# A/B Testing and Statistical Power Analysis

S. A. Owerre[*]

(Dated: August 17, 2021)

In this report, I jot down my thoughts on A/B testing and statistical power analysis focusing mainly on the statistics behind them. In addition, I explore parametric t-tests and non-parametric tests. This report is compiled based on my research in this subject by reading different textbooks, papers, and blogs.

## I. A/B TESTING

Suppose we want to know if adding a new feature $X$ in a retail company's webpage will increase revenue. To do this, we randomly split users (i.e., customers) into two groups. We expose one group to the existing webpage without the new feature $X$, and the other group to the existing webpage with the new feature $X$. Users assigned to the existing webpages with and without the new feature $X$ are called the treatment and control groups respectively as shown in Fig. (1). A user should consistently see the same webpage during the course of the experiment, to guarantee that the observations in the treatment and control groups are independent [1].

In practice, power analysis should be conducted first to determine the minimum sample size needed to detect a particular effect size. This will be discussed in later sections. It is also recommended to start with small number of subjects (e.g. users) and increase it over time, in case something goes wrong and you need to abort the experiment abruptly.

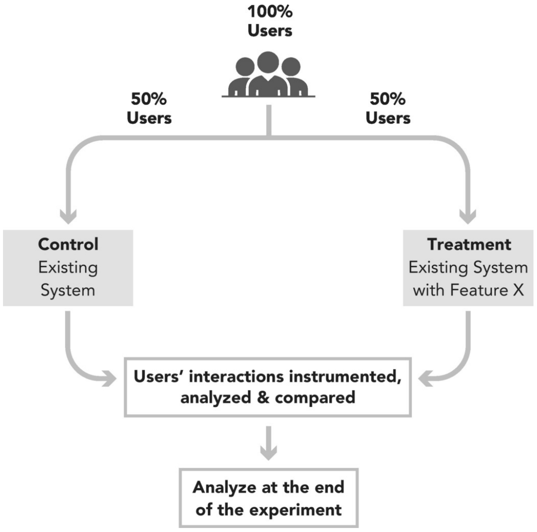

FIG. 1. A simple controlled experiment: An A/B Test. Adapted from Ref. [1].

Assuming everything goes well and we collect data for the revenue per user generated in each group (say, after about one week). In this case, the revenue per user is our overall evaluation criterion or the metric of success. It is also called the target or dependent variable. Usually, we assume that the treatment and control samples are drawn independently from two normally distributed populations with equal variances (or without equal variances). This leads to a two-sample t-test which will be described in later sections.

To test the effectiveness of the new feature $X$, we can formulate the following two hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2, \tag{1}$$

where $\mu_1, \mu_2$ are the population means. The hypotheses in Eq. (1) are called a two-tailed (nondirectional) or a two-sided hypothesis testing. We call $H_0$ the null hypothesis and $H_1$ the alternative hypothesis. The null hypothesis states that there is no difference between the two population means, whereas the alternative hypothesis states that there is a difference between the two population means.

Evidently, the hypothesis makes a statement about the population parameters. The goal of a hypothesis test is to decide, based on samples drawn from the populations (in this case the treatment and control samples), which of two complementary hypotheses is true [2]. In hypothesis testing, there are two types of errors that can happen as shown in Table. (I):

TABLE I. Outcomes of hypothesis testing

| | Accept $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ true | ✓ | type I error |
| $H_1$ true | type II error | ✓ |

1. **type I error** is to reject $H_0$, when it is in fact true. This means concluding there is a significant difference between treatment and control, when there is none.
2. **type II error** is to accept $H_0$, when it is in fact false. This means concluding there is no significant difference between treatment and control, when there is one.

For any test, the probability of making type I error is denoted by $\alpha$, and it is identical to the significance level, usually set at 0.05, implying that it is acceptable

[*] alaowerre@gmail.com

to have a 5% probability of committing type I error. The probability of making type II error is denoted by $\beta$ [3, 4]. Mathematically,

$$\mathcal{P}(\text{rejecting } H_0 | H_0 \text{ true}) = \alpha, \tag{2}$$

$$\mathcal{P}(\text{accepting } H_0 | H_1 \text{ true}) = \beta. \tag{3}$$

## II. P-VALUE

In hypothesis testing, one of the crucial questions one usually asks is: does our data provide enough evidence for us to reject the null hypothesis $H_0$ [5]? This question is usually answered with the help of the p-value. The p-value is a measure of the evidence against the null hypothesis $H_0$: the smaller the p-value, the stronger the evidence against $H_0$. A test result that leads to the correct rejection of $H_0$ is said to be statistically significant.

Formally, the p-value is the probability of observing a test statistic at least as extreme as what was observed, assuming that the null hypothesis is true.

Suppose we have a fixed significance level $\alpha$ (i.e., the probability of type I error), if

$$\text{p-value} \leq \alpha, \tag{4}$$

then we reject $H_0$ at level $\alpha$. This means that the p-value is the smallest $\alpha$ at which we reject $H_0$. It is important to note that a large p-value $> \alpha$ is not a strong evidence in favour of $H_0$. A large p-value can occur for two reasons: $(i)$ $H_0$ is true or $(ii)$ $H_0$ is false, but the test has low power [3]. In this case (i.e., p-value $> \alpha$), it is safe to say that there is not enough data to reach any conclusion.

## III. PARAMETRIC T-TESTS

The t-test is *parametric* in that it makes assumptions on the population distributions (or the parameters of the populations) from which the samples were drawn. The meaningfulness of the result of a t-test depends on the validity of the assumptions. Thus, before using a particular t-test, it is advisable to examine if the assumptions that a test is based upon is valid using the sample data.

### A. The Student's T-Test

In the development of two-sample t-test, we make the following assumptions:

1. *Normality*: The two samples are assumed to be drawn from two normally distributed populations with unknown variances or standard deviations. To test this assumption in practice, one should plot the histogram of each sample to see if it follows a normal distribution especially when the sample size is large enough. If the distribution is right skewed, a log transformation can be applied. However, performing t-test on the log transformed data can be difficult to interpret since you cannot map the result back to the untransformed data unlike in regression prediction.

2. *Equal variances*: The two populations are assumed to have equal variance, i.e., homoscedasticity. To test this assumption requires an examination of the variance of each sample. If the sample variances are roughly the same, then it is reasonable to assume that they are estimating a common population variance. In this case, we can pool the two sample variances as the estimate of the population common variance.

3. *Independence*: The two samples are assumed to be drawn independently. This assumption is tested by the experimenter when designing the experiment. The experimenter should make sure that during the course of the experiment each subject (randomization unit) is assigned to either the treatment or the control to avoid repeated measurement from the same subject. In other words, a subject cannot generate an observation in both the treatment and the control. This guarantees that there is no relationship between the observations in one group as compared to the other. When this assumption is violated one should consider the paired t-test.

Formally, let $x_1^i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $i = 1, 2, \cdots, n_1$; and $x_2^i \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $i = 1, 2, \cdots, n_2$, where $x_1^i$ and $x_2^i$ are independent, and $\sigma_1^2 = \sigma_2^2 = \sigma^2$ with $\sigma$ unknown. We can think of the samples $x_1$'s and $x_2$'s as observations from the treatment and control groups respectively. The test statistic is defined as

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}, \tag{5}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the two sample means (i.e., the estimates of $\mu_1$ and $\mu_2$), and $SE(\bar{x}_1 - \bar{x}_2)$ is the standard error of $\bar{x}_1 - \bar{x}_2$, given by

$$\begin{aligned} SE(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}}, \\ &= s_{\text{pooled}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \end{aligned} \tag{6}$$

where $s^2_{\text{pooled}}$ is the sample pooled variance (i.e., the estimate of the common population variance $\sigma^2$) given by

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \tag{7}$$

where $n_1, n_2$ are the two sample sizes, and $s_1^2, s_2^2$ are the two sample variances, given by

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_j - \bar{x}_i)^2, \quad i = 1, 2. \tag{8}$$

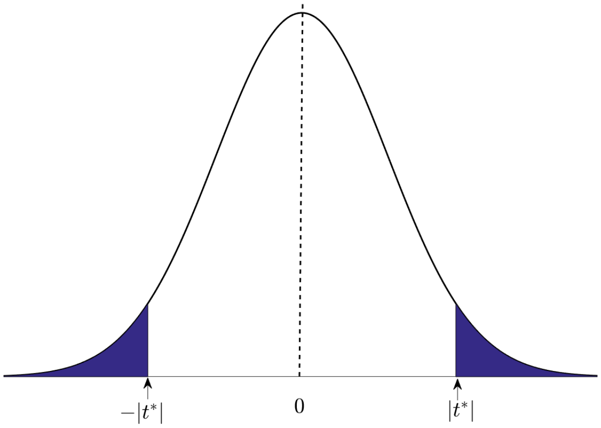

FIG. 2. A plot of the t-distribution (under $H_0$) for sample size $\nu = 50$. The area under the shaded regions is the p-value.

The test statistic $t^*$ measures the number of standard deviations that $\bar{x}_1 - \bar{x}_2$ is away from the null hypothesis $\mu_1 - \mu_2 = 0$. Assuming that the null hypothesis is true, the probability of observing any value equal to $|t^*|$ or large is the p-value. To compute this p-value, we need to know the sampling distribution from which $t^*$ was drawn (i.e., the distribution of $t^*$ under $H_0$). The distribution is the Student's t-distribution given by [6, 7]

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{9}$$

where $\nu$ is the number of degrees of freedom.

The test-statistic $t^*$ is drawn from the t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom, and for large samples, i.e., $\nu \to \infty$, $f_\nu(t) \to \mathcal{N}(0,1)$. The p-value is the area of the two shaded regions in Fig. (2), given by

$$\begin{aligned}\text{p-value} &= \mathcal{P}\Big(|T| \geq |t^*| \Big| T = f_\nu(t)\Big) \\ &= \mathcal{P}\Big(T \leq -|t^*| \text{ or } T \geq |t^*| \Big| T = f_\nu(t)\Big) \\ &= 2\mathcal{P}\Big(T \leq -|t^*| \Big| T = f_\nu(t)\Big) \\ &= 2\Phi_\nu\big(-|t^*|\big), \end{aligned} \tag{10}$$

where $\Phi_\nu(x)$ is the cumulative distribution function (CDF) of the t-distribution.

The $100(1-\alpha)\%$ confidence interval for $\delta = \mu_1 - \mu_2$ is given by

$$\hat{\delta} - t^*_{\nu,\alpha/2} SE(\hat{\delta}) \leq \delta \leq \hat{\delta} + t^*_{\nu,\alpha/2} SE(\hat{\delta}) \tag{11}$$

where $\hat{\delta} = \bar{x}_1 - \bar{x}_2$ and $t^*_{\nu,\alpha/2}$ is a critical value of $t^*$ with $\nu = n_1 + n_2 - 2$ degrees of freedom.

If the confidence interval does not contain zero, it means that the null hypothesis $\delta = 0$ is false and we reject it. Therefore, $\hat{\delta}$ is statistically significant at level $\alpha$ if the $100(1-\alpha)\%$ confidence interval does not contain zero, which is equivalent to p-value $< \alpha$. Note that for large sample sizes ($\nu \to \infty$) and $\alpha = 0.05$, the critical value $t^*_{\nu\to\infty,0.025} \approx 1.96$, so $\delta$ is about 2-standard deviation away from $\hat{\delta}$ on both sides.

### B. The Welch's T-Test

The assumption of equal variances in the Student's t-test is not always tenable. In this case, the distribution of the test statistic is no longer a t-distribution. This is called the Behrens-Fisher problem [8]. The Welch approximation for this problem is called the Welch's t-test [9]. Suppose $\sigma_1^2 \neq \sigma_2^2$ and both population variances are unknown, the test-statistic is given by

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}, \tag{12}$$

where $SE(\bar{x}_1 - \bar{x}_2)$ is now given by

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{13}$$

where $s_1^2, s_2^2$ are the two sample variances in Eq. (8), which are the estimates of the two unequal population variances. The exact distribution of $t'$ is very complex. However, it can be approximated to a t-distribution with $\nu$ degrees of freedom, given by

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}. \tag{14}$$

Note that $\nu$ is generally non-integer unlike the Student's degrees of freedom.

### C. Summary of Procedure for T-Test

In this section, I summarize the steps that should be taken when conducting a t-test. Given an experimental data, the following steps can be performed:

1. Verify the assumptions that the test is based upon.
2. State the two complementary hypotheses.
3. Pick a significance level, $\alpha$.
4. Compute the test statistic.
5. Find the degrees of freedom, $\nu$.
6. Compute the p-value and compare it to $\alpha$.
7. State the conclusion of the test results.

*Programming package*: In principle, the statistical analyses of t-test can be coded manually. However, it is advisable to use programming softwares especially for large data. In this regard, the Python SciPy package scipy.stats.ttest_ind can be used to compute test statistic and its p-value for the Student's t-test, and for Welch's t-test set the equal_var parameter to False.

## IV. NON-PARAMETRIC TESTS

In the *parametric* t-test, assumptions are made in advance about the population distributions. However, in some instances these assumptions may not be valid. It is often difficult to have access to normally distributed samples. When the data fail the normality assumption in which the two-sample t-tests are based upon, a *non-parametric* statistical test is recommended. In this case, the population distributions from which the samples were drawn are not assumed in advance. Therefore, *non-parametric* tests are "distribution-free" [4].

### A. The Mann-Whitney U Test

The Mann-Whitney U test is a *non-parametric* test that can be used in this scenario to compare two groups, which are assumed to be independent [4, 10–12]. In this case, the null hypothesis $H_0$ states that the two independent samples come from the same population, whereas the alternative hypothesis $H_1$ states that the two independent samples come from different populations.

Suppose we have $n_1$ observations $x_1^i (i = 1, 2, \cdots, n_1)$ in one sample (i.e. from one population) and $n_2$ observations $x_2^i (i = 1, 2, \cdots, n_2)$ in another sample (i.e., from another population). The Mann-Whitney test is based on the comparison of each observation $x_1^i$ from sample 1 with each observation $x_2^i$ from sample 2. This means that the data (sample 1 & sample 2 combined) must be sorted in ascending order and each observation is indexed from 1 to $N = n_1 + n_2$. If two or more observations are the same, the indexes (ranks) of those observations should be the average of the indexes. The total number of pairwise comparisons that can be made is $n_1 n_2$.

If the samples have the same median then each $x_1^i$ has an equal chance (i.e., probability 1/2) of being greater or smaller than each $x_2^i$. Hence, the hypotheses can be formally stated as

$$H_0 : \mathcal{P}\big(x_1^i > x_2^j\big) = \frac{1}{2} \text{ versus } H_1 : \mathcal{P}\big(x_1^i > x_2^j\big) \neq \frac{1}{2}. \tag{15}$$

The Mann-Whitney U statistic is defined as

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \tag{16}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2, \tag{17}$$

where $R_1$ is the sum of the ranks assigned to sample 1 (i.e., sum of the indexes assigned to the observations after sorting the combined samples), and $R_2$ is the sum of the ranks assigned to sample 2. Here, $U_1$ is the number of times an $x_1^i$ from sample 1 is greater than an $x_2^i$ from sample 2. Similarly, $U_2$ is the number of times an $x_1^i$ from sample 1 is smaller than an $x_2^i$ from sample 2. The sum of the ranks of the two samples is given by

$$(R_1 + R_2) = 1 + 2 + 3 + \cdots + N = \frac{N(N+1)}{2}. \tag{18}$$

One can check that $U_1 + U_2 = n_1 n_2$, hence once either $U_1$ or $U_2$ is found, the other one can be calculated from the addition rule.

Now define $U = \min(U_1, U_2)$. The distribution of $U$ under $H_0$ has been found by Mann & Whitney, and the probabilities for small values of $n_1, n_2$ have a table known as the Mann-Whitney U tables. If one finds that the probability $\mathcal{P}\big(U = \min(U_1, U_2)\big) < \alpha$, then the null hypothesis $(H_0)$ is rejected at level $\alpha$.

#### 1. *Normal Approximation*

In most A/B testing experiments, the obtained samples sizes are usually large (i.e., $n_1 n_2 > 20$). In this case, a normal approximation can be used by defining:

$$\mu_U = \frac{n_1 n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 n_2 (N+1)}{12}}, \tag{19}$$

where $\mu_U$ and $\sigma_U$ are the mean and standard deviation of the $U$ distribution respectively. The corresponding test statistic is given by

$$z = \frac{U - \mu_U}{\sigma_U}, \tag{20}$$

where $z \sim \mathcal{N}(0, 1)$. In the normal approximation, the computation of $U$ can be done by either $U_1$ or $U_2$, because $|z|$ is the same whether $U_1$ or $U_2$ is used in computing $U$.

The p-value is calculated from the standard normal distribution under $H_0$, given by

$$\begin{aligned} \text{p-value} &= \mathcal{P}\Big(|Z| \geq |z| \Big| Z = \mathcal{N}(0,1)\Big) \\ &= 2\mathcal{P}\Big(Z < -|z| \Big| Z = \mathcal{N}(0,1)\Big) \\ &= 2\Phi(-|z|), \end{aligned} \tag{21}$$

where $\Phi$ is the CDF of normal distribution.

1. *Dealing with ties*: When two or more observations have the same value, it is called a tie. If the ties occur between two or more observations in the same sample, the value of $U$ is unaffected. However, if the ties occur between two or more observations in both samples, the value of $U$ is affected. In this

case, the standard deviation is renormalized as

$$\sigma_U = \sqrt{\left(\frac{n_1 n_2}{N(N-1)}\right)\left(\frac{N^3-N}{12} - \sum_j^{n_s} \frac{t_j^3 - t_j}{12}\right)}, \qquad (22)$$

where $n_s$ is the number of samples in which ties occur, and $t_j$ is the number of observations tied for a given rank in sample $j$. We can see that in the absence of ties (i.e., $t = 0$), Eq. (22) reduces to $\sigma_U$ in Eq. (19).

2. *Programming package*: The Python SciPy package scipy.stats.mannwhitneyu can be used to compute the Mann-Whitney U statistic and its p-value.

## V. POWER ANALYSIS

### A. Definition

The power of a statistical test is the probability that the test correctly rejects the null hypothesis, when the alternative hypothesis is true. In other words, it is the probability of not making the type II error. It is also stated as the probability of detecting a treatment effect, when the effect is really there. Mathematically,

$$\text{Power} = \mathcal{P}(\text{rejecting } H_0 | H_1 \text{ true}) = 1 - \beta. \qquad (23)$$

It is common to conduct power analysis before starting an experimental design, to determine the (minimum) sample size needed to detect a given effect size. The use of post hoc power analysis, that is after the experiment has been conducted and found to be nonsignificant (say, p-value $> \alpha$), is discouraged [13]. The reason being that the p-value is directly related to the power of the test. Therefore, when a test yields p-value $> \alpha$, it is safe to conclude that "there is not enough data to reach any conclusion", as opposed to concluding "there is no significant difference between the two groups".

### B. Properties

Power analysis answers questions like "how big a sample size do I need?" and "how much statistical power does my study have [14, 15]?" The power analysis of a statistical test depends upon four main parameters: the effect size, the sample size, the alpha significance level, and the power of the statistical test. They are related that each one of them is a function of the other three, hence when any three of them are fixed, the fourth one is completely determined [14].

1. *Effect size or minimum detectable effect* - is the degree to which the phenomenon is present in the population or the degree to which the null hypothesis is false (note: the null hypothesis means that effect size is zero). The effect size is defined as

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}}, \qquad (24)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the two sample means, and $s_{\text{pooled}}$ is the sample pooled standard deviation estimate of the population standard deviation given by Eq. (7). Combining Eqs. (5) and (24), we obtain the relationship

$$t^* = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}. \qquad (25)$$

2. *Sample size* - is the number of observations which determines the amount of sample error in the result.

3. *Significance level* $\alpha$ - is the probability of committing type I error. Normally, we set $\alpha = 0.05$.

4. *Statistical power* - is the probability that the test correctly rejects the null hypothesis. If the acceptable level of type II error is $\beta = 0.20$, then the desired power is $1 - \beta = 0.80$. The industry recommended power is at least 0.80 or 80%.

### C. Use Cases

Here, I will list two cases in which power analysis can be applied in most cases of physical interests.

1. *Sample size as a function of effect size, power, and* $\alpha$: In this scenario, the experimenter has to come up with educated guesses and assumptions for the effect size. This is usually based on past experience of similar studies. Next, he sets a significance level, say $\alpha = 0.05$, and specifies the amount of power he desires, usually at least 80% power. Then the minimum sample sizes that meet the specifications can be determined. The use of power analysis in this manner is mainly before an experiment is conducted to determine the (minimum) sample size needed to detect a particular effect size.

2. *Power as a function of effect size, sample size, and* $\alpha$: Power analysis in this case can also be used before an experiment is conducted, if the experimenter already knows that only a fixed number of subjects or randomization units (e.g., users) are available (or that he only has the budget for fixed number subjects). The experimenter may want to know if he will have enough power to justify actually doing the study. In most cases of physical interest, there is really no point to conduct a study that is seriously underpowered [16].

   Power analysis in this case can also be post hoc, i.e., after the experiment has been conducted and found to be nonsignificant. As we mentioned above, the use of power analysis in this manner is not recommended [13].

*Programming package*: To conduct power analysis, one can use the Python library statsmodels.stats.power.TTestIndPower.

[1] R. Kohavi, D. Tang, and Y. Xu (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (1 ed.). Cambridge University Press.

[2] G. Casella and R. L. Berger (2002). Statistical Inference (2 ed.). Duxbury.

[3] L. Wasserman (2004). All of Statistics: A Concise Course in Statistical Inference. Springer New York.

[4] S. Siegel (1956). Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Company, INC.

[5] R. Willett (2016). t-tests and p-values.

[6] R. A. Fisher, Metron, **5**, 90 (1925). See also Wikipedia page for Student's t-distribution.

[7] Student, Biometrika, **6**, 1 (1908).

[8] R. A. Fisher, Annals of Eugenics, **8** 391 (1935).

[9] B. L. Welch, Biometrika. **34**, 28 (1947).

[10] H. B. Mann and D. R. Whitney, Ann. Math. Statist. **18**, 50 (1947).

[11] Rosie Shier (2004). Statistics: 2.3 The Mann-Whitney U Test, Mathematics Learning Support Centre.

[12] N. Nachar, Tutorials in Quantitative Methods for Psychology **4**, 13 (2008).

[13] J. M. Hoenig and D. M. Heisey, The American Statistician, **55**, 1 (2001).

[14] J. Cohen (1988). Statistical Power Analysis for the Behavioral Sciences (2 ed.). Lawrence Erlbaum Associates.

[15] P. D. Ellis (2010). The Essential Guide to Effect Sizes. Cambridge University Press.

[16] UCLA Statistical Consulting, blog link