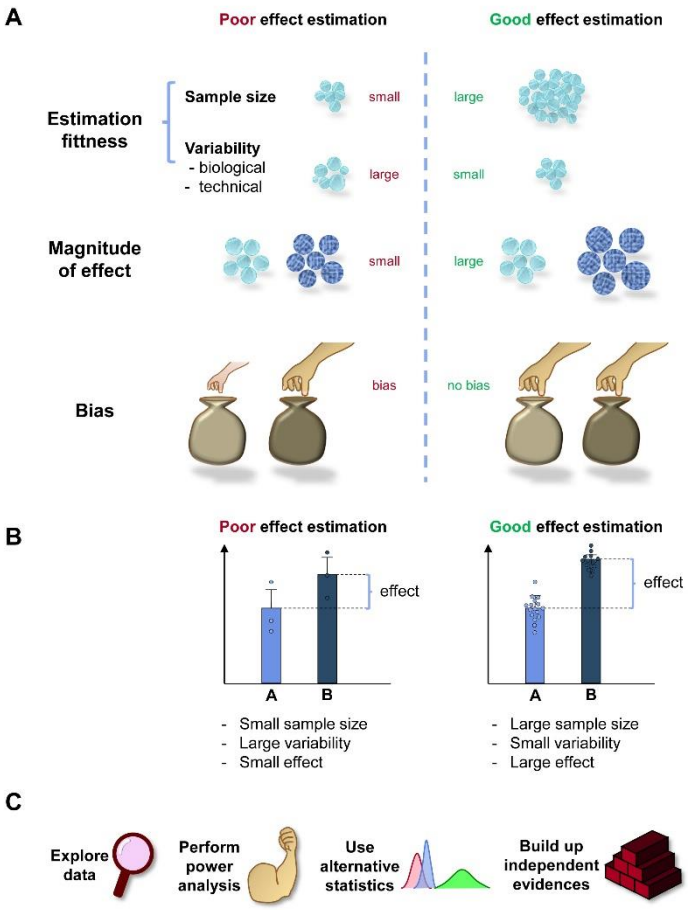

Figure 3. The reliability of t-tests depends on several parameters. **(A)** The estimation of population parameters is affected by both sample size and variability, which depends both on the biological variability of the population as well as on the accuracy of the technique used for the measurement. Large sample sizes and smaller sample variances allow good estimations of population parameters. In addition, it is easier to detect differences between populations when the magnitude of the effect is large (in particular, when it is larger than the sample variability). Finally, the presence of known or unknown biases will also affect our capacity to detect significant changes. **(B)** Examples of experimental conditions that impact our estimation of the effect, that is, the difference between the groups A and B. The bars represent the mean ± SEM (standard error of the mean) that result from the individual data points (represented as dots). On the left, small samples size, large variability and a small effect lead to a poor estimation and render the results of the t-test very unreliable (regardless of whether the p-value is significant or not). On the right, the sampling is improved and the effect is larger. In addition, the A group has the same spread than on the right panel (the extreme values are identical) but now the calculated SEM is reduced because of the larger sample size. In the B group, the spread of the data points is reduced on the left panel compared to the right panel. These three conditions, increased sampling, reduced variability and larger effect magnitude, allow us to better estimate whether A and B are, in fact, different. **C**, examples of routes to take to increase our confidence in the statistical analysis performed: explore the data to understand causes of variability and bias; perform power analysis to determine the required sample size to reach significance; complement the analysis with other types of statistics; and most of all, build up independent pieces of evidence to support our hypothesis.

If this is so evident, why do generations of Neuroscience researchers keep falling in this trap? Because it seems intuitive (although incorrect) that if low p values are interpreted as statistical support for A ≠ B (alternative hypothesis), high p-values should be interpreted as statistical support for A = B (null hypothesis). Intuition, however, would lead us astray here: due to the underlying mathematics of the t-distribution, higher p-values do not provide support for the null hypothesis (Keysers et al., 2020) and therefore a t-test can never be used to conclude that an experimental manipulation had no effect. Therefore, as initially posed by Altman and Bland already in 1995: "Absence of evidence is not evidence of absence" (Altman and Bland, 1995). Or as more recently stated by Armhein and colleagues: "We should never conclude there is 'no difference' or 'no association' just because p>0.05" (Amrhein et al., 2019).

The only valid conclusion of a t-test resulting in a p-value above the threshold for significance is that no significant effects were found (Makin and Orban de Xivry, 2019). And it should be acknowledged that this lack of significance could be either because in fact there was no effect; or because poor sampling, large variability, an effect of small magnitude, biases, or not optimized experimental conditions prevented us from reaching a conclusion. How can we then extract useful information from non-significant experiments? I am proposing here three possible routes: 1, explore your data; 2, perform power analysis; and 3, investigate novel statistical tools (**Figure 3C**).

The first step is obviously spending some quality time with your data to determine whether the problem is small sample size or large variability (including the presence of biological or technical outliers), and discover potential biases. It is also important to understand the magnitude of your effect (**Note 5**): are you discussing a 10% effect or a 200% effect? In the second case, even if no significant effects were found it may be worth pursuing further tests. To help us decide whether increasing sample size would allow us to reach statistical significance, given the expected effect magnitude observed in our preliminary experiments, we can perform power calculations (Button et al., 2013). A power above 0.8 (i.e., when 8 out of 10 experiments find statistical difference when the groups are in fact different) is usually considered sufficient. The larger the magnitude of the effect and the smaller the biological variability of the samples, the smaller the sample size needed to reach statistical significance.

Finally, researchers should be confident to explore beyond the t-test (Bernard, 2019). The information provided by t-test is limited, as they can only produce yes-or-no responses: are marbles in the two handfuls are the same (or not)? In contrast, estimation tests that measure the magnitude of the effect paint a broader picture: how much larger is the second handful of marbles? (Calin-Jageman and Cumming, 2019). Estimation statistics compare the effect size of

Note 5. For both t-test and ANOVA, an interesting parameter that allows comparing the magnitude of the effect of different variables is eta squared (η^2) (Lakens, 2013).

the experimental manipulation using confidence intervals determined by bootstrap resampling of the observations (https://www.estimationstats.com/#/). Another interesting approach is Bayesian hypothesis testing, which directly tests the plausibility of both the null and the alternate hypothesis (Keysers et al., 2020) and therefore can be used to attest both positive and negative effects.

At this point, some researchers may be pulling their hairs off and thinking that experimental Neurobiology is doomed. But do not despair, my friends. How do we live with the intrinsic uncertainty of Statistics in the experimental world? Easy enough, by performing complementary sets of experiments using different techniques that then allow us to build up independent pieces of evidence to support our initial hypothesis (**Figure 3C**).

Concluding remarks

In summary, in a t-test, low p-values can be used to determine that a certain experimental manipulation had a significant effect, but we need to be reminded that it may have happened by chance, just as it would if we were pulling handfuls of marbles out of a bag. Even more importantly, high p-values are inconclusive and cannot be used to claim that the manipulation had no effect; rather, they may result from bad sampling or poor experimental design. And, of course, never forget that setting the alpha level to 0.05 is simply arbitrary: there is nothing magic about the 5% threshold. P-values of 0.049 and 0.051 surely imply highly similar effects, even if one is just below and the other is just above the threshold. Furthermore, researchers should be aware that statistical significance does not necessarily involve biological relevance. The p-value cannot summarize the whole complexity of the biological data in our experiments and this is the reason why many experts claim to start using confidence intervals to measure the magnitude of the effect. Overall, we need to acknowledge that Statistics is just a tool that we use to agree on whether something is likely to happen or not. I hope that from visualizing marbles researchers can extrapolate the meaning of the p-value to their own experiments.

ACKNOWLEDGEMENTS

I would like to dedicate this Commentary to the memory of my dear Agora Highschool Math teacher, Eugenio Rodrigo, who taught me to love Mathematics. I am deeply grateful for enriching discussions to Eva Benito, Luis Miguel García-Segura, Carlos Matute, students of the Achucarro Introductory course on Statistics for Neurobiologists, and researchers of the Sierra lab, who inspired this article.

This work was supported by grants from the Spanish Ministry of Science and Innovation (https://www.ciencia.gob.es/) with FEDER funds to A.S. (RTI2018-099267-B-I00) and a Tatiana Foundation project (P-048-FTPGB 2018).

REFERENCES

Altman DM, Bland JM (1995) Absence of evidence is not evidence of absence. BMJ.

Amrhein V, Greenland S, McShane B (2019) Scientists rise up against statistical significance. Nature 567:305-307.

Bernard C (2019) Changing the Way We Report, Interpret, and Discuss Our Results to Rebuild Trust in Our Research. eNeuro 6.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafo MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365-376.

Calin-Jageman RJ, Cumming G (2019) Estimation for Better Inference in Neuroscience. eNeuro 6.

Jafari M, Ansari-Pour N (2019) Why, When and How to Adjust Your P Values? Cell J 20:604-607.

Keysers C, Gazzola V, Wagenmakers EJ (2020) Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. Nat Neurosci 23:788-799.

Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 4:863.

Makin TR, Orban de Xivry JJ (2019) Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. Elife 8.

Nieuwenhuis S, Forstmann BU, Wagenmakers EJ (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. Nat Neurosci 14:1105-1107.