# Statistics for Neurodummies II

Oct 2020

## Statistical tongue twisters: on normality and homoscedasticity and why they are important in t-test and ANOVA

**Amanda Sierra**

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

Department of Neuroscience, University of the Basque Country, Spain

Ikerbasque Foundation, Spain

Once upon a time, when you were in graduate school, you were told that before applying statistic parametric tests, such as t-test and ANOVA, you should make sure that your data complied with normality and homoscedasticity. Surely, you were told. Then, why does this basic statistical knowledge seems evaporated from many current Neurobiology papers? As a journal reviewer, I find that, most often than not, research paper authors skip assessing normality and homoscedasticity and go straight into applying t-tests and ANOVA. In this essay, I will explain in simple terms the meaning of these two concepts, normality and homoscedasticity, and their impact on the power of parametric tests. 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. Finally, I will discuss alternatives when the data does not comply with normality and homoscedasticity, such as studying outliers, using non-parametric tests, and transforming the data.

### Homoscedasticity

Let us start with the battle horse: the tongue-twisting word, homoscedasticity. Homoscedasticity is a Greek word that means equal dispersion and homoscedastic groups are those with equal (or homogeneous) variances. Variability is more layman term to refer to this property or experimental samples: how far away are the individual data points from the mean or median of their group (**Figure 1A**). Visually, sample variability is related to error bar length in bar graphs and whisker length in boxplots. In addition, there are statistical tests that allow us to determine whether the variances between groups are homogeneous, such as the Levene test and modified versions (Kim and Cribbie, 2018). Variability has many sources: the intrinsic variability of the biological parameter measured; the variability imposed by the quantification technique and device; and the existence of unknown biases (**Figure 1B**). For instance, if you perform an experiment to measure the individual height of researchers in your work place you will find that it depends on three factors:

**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!**

1, the biological variability of individuals, based on their age, gender, race, medical history, etc. 2, the precision of your height calculator. And 3, whether all measurements were taken at the same time of the day (no bias); or some were taken early in the morning, fresh from a night of rest, and others late in the afternoon, after a day of crouching in front of the bench (bias due to diurnal height reduction). Another source of bias could be the expertise of the experimenter using the height calculator. And of course, there can be other less suspicious bias not so easy to spot. Variability is thus a parameter just as important as the mean or median to define our sample, as it has a strong impact in our capacity to determine whether our variable of interest is different between our experimental groups.

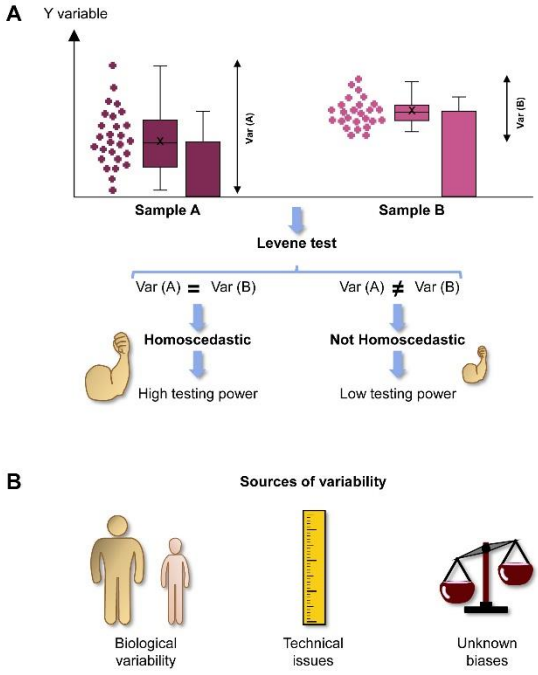



**Figure 1. Effect of homoscedasticity on parametric tests. A,** Distributions of two samples (A and B) shown by their data points (left), whisker plots with median (middle line) and mean (X)(center), and bar plot with mean plus standard deviation of the mean (right). The spread or variability (var) of each sample is indicated as a two-arrow line. The Levene test is used to assess homoscedasticity between samples. If samples are homoscedastic, parametric tests with high power such as the t-test and ANOVA can be used. If samples are not homocedastic, the power of parametric tests is considerably reduced. **B**, The three main sources of variance are the biological variability of the sample, the device and technique used to measure our variable of interest, and unknown biases.

But, why is data homoscedasticity important when running a t-test or an ANOVA? Because variability analysis lies at the core of these parametric tests: visually, when the variability between groups is larger than the variability within each group (sample variance), the groups are found to be significantly different (for a full explanation see *Statistics for Neurodummies I. On the meaning of the p-value in t-test and ANOVA: sampling, bias, and estimation*). In mathematical terms, this comparison is performed by fitting the individual data points from all groups combined to a linear regression model, a fitting that imposes the need for equal variances and normality (**Note 1**). A more practical way to see why homoscedasticity is important in t-tests and ANOVA is by performing data simulations. When p-values are calculated from sets of simulated data it becomes evident that lack of homoscedasticity reduces the chances of obtaining significant results (Yang

**Note 1.** The "error" in this model is distance of each data point to the regression line (which is related to the groups´ variance), and needs to be normally distributed and have a variance of $\sigma 2$ (http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis).

et al., 2019). Therefore, if the measurement of our variable of interest shows different variances between the experimental groups, due to either biological variability or technical issues, our statistical analysis will have reduced power and an increased risk of false negatives, concluding that there were no significant differences when in fact there may be (**Figure 1A**).

## Normality

Intuitively, we all can visually identify a normal or Gaussian distribution when we see a bell-shaped curve.

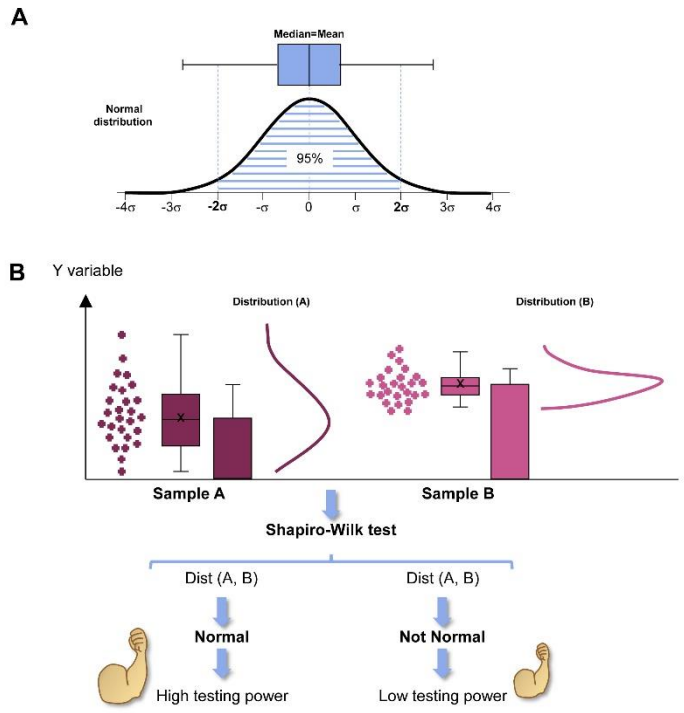



**Figure 2. Effect of normality on parametric tests.** **A**, The normal distribution is a symmetric around the mean (which coincides with the median) and 95% of the data points are located in the interval of plus/minus two standard deviations (σ). **B**, distributions of the same two variables A and B shown in Figure 1, to assess their similarity to the normal distribution shown in A. B, The Shapiro-Wilk test is used to assess normality. If the data follows a normal distribution, parametric tests with high power such as the t-test and ANOVA can be used**.** If the data is not normal, the power of parametric tests is decreased but to a lesser extent than the lack of homoscedasticity.

This curve is a histogram, in which the x-axis represents the values of the variable and the y-axis represents the frequency with which each value appears. In the normal distribution, the data points are symmetrically distributed around the mean value, i.e., the mean and the median coincide. The width of the curve relates to the standard deviation (Std Dev) of the mean, so that 95% of the data points are located in the interval (mean ± 2 Std Dev) (**Figure 2A**) (https://www.statisticshowto.com/probability-and-statistics/normal-distributions/). This distribution is also easy to identify when looking at whisker plots, as the median is placed on the middle of the whisker box and the whiskers are symmetrical and of the same length as the box (**Figure 2A, B**) (https://www.simplypsychology.org/boxplots.html). In contrast, it is not obvious to determine

whether a distribution is normal by looking at bar graphs. And of course, most statistical packages contain tests to assess normality, such as the Shapiro-Wilk test (Ghasemi and Zahediasl, 2012).

As in the case of homoscedasticity, lack of normality may lead to reduced power to detect significant differences (http://goodsciencebadscience.nl/?p=492). However, in most cases, ANOVA and t-test are robust to departures from normality, as shown by data simulations (Yang et al., 2019) (**Figure 2B**). In fact, testing for normality can be disregarded when working with large sample sizes (n> 50) (Yang et al., 2019), because large samples tend to behave like a normal distribution (Kim and Park, 2019). It is important to realize, however, that in many experimental conditions in Neurobiology, our sample size is simply too small to determine whether it follows a normal distribution (**Note 2**).

### A word on outliers

As we have seen above, t-tests and ANOVA are quite resistant to lack of normality, whereas lack of homoscedasticity may have a strong impact on the power of our data analysis. Therefore, the first step after assessing that our sample groups are not homoscedastic is to perform a detailed analysis of the causes behind such different data variance. Is it all due to biological variability? Are there any technical issues? Are there any suspected biases? Is the data sparsely but homogeneously distributed or are there any clear outliers? Outliers can be defined as data located outside the 95 or 99% confidence interval or the mean (**Figure 3**), and to help in their identification statistical tests such as ROUT or Grubbs can be used (https://www.graphpad.com/guides/prism/7/statistics/STAT_How_to_Removing_outliers.htm). But, once identified, what should we do with outliers?

This is a critical ethical problem in research. My rule of thumb is that unless there is a sound biological or technical reason to do so, outliers should not be removed, as they represent the biological variability of the sample. For example, imagine that you performed an experiment to test the effect of enriched environment on the dendritic arbor length of your neurons of interest. One of the enriched environment mice showed very short dendritic arbor length compared to the rest of the animals in its group, and it turned out that this particular mice was sickly beaten by the other mice in the cage. In this case, it seems likely that the stress conditions suffered by that mouse overpowered the effects of the enriched environment and thus it is probably safe to consider the mouse an outlier and remove it from the study. Nonetheless, when in doubt, it is best to increase the sample size to have a better representation of the biological variability. Importantly, under no circumstances can outliers be removed with the purpose of reaching statistical difference between experimental groups, as it introduces sample bias and violates all principles of research ethics.

---

**Note 2.** A small sample size leads to low power in the normality test and as a result, the alternative hypothesis of non-normality is rejected, thus leading us to wrongly believe that our distribution is normal (Ghasemi and Zahediasl, 2012).

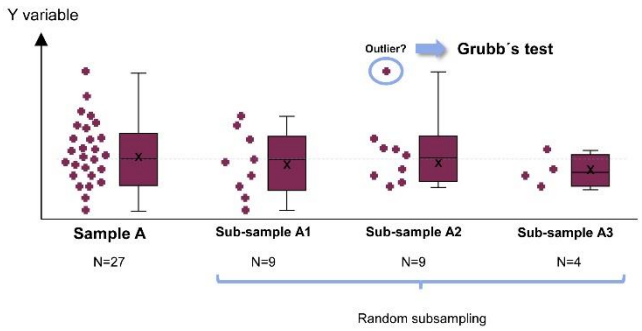
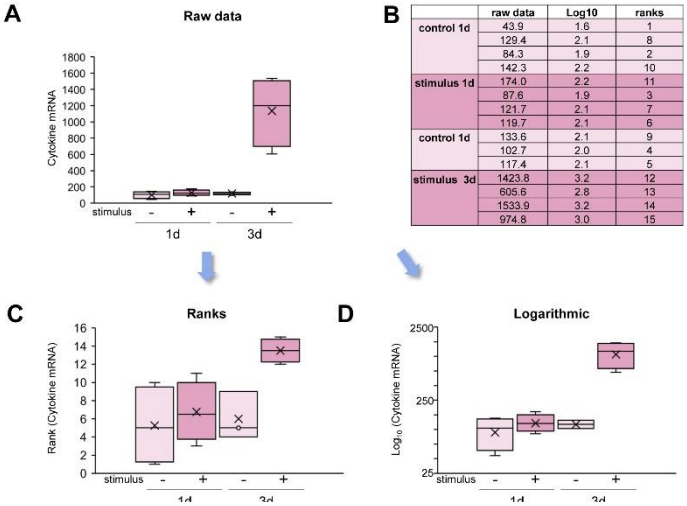



**Figure 3. Effect of outliers and sample size on statistical analysis.** The data points (left) and whisker plot (right) of sample A (N=27 data points) and three subsamples (A1, A2: N=9; A3: N=4) obtained by randomly removing points from A. A1 is a good subsample of A, as it has a similar median (blue dotted line) although it has a slightly smaller mean (X). A2 is a also a good subsample of A, as it also has similar median and mean; however, in A2 the upper most extreme data point is far from the rest of the points and would be a suspected outlier (and could be assessed using Grubb´s test) whereas in fact is part of the original A sample. A3 in contrast is a very small sample and as it not contains a good representation of all data points in A, its median and mean are considerably smaller.

Nonetheless, in some experiments, homoscedasticity is never reached, no matter how large the group size is. In my own experience, one example of this type of situation is the mRNA expression of cytokines by microglia under an inflammatory stimulus. In control conditions, the expression of tumor necrosis factor alpha (TNF$\alpha$) is almost undetectable, whereas after inflammatory challenge it rises several orders of magnitude and with large variability between samples. In fact, the effect is so large that I usually present the data in a logarithmic scale (**Figure 4A**). When this is the case, we have two alternatives to assess whether our experimental manipulation had a significant effect: use non-parametric tests or transform the data (**Figure 4B-D**) (**Note 3**).

### Using non-parametric tests or transforming data?

The most conventional route when data is not homoscedastic and parametric tests should not be used is to perform a non-parametric test. Examples are the Mann-Whitney-Wilcoxon U test or rank test, in substitution for a t-test when comparing two samples; or the Kruskal-Wallis test, in substitution for an ANOVA when comparing more than two samples (Grech and Calleja, 2018). However, before using these tests, it is important to understand how they operate: all data points from the groups combined are ranked from smallest to largest, and their original value is replaced by their place in the rank. The data points are then split back into their original groups and the test is performed not on their original values but on their ranks. A major drawback of tests based on ranks is that the effect size is lost, that is, that the relative spread between data points is not maintained when ranks are considered (**Figure 4B,C**). As a result, non-parametric tests have very little statistical power (Grech and Calleja, 2018). Since using non-parametric tests reduces your chances of detecting statistical significant differences because it transforms your data based on ranks, it is a good idea to consider alternative mathematical functions to transform your data.




| | raw data | Log10 | ranks |
|---|---|---|---|
| control 1d | 43.9 | 1.6 | 1 |
| | 129.4 | 2.1 | 8 |
| | 84.3 | 1.9 | 2 |
| | 142.3 | 2.2 | 10 |
| stimulus 1d | 174.0 | 2.2 | 11 |
| | 87.6 | 1.9 | 3 |
| | 121.7 | 2.1 | 7 |
| | 119.7 | 2.1 | 6 |
| control 1d | 133.6 | 2.1 | 9 |
| | 102.7 | 2.0 | 4 |
| | 117.4 | 2.1 | 5 |
| stimulus 3d | 1423.8 | 3.2 | 12 |
| | 605.6 | 2.8 | 13 |
| | 1533.9 | 3.2 | 14 |
| | 974.8 | 3.0 | 15 |

**Figure 4. Effect of data transformation on statistical analysis**. **A**, Experimental example of TNFa cytokine mRNA in four experimental groups: in absence and presence of an inflammatory stimulus, in two time points (1 and 3 days). **B**, Table showing raw TNFα data as well as logarithmic (base 10) and rank transformation. **C**, Graph of ranks-transformed data. D, Graph of Log10-transformed data.

Widely used transformations include functions such as sigmoid, logarithmic, and logarithmic plus one if the samples have zeroes (https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16). It is of paramount importance checking whether the transformation actually improved the data distribution. If they do and you end applying a test or an ANOVA it is important to realize that you may find that you have significant differences – but on the transformed data instead that on the raw data. For instance, in the cytokine example above, I may find that under inflammatory conditions microglia significantly increased the logarithmic expression of TNFα mRNA. In this case, this result can be graphically interpreted as if the statistical test were applied on the data showed in the logarithmic graph (**Figure 4D**). However, interpreting statistical data from transformed variables may not be always evident; more importantly, it may not even be relevant for the non-transformed data (Feng et al., 2014) (**Note 3**).

### In summary, be honest

I hope to have convinced you of the importance of assessing normality and homoscedasticity. Not only because not complying with these assumptions leads to reduced power to detect significant differences; but also because they are intrinsic properties of the samples. The more you know your data, the better you will be able to understand the effect of your experimental conditions. A summary of data analysis and decisions to be made is shown in **Figure 5**. As a take-home message, I would recommend to pamper the Statistics methods section in your

**Note 3.** As an alternative to data transformation, novel tests have been developed that do not depend on the data distribution, such as generalized estimating equations (Feng et al., 2014)

research paper: clearly state whether normality and homoscedasticity were tested; how were outliers identified were removed and whether they were removed (and why); whether data was transformed; which statistics were used; and which criteria was used to determine statistical significance. In summary, be honest and let your reader decide for him/herself whether the data supports the conclusions.

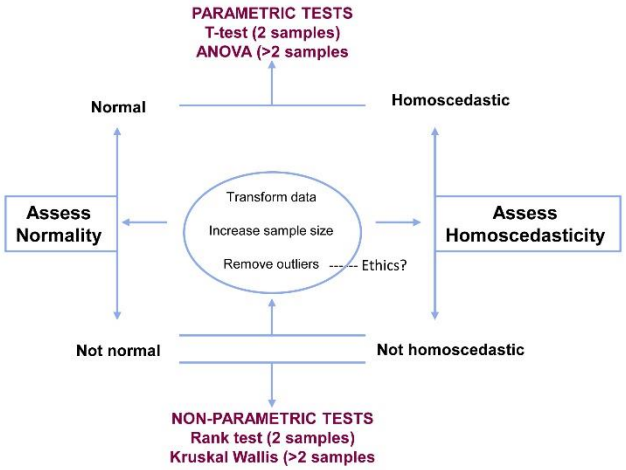



**Figure 5. Normality and homoscedasticity decision tree.**

**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**

Feng C, Wang H, Lu N, Chen T, He H, Lu Y, Tu XM (2014) Log-transformation and its implications for data analysis. Shanghai Arch Psychiatry 26:105-109.

Ghasemi A, Zahediasl S (2012) Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab 10:486-489.

Grech V, Calleja N (2018) WASP (Write a Scientific Paper): Parametric vs. non-parametric tests. Early Hum Dev 123:48-49.

Kim TK, Park JH (2019) More about the basic assumptions of t-test: normality and sample size. Korean J Anesthesiol 72:331-335.

Kim YJ, Cribbie RA (2018) ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. Br J Math Stat Psychol 71:1-12.

Yang K, Tu J, Chen T (2019) Homoscedasticity: an overlooked critical assumption for linear regression. Gen Psychiatr 32:e100148.