machine learning &
knowledge extraction

MDPI

Review

Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference

Frank Emmert-Streib [1,2,*] **and Matthias Dehmer** [3,4,5]

[1] Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland
[2] Institute of Biosciences and Medical Technology, Tampere University, 33520 Tampere, Finland
[3] Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Steyr Campus, 4040 Steyr, Austria
[4] Department of Mechatronics and Biomedical Computer Science, University for Health Sciences, Medical Informatics and Technology (UMIT), 6060 Hall, Tyrol, Austria
[5] College of Computer and Control Engineering, Nankai University, Tianjin 300000, China
[*] Correspondence: v@bio-complexity.com; Tel.: +358-50-301-5353

Received: 27 July 2019; Accepted: 9 August 2019; Published: 12 August 2019

check for updates

Abstract: Statistical hypothesis testing is among the most misunderstood quantitative analysis methods from data science. Despite its seeming simplicity, it has complex interdependencies between its procedural components. In this paper, we discuss the underlying logic behind statistical hypothesis testing, the formal meaning of its components and their connections. Our presentation is applicable to all statistical hypothesis tests as generic backbone and, hence, useful across all application domains in data science and artificial intelligence.

Keywords: hypothesis testing; machine learning; statistics; data science; statistical inference

1. Introduction

We are living in an era that is characterized by the availability of big data. In order to emphasize the importance of this, data have been called the 'oil of the 21st Century' [1]. However, for dealing with the challenges posed by such data, advanced analysis methods are needed. A very important type of analysis method on which we focus in this paper is statistical hypothesis tests.

The first method that can be considered a hypothesis test is related back to John Arbuthnot in 1710 [2,3]. However, the modern form of statistical hypothesis testing originated from the combination of work from R. A. Fisher, Jerzy Neyman and Egon Pearson [4–8]. It can be considered one of the first statistical inference methods and it is till this day widely used [9]. Examples for applications can be found in all areas of science, including medicine, biology, business, marketing, finance, psychology and social sciences. Specific examples in biology include the identification of differentially expressed genes or pathways [10–14], in marketing it is used to identify the efficiency of marketing campaigns or the alteration of consumer behavior [15], in medicine it can be used to assess surgical procedures, treatments or the effectiveness of medications [16–18], in pharmacology to identify the effect of drugs [19] and in psychology it has been used to evaluate the effect of meditation [20].

In this paper, we provide a primer of statistical hypothesis testing and its constituting components. We place a particular focus on the accessibility of our presentation due to the fact that the understanding of hypothesis testing causes in general widespread problems [21,22].

A problem with explaining hypothesis testing is that either the explanations are too mathematical [9] or too non-mathematical [23,24]. However, a middle ground is needed for the beginner and interdisciplinary scientist in order to avoid the study from becoming tedious and frustrating yet delivering all needed details for a thorough understanding. For this reason we are aiming at an intermediate level that is accessible for data scientists having a mixed background [25].

In the following, we first discuss the basic idea of hypothesis testing. Then we discuss the seven main components it consists of and their interconnections. After this we address potential errors resulting from hypothesis testing and the meaning of the power. Furthermore, we show that a confidence interval complements the value provided by a test statistic. Then we present an example that serves also as a warning. Finally, we provide some historical notes and discuss common misconceptions of *p*-values.

2. Basic Idea of Hypothesis Testing

The principle idea of a statistical hypothesis test is to decide if a data sample is typical or atypical compared to a population assuming a hypothesis we formulated about the population is true. Here a data sample refers to a small portion of entities taken from a population, for example, via an experiment, whereas the population comprises all possible entities.

In Figure 1 we give an intuitive example for the basic idea of hypothesis testing. In this particular example the population consists of all ducks and the data sample is one individual duck randomly drawn from the entire population. In statistics 'randomly drawn' is referred to as 'sampling'. In order to perform the comparison between the data sample and the population one needs to introduce a quantification of the situation. In our case this quantification consists in a mapping from a duck to a number. This number could correspond to, for example, the body weight, the beak size, the body size or the hair length of a duck. In statistics this mapping is called test statistic.

A key component in hypothesis testing is of course a 'hypothesis'. The hypothesis is a quantitative statement we formulate about the population value of the test statistic. In our case it could be about the body parts of a duck, for example, body size. A particular hypothesis we can formulate is: The mean body size equals 20 cm. Such a hypothesis is called the null hypothesis H_0.

Assuming now we are having a population of ducks having a body size of 20 cm including natural variations. Due to the fact that the population consists of (infinite) many ducks and for each we are obtaining such a quantification this results in a probability distribution, called the sampling distribution, for the mean body size. Here it is important to note that our population is a hypothetical population which obeys our null hypothesis. In other words, the null hypothesis specifies the population completely.

Having now a numerical value of the test statistic, representing the data sample and the sampling distribution, representing the population, we can compare both with each other in order to evaluate the null hypothesis that we have formulated. From this comparison we obtain another numerical value, called the *p*-values, which quantifies the typicality or atypicality of the configuration assuming the null hypothesis is true. Finally, based on the *p*-values a decision is made.

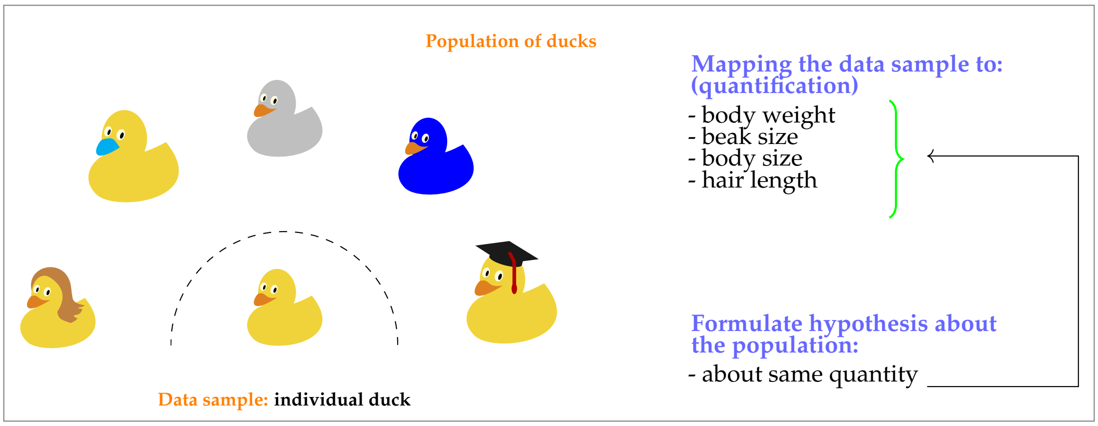
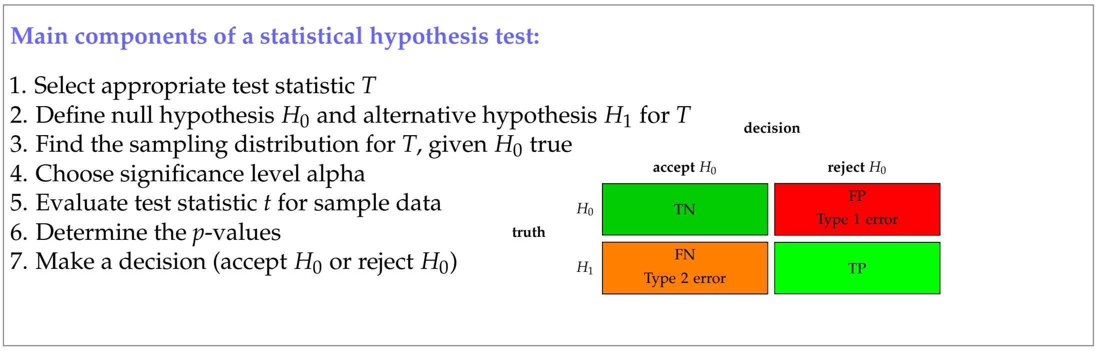


Figure 1. Intuitive example explaining the basic idea underlying an one-sample hypothesis test.

On a technical note, we want to remark that due to the fact that in the above problem there is only one population involved this is called an one-sample hypothesis test. However, the principal idea extends also to hypothesis tests involving more than population.

3. Key Components of Hypothesis Testing

In the following sections, we will formalize the example discussed above. In general, regardless of the specific hypothesis test one is conducting, there are seven components common to all hypothesis tests. These components are summarized in Figure 2. We listed these components in the order they are entering the process when performing a hypothesis test. For this reason they can be also considered as steps of a hypothesis test. Due to the fact that they are interconnected with each other their logical order is important. Overall this means a hypothesis test is a procedure that needs to be executed. In the following subsections, we will discuss each of these seven procedural components in detail.



Main components of a statistical hypothesis test:

1. Select appropriate test statistic T
2. Define null hypothesis H_0 and alternative hypothesis H_1 for T
3. Find the sampling distribution for T, given H_0 true
4. Choose significance level alpha
5. Evaluate test statistic t for sample data
6. Determine the p-values
7. Make a decision (accept H_0 or reject H_0)

	accept H_0	reject H_0
H_0	TN	FP Type 1 error
H_1	FN Type 2 error	TP

Figure 2. Main components that are common to all hypothesis tests.

3.1. Step 1: Select Test Statistic

Put simply, a test statistic quantifies a data sample. In statistics the term 'statistic' refers to any mapping (or function) between a data sample and a numerical value. Popular examples are the mean value or the variance. Formally, the test statistic can be written as

$$t_n = T(D(n)) \tag{1}$$

whereas $D(n) = \{x_1, \ldots, x_n\}$ is a data sample with sample size n. Here we denoted the mapping by T and the value we obtain by t_n. Typically the test statistic can assume real values, that is, $t_n \in \mathbb{R}$ but restrictions are possible.

A test statistic assumes a central role in a hypothesis test because by deciding which test statistic to use one determines a hypothesis test to a large extend. The reason is that it will enter the hypotheses we formulate in step 2. For this reason one needs to carefully select a test statistic that is of interest and importance for the conducted study.

We would like to emphasize that in this step, we select the test statistics but we neither evaluate it nor we use it yet. This is done in step 5.

3.2. Step 2: Null Hypothesis H_0 and Alternative Hypothesis H_1

At this step, we define two hypotheses which are called the null hypothesis H_0 and the alternative hypothesis H_1. Both hypotheses make statements about the population value of the test statistic and are mutually exclusive. For the test statistic $t = T(D)$ we selected in step 1, we call the population value of t as θ. Based on this we can formulate the following hypotheses:

null hypothesis: H_0: $\theta = \theta_0$
alternative hypothesis: H_1: $\theta > \theta_0$

As one can see, the way the two hypotheses are formulated, the value of the population parameter θ can only be true for one statement but not for both. For instance, either $\theta = \theta_0$ is true but then the alternative hypothesis H_1 is false or $\theta > \theta_0$ is true but then the null hypothesis H_0 is false.

In Figure 2, we show the four possible outcomes of a hypothesis test. Each of these outcomes has a specific name that is commonly used. For instance, if the null hypothesis is false and we reject H_0 this is called a 'true positive' (TP) decision. The reason for calling it 'positive' is related to the asymmetric meaning of a hypothesis test, because rejecting H_0 when H_0 is false is more informative than accepting H_0 when H_0 is true. In this case one can consider the outcome of a hypothesis test a positive result.

The alternative hypothesis formulated above is an examples for a one-side hypothesis. Specifically, we formulated a right-sided hypothesis because the alternative assumes values larger than θ_0. In addition, we can formulate a left-sided alternative hypothesis stating

alternative hypothesis: H_1: $\theta < \theta_0$

Furthermore, we can formulate a two-side alternative hypothesis that is indifferent regarding the side by

alternative hypothesis: H_1: $\theta \neq \theta_0$

Despite the fact that there are hundreds of different hypothesis tests [26], the above description principally holds for all of them. However, this does not mean that if you understand one hypothesis test you understand all but if you understand the *principle* of one hypothesis test you understand the *principle* of all.

In order to connect the test statistic t, which is a sample value, with its population value θ one needs to know the probability distribution of the test statistic. Because of this connection, this probability distribution received a special name and is called the *sampling distribution* of the test statistic. It is important to emphasize that the sampling distribution represents the values of the test statistic assuming the null hypothesis is true. This means that in this case the population value of θ is θ_0.

Let's assume for now that we know the sampling distribution for our test statistic. By comparing the particular value t of our test statistic with the sampling distribution in a way that is determined by the way we formulated the null and the alternative hypothesis, we obtain a quantification for the 'typicality' of this value with respect to the sampling distribution, assuming the null hypothesis is true.

3.3. Step 3: Sampling Distribution

In our general discussion about the principle idea of a hypothesis test above, we mentioned that the connection between a test statistic and its sampling distribution is crucial for any hypothesis test. For this reason, we elaborate in this section on this point in more detail.

In this section, we want to answer the following questions:

1. What is the sampling distribution?
2. How to obtain the sampling distribution?
3. How to use the sampling distribution?

To 1.: First of all, the sampling distribution is a probability distribution. The meaning of this sampling distribution is that it is the distribution of the test statistic T, which is a random variable, given some assumptions. We can make this statement more precise by defining the sampling distribution of the null hypothesis as follows.

Definition 1. *Let $X(n) = \{X_1, \ldots, X_n\}$ be a random sample from a population with $X_i \sim P_{pop}$ $\forall i$ and $T(X(n))$ be a test statistic. Then the probability distribution $f_n(x|H_0$ true) of $T(X(n))$, assuming H_0 is true, is called the sampling distribution of the null hypothesis or the null distribution.*

Similarly, one defines the sampling distribution of the alternative hypothesis by $f_n(x|H_1$ true). Since there are only two different hypotheses, H_0 and H_1, there are only two different sampling distributions in this context. However, we would like to note that sampling distributions are also playing a role outside statistical hypothesis testing, for example, for estimation theory or data Bootstrapping [27].

There are several points in the above definition we would like to highlight. First, the distribution P_{pop} from which the random variables X_i are sampled can assume any form and is *not limited to*, for example, a normal distribution. Second, the test statistic is a random variable itself because it is a function of random variables. For this reason there exists a distribution that belongs to this random variable in a way that the values of this random variable are samples thereof. Third, the test statistic is a function of the sample size n and for this reason also the sampling distribution is a function of n. That means, if we change the sample size n, we change the sampling distribution. Fourth, the fact that $f_n(x|H_0 \text{ true})$ is the probability distribution of $T(X(n))$ means that by taking infinite many samples from $f_n(x|H_0 \text{ true})$ in the form, $T(X(n)) \sim f_n(x|H_0 \text{ true})$, we can perfectly reconstruct the distribution $f_n(x|H_0 \text{ true})$ itself. The last point allows under certain conditions a numerical approximation of the sampling distribution, as we will see in the following example.

Examples

Suppose we have a random sample $X(n) = \{X_1, \ldots, X_n\}$ of size n whereas each data point X_i is sampled from a gamma distribution with $\alpha = 4$ and $\beta = 2$, that is, $X_i \sim gamma(\alpha = 4, \beta = 2)$. Furthermore, let's use the mean value as a test statistic, that is,

$$t_n = T(X(n)) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2}$$

In Figure 3A–C, we show three examples for three different values of n (in A $n = 1$, in B $n = 3$ and in C $n = 10$) when drawing E = 100,000 samples $X(n)$, from which we estimate E = 100,000 different mean values T. Specifically, in Figure 3A–C we show density estimates of these 100,000 values. As indicated above, in the limit of infinite many samples E, the approximate sampling distribution $P_s(n, E)$ will become the (theoretical) sampling distribution, that is,

$$f_n(x|H_0 \text{ true}) = \lim_{E \to \infty} P_s(n, E) \tag{3}$$

as a function of the sample size n.

For $n = 1$, we obtain the special case that the sampling distribution is the same as the underlying distribution of the population P_{pop}, which is in our case a gamma distribution with the parameters $\alpha =$ 4 and $\beta = 2$, shown in Figure 3A. For all other $n > 1$, we observe a transformation in the distributional shape of the sampling distribution, as seen in Figure 3B,C. However, this transformation should be familiar to us because from the Central Limit Theorem we know that the mean of $\{X_1, \ldots, X_n\}$ independent samples with mean μ and variance σ^2 follows a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, that is,

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}). \tag{4}$$

We notice that this result is only strictly true in the limit of large n. However, in Figure 3D, we show a qq-plot that demonstrates that already for $n = 10$ the resulting distribution, $P_s(n = 10, E$ = 100,000), is quite close to such a normal distribution (with the appropriate parameters).

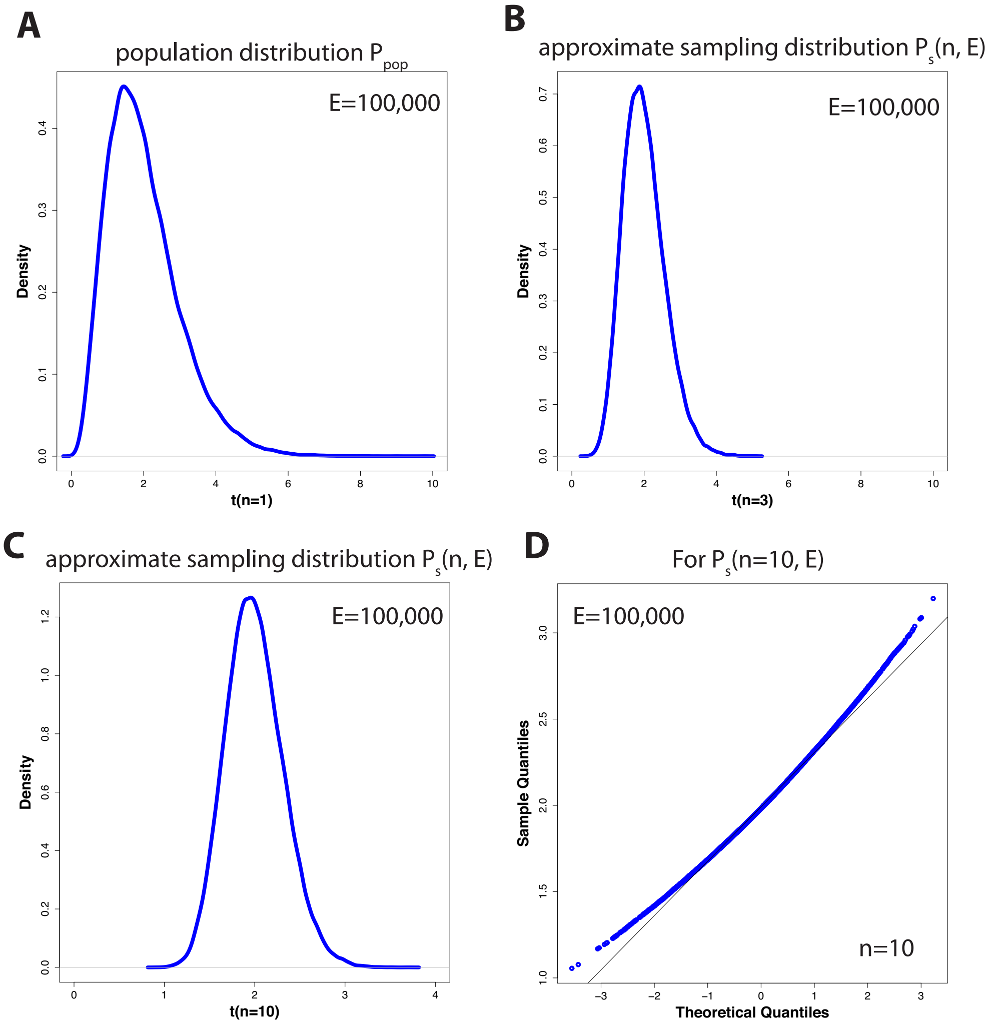


Figure 3. In Figure (**A**–**C**) we show approximate sampling distributions for different values of the sample size n. Figure (**A**) shows $P_s(n = 1, E = 100{,}000)$ which is equal to the population distribution of X_i. Figure (**D**) shows a qq-plot comparing $P_s(n = 10, E = 100{,}000)$ with a normal distribution.

We would like to remind that the Central Limit Theorem holds for arbitrarily iid (independent and identically distributed) random variables $\{X_1, \ldots, X_n\}$. Hence, the sampling distribution for the mean is always the normal distribution given in Equation (4).

There is one further simplification we obtain by applying a so called z-transformation of the mean value of $\bar{X}$ to Z by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{5}$$

because the distribution of Z is a standard normal distribution, that is,

$$Z \sim N(0,1). \tag{6}$$

Now, we reached an important point where we need to ask ourself if we are done. This depends on our knowledge about the variance. If we know the variance σ^2 the sampling distribution of our transformed mean $\bar{X}$, we called Z, is a standard normal distribution. However, if we do not know the variance σ^2, we cannot perform the z-transformation in Equation (5), because this transformation depends on σ. In this case, we need to estimate the variance of the random sample $\{X_1, \ldots, X_n\}$ by

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \hat{X} \right)^2. \tag{7}$$

Then we can use the estimate for the variance to use it for the following t-transformation

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}. \tag{8}$$

Despite the fact that this t-transformation is formally similar to the z-transformation in Equation (5) the resulting random variable T does not follow a standard normal distribution but a Students' t-distribution with $n-1$ degrees of freedom (dof). We want to mention that this holds strictly for $X_i \sim N(\mu, \sigma)$, that is, normal distributed samples.

The following Table 1 summarizes the results from this section regarding the sampling distribution of the z-score (Equation (5)) and the t-score (Equation (8)).

Table 1. Sampling distribution of the z-score and the t-score.

Test Statistic	Sampling Distribution	Knowledge about Parameters
z-score	N(0,1)	σ^2 needs to be known
t-score	Students' t-distribution, $n-1$ dof	none

3.4. Step 4: Significance Level α

The significance level α is a number between zero and one, that is, $\alpha \in [0, 1]$. It has the meaning

$$\alpha = \text{P(Type 1 error)} = \text{P(reject } H_0 | H_0 \text{ true)} \tag{9}$$

giving the probability to reject H_0 provided H_0 is true. That means it gives us the probability of making a Type 1 error resulting in a false positive decision.

When conducting a hypothesis test, we have the freedom to choose this value. However, when deciding about its numerical value one needs to be aware of potential consequences. Possibly the most frequent choice of α is 0.05, however, for Genome-Wide Association Studies (GWAS) values as low as 10^{-8} are used [28]. The reason for such a wide variety of used values is in the possible consequences in the different application domains. For GWAS, Type 1 errors can result in wasting millions of Dollars because follow-up experiments in this field are very costly. Hence, α is chosen very small.

Finally, we want to remark that formally we obtain the value of the right-hand side of Equation (9) by integrating the sampling distribution, as given by Equation (13) (discussed below).

3.5. Step 5: Evaluate Test Statistic from Data

This step is our connection to the real world, as represented by the data, because everything until here has been theoretical. For $D(n) = X(n) = \{x_1, \ldots, x_n\}$ we estimate the numerical value of the test statistic selected in Step 1 giving

$$t_n = T(D(n)). \tag{10}$$

Here t_n represents a particular numerical value obtained from the observed data $D(n)$. Due to the fact that our data set depends on the number of samples n, also this numerical value will be dependent on n. This is explicitly indicated by the subscript.

3.6. Step 6: Determine the p-Values

For determining the p-values of a hypothesis test, we need to use the sampling distribution (Step 3) and the estimated test statistic t_n (Step 5). That means the p-values results from a comparison of theoretical assumptions (sampling distribution) with real observations (data sample) assuming H_0 is true. This situation is visualized in Figure 4 for a right-sided alternative hypothesis. The p-values is the probability for observing more extreme values than the test statistic t_n assuming H_0 is true

$$p = P(\text{observe} \times \text{ at least as extreme as } |t| \; |H_0 \text{ is true}) = P(x \geq |t| \; |H_0 \text{ is true}) \tag{11}$$

Formally it is obtained by an integral over the sampling distribution

$$p = \int_{t_n}^{\infty} f_n(x'|H_0 \text{ true}) dx' \tag{12}$$

The final decision if we reject or accept the null hypothesis will be based on the numerical value of p.

Furthermore, we can use the following integral

$$\alpha = \int_{\theta_c}^{\infty} f_n(x'|H_0 \text{ true}) dx' \tag{13}$$

to solve for θ_c. That means, the significance level α implies a threshold θ_c. This threshold can also be used to make a decision about H_0.

We would like to emphasize that due to the fact that the test statistic is a random variable also the p-values is a random variable since it depends on the test statistic [29].

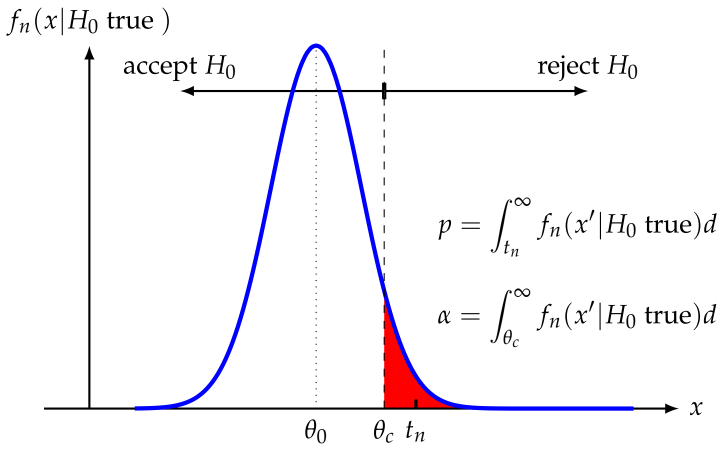

$f_n(x|H_0 \text{ true })$
accept H_0
reject H_0
$p = \int_{t_n}^{\infty} f_n(x'|H_0 \text{ true}) dx'$
$\alpha = \int_{\theta_c}^{\infty} f_n(x'|H_0 \text{ true}) dx'$
x
θ_0
θ_c
t_n

Figure 4. Determining the p-values from the sampling distribution of the test statistic.

Remark 1. *The sample size n has an influence on the numerical analysis of the problem. For this reason the test statistic and the sampling distribution are indexed by it. However, it has no effect on the formulation and expression of the hypothesis because we make statements about a population value that hold for all n.*

3.7. Step 7: Make a Decision about the Null Hypothesis

In the final step we are making a decision about the null hypothesis. In order to do this there are two alternative ways. First, we can make a decision based on the p-values or, second, we make a decision based on the value of the test statistic t_n.

1. Decision based on the p-values:

$$\text{If } p < \alpha \text{ reject } \mathrm{H}_0 \tag{14}$$

2. 2. Decision based on the threshold θ_c:

$$\text{If } t_n > \theta_c \text{ reject } \mathrm{H}_0 \tag{15}$$

In case we cannot reject the null hypothesis we accept it.

4. Type 2 Error and Power

When making binary decisions there is a number of errors one can make [30]. In this section, we go one step back and take a more theoretical look on a hypothesis test with respect to the possible errors one can make. In section 'Step 2: Null hypothesis H_0 and alternative hypothesis H_1' we discussed that there are two possible errors one can make, a false positive and a false negative and when discussing Step 4, we introduced formally the meaning of a Type 1 error. Now we extend this discussion to the Type 2 error.

As mentioned previously, there are only two possible configurations one needs to distinguish. Either H_0 is true or it is false. If H_0 is true (false) it is equally correct to say H_1 is false (true). Now, let's assume H_1 is true. For evaluating the Type 2 error we require the sampling distribution assuming H_1 is true. However, for performing a hypothesis test, as discussed in the previous sections (see Figure 2), we do not need to know the sampling distribution assuming H_1 is true. Instead, we need to know the sampling distribution assuming H_0 is true because this distribution corresponds to the null hypothesis. The good news is the sampling distribution assuming H_1 is true can be easily obtained if we make the alternative hypothesis more precise. Let's assume we are testing the following hypothesis.

null hypothesis: H_0: $\theta = \theta_0$
alternative hypothesis: H_1: $\theta > \theta_0$

In this case H_0 is precisely specified because it sets the population parameter θ to θ_0. In contrast, H_1 limits the range of possible values for θ but does not set it to a particular value.

For determining the Type 2 error we need to set θ in the alternative hypothesis to a particular value. So let's set the population parameter $\theta = \theta_1$ in H_1 for $\theta_1 > \theta_0$. In Figure 5 we visualize the sampling distribution for H_1 and H_0.

If we reject H_0 when H_1 is true, this is a correct decision and the green area in Figure 5 represents the corresponding probability for this, formally given by

$$1 - \beta = P(\text{reject } H_0 | H_1 \text{ is true}) = \int_{\theta_c}^{\infty} f_n(x' | H_1 \text{ true }) dx'. \tag{16}$$

For short this probability is usually denoted by $1 - \beta$ and called the *power* of a test.

On the other hand, if we do not reject H_0 when H_1 is true, we make an error, given by

$$\beta = P(\text{Type 2 error}) = P(\text{do not reject } H_0 | H_1 \text{ is true}). \tag{17}$$

This is called a Type 2 error. In Figure 5, we highlight the Type 2 error probability in orange.

We would like to emphasize that the Type 1 error and the Type 2 error are both long-run frequencies for repeated experiments. That means both probabilities give the error when repeating the exact same test many times. This is in contrast to the p-values, which is the probability for a given data sample. Hence, the p-values does not allow to draw conclusions for repeated experiments.

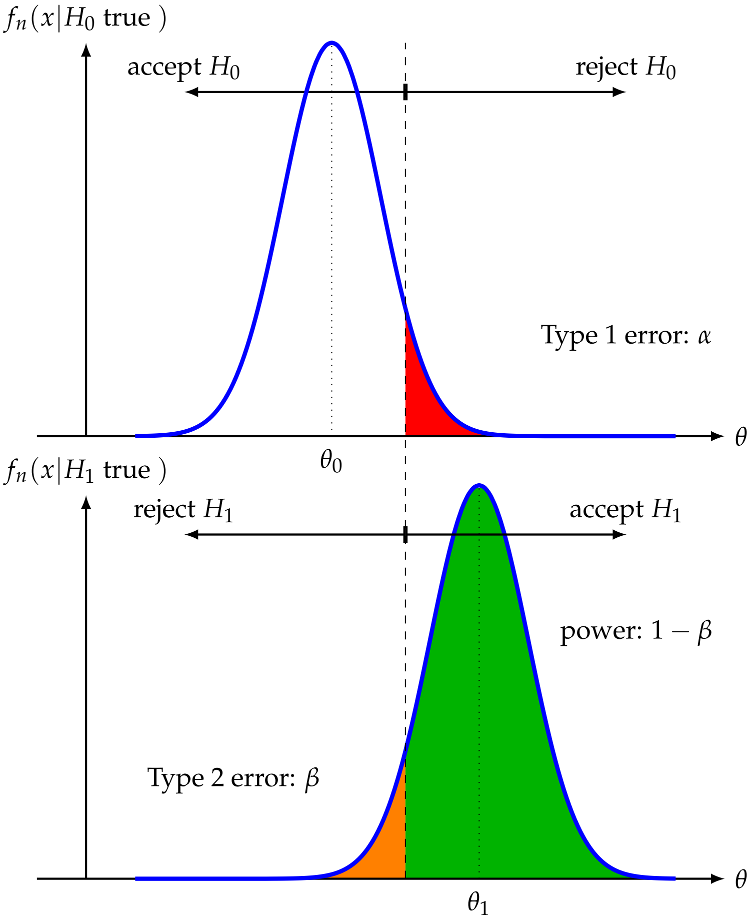


Figure 5. Visualization of the sampling distribution for H_0 and H_1 assuming a fixed sample size n.

Connections between Power and Errors

From Figure 5 we can see the relation between power ($1-\beta$), Type 1 error (α) and Type 2 error (β), summarized in Figure 6. Ideally, one would like to have a test with a high power and low Type 1 error and low Type 2 error. However, from Figure 5 we see that these three entities are not independent from each other. Specifically, if we increase the power ($1-\beta$) by changing α we increase the Type 1 error (α) because this will reduce the critical value θ_c. In contrast, reducing α leads to an increase in Type 2 error (β) and a reduction in power. Hence, in practice, one needs to make a compromise between the ideal goals.

		decision	
		accept H_0	**reject** H_0
truth	H_0	$1-\alpha$ =P(accept H_0| H_0 true)	α = P(reject H_0| H_0 true)
	H_1	β = P(reject H_0| H_1 true)	$1-\beta$ =P(accept H_0| H_1 true)

Figure 6. Overview of the different errors as a result from hypothesis testing and their probabilistic meaning.

For the discussion above, we assumed a fixed sample size n. However, as we discussed in the example of section 'Step 3: Sampling distribution', the variance of the sampling distribution depends on the sample size via the standard error in the way

$$\sigma(n)^2 = \frac{\sigma_{pop}^2}{n} \tag{18}$$

This opens another way to increase the power and to minimize the Type 2 error by increasing the sample size n. By keeping the population means θ_0 and θ_1 unchanged but increasing the sample size n' to a value larger than n, that is, $n' > n$, the sampling distributions for H_0 and H_1 become narrower because their variances decrease according to Equation (18). Hence, as a consequence of an increased sample size the overlap between the distributions, as measured by β, is reduced leading to an increase in the power and a decrease in Type 2 error for an unchanged value of the significance level α. In the extreme case for $n \to \infty$ the power approaches 1 and the Type 2 error 0, for a fixed Type 1 error α.

From this discussion the importance of the sample size in a study becomes apparent as a control mechanism to influence the resulting power and the Type 2 error.

5. Confidence Intervals

The test statistic is a function of the data (see Step 1 in Section 3.1) and, hence, it is a random variable. That means there is a variability of a test statistic because its value changes for different samples. In order to quantify the interval within which such values fall, one uses a confidence interval (CI) [31,32].

Definition 2. *The interval $I = [a, b]$ is called a confidence interval for parameter θ if it contains this parameter with probability $1 - \alpha$ for $\alpha \in [0, 1]$, that is,*

$$P\big(a \leq \theta \leq b\big) = 1 - \alpha. \tag{19}$$

The interpretation of a CI $I = [a, b]$ is that for repeated samples the confidence intervals of these are expected to contain the true θ with probability $1 - \alpha$. Here it is important to note that θ is fixed because it is a population value. What is random is the estimate of the boundaries of the CI, that is, a and b. Hence, for repeated samples, θ is fixed but I is a random interval.

The connection between a $1 - \alpha$ confidence interval and a hypothesis test for a significance level of α is that if the value of the test statistic falls within the CI then we do not reject the null hypothesis. On the other hand, if the confidence interval does not contain the value of the test statistic, we reject the null hypothesis. Hence, the decisions reached by both approaches agree always with each other.

If one does not make any assumption about the shape of the probability distribution, for example, symmetry around zero, there are infinite many CIs because neither the starting nor the ending values of a and b are uniquely defined but follow from assumptions. Frequently, one is interested in obtaining a CI for a quantile separation of the data in the form

$$P\left(q_{\alpha/2} \leq \theta \leq q_{1-\alpha/2}\right) = 1 - \alpha \tag{20}$$

whereas $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are quantiles of the sampling distribution with respect to $100\alpha/2\%$ respectively $100(1 - \alpha/2)\%$ of the data.

5.1. Confidence Intervals for a Population Mean with Known Variance

From the central limit theorem we know that the sum of random variables $\hat{\theta} = 1/n \sum x_i$ is normal distributed. If we normalize this by

$$Z = \frac{\hat{\theta} - \mathbb{E}[\hat{\theta}]}{\sigma(\hat{\theta})} \tag{21}$$

then Z follows a standard normal distribution, that is, $N(0,1)$, whereas $\sigma(\hat{\theta})$ is the standard error

$$\sigma(\hat{\theta}) = \frac{\sigma}{\sqrt{n}} \tag{22}$$

of $\hat{\theta}$.

Adjusting the definition of a confidence interval in Equation (20) to our problem gives

$$P\left(q_{\alpha/2} \leq Z \leq q_{1-\alpha/2}\right) = 1 - \alpha \tag{23}$$

with

$$q_{\alpha/2} = -z_{\alpha/2} \tag{24}$$

$$q_{1-\alpha/2} = z_{\alpha/2} \tag{25}$$

Here the values of $\pm z_{\alpha/2}$ are obtained by solving the equations for a standard normal distributed probability

$$P\big(Z < -z_{\alpha/2}\big) = \alpha/2 \tag{26}$$

$$P\big(Z > z_{\alpha/2}\big) = \alpha/2 \tag{27}$$

Using these and solving the inequality in Equation (23) for the expectation value gives the confidence interval $I = [a, b]$ with

$$a = \hat{\theta} - z_{\alpha/2}\sigma(\hat{\theta}) = \hat{\theta} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{28}$$

$$b = \hat{\theta} + z_{\alpha/2}\sigma(\hat{\theta}) = \hat{\theta} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{29}$$

Here we assumed that σ is know. Hence, the above CI is valid for a z-test.

5.2. Confidence Intervals for a Population Mean with Unknown Variance

If we assume that σ is not know then the sampling distribution of a population mean is Student's T-distribution and σ needs to be estimated from samples by the sample standard deviation s. In this case a similar derivation as above results in

$$a = \hat{\theta} - t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{30}$$

$$b = \hat{\theta} + t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{31}$$

Here $\pm t_{\alpha/2}$ are critical values for a Student's T-distribution, obtained similarly as in Equations (26) and (27). Such a CI is valid for a t-test.

5.3. Bootstrap Confidence Intervals

In case a sampling distribution is not given in analytical form numerical approaches need to be used. In such a situation a CI can be numerically obtained via nonparametric Bootstrap [33]. This is the most generic way to obtain a CI. By utilizing the augmented definition in Equation (20) for any test statistic $\hat{\theta}$ the CI can be obtained from

$$P\left(\hat{q}_{\alpha/2} \leq \hat{\theta} \leq \hat{q}_{1-\alpha/2}\right) = 1 - \alpha \tag{32}$$

whereas the quantiles $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ are directly obtained from the data resulting in $I = [\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}]$. Such a confidence interval can be used for any statistical hypothesis test.

We would like to emphasize that in contrast to Equation (20) here the quantiles $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ are estimates of the quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ from the sampling distribution. Hence, the obtained CI is merely an approximation.

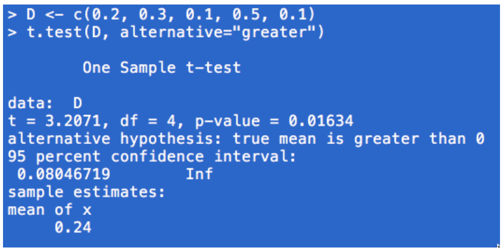
6. An Example and a Warning

Finally, we are providing a practical example for an one-sample *t*-test that will also serve as a warning. In Figure 7 we show a worked example for a data set D defining the major components of a *t*-test.

Data from experiment: D$= \{0.2, 0.3, 0.1, 0.5, 0.1\}$

Main components of an one-sample *t*-test:

1. Select appropriate test statistic T: t-score
2. Define null hypothesis H_0 and alternative hypothesis H_1 ⟵ $\begin{cases} H_0\colon \theta = 0 \\ H_1\colon \theta > 0 \end{cases}$
3. Find the sampling distribution for T, given H_0 true: Student t-distribution
   ⟶ use t.test
4. Choose significance level alpha: $\alpha = 0.05$
5. Evaluate test statistic t for sample data: 3.2071
6. Determine the p-values: 0.01634
7. Make a decision (accept H_0 or reject H_0):
   ⟶ **Reject H_0**



```
> D <- c(0.2, 0.3, 0.1, 0.5, 0.1)
> t.test(D, alternative="greater")

        One Sample t-test

data:  D
t = 3.2071, df = 4, p-value = 0.01634
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.08046719        Inf
sample estimates:
mean of x
     0.24
```

Figure 7. Example for an one-sample *t*-test conducted by using the statistical programming language R. The test can be performed by using the shown data D.

On the left-hand side of Figure 7 a summary of the test is presented and on the right-hand side of Figure 7 we show a script in the programming language R providing the numerical solution to the problem. R is a widespreadly used programming language to study statistical problems [34]. The solution script is only two lines, in the first the data sample is defined and in the second the hypothesis test is conducted. The command 't.test' has arguments that specify the used data and the type of the alternative hypothesis. In our case we are using a right-sided alternative indicated by 'greater'. In addition, the null hypothesis needs to be specified. In our case we used the default which is $\theta = 0$, however, by using the argument 'mu' one can set different values.

From this example one can learn the following. First, the practical execution of a hypothesis test with a computer is very simple. In fact, every hypothesis test assumes a similar form as the provided example. Second, due to the simplicity, all complexity of a hypothesis test, as discussed in the previous sections of this paper, is hidden behind the abstract computer command 't.test'. However, from this follows that a deeper understanding of a hypothesis test cannot be obtained by the practical execution of problems if cast into a black-box frame (in the above example 't.test' is the black-box). The last point maybe counterintuitive if one skips the above discussion, however, we consider this one cause for the widespread misunderstanding of statistical hypothesis tests in general.

7. Historical Notes and Misinterpretations

The modern formulation of statistical hypothesis testing, as discussed in this paper, has not been introduced as one theory but it evolved from two separately introduced theories and accompanied concepts. The first method is due to Fisher [4] and the second due to Neyman and Pearson [8]. Since about the 1960s an unified form was established (some call this null hypothesis significance testing (NHST)) in the literature as it is used to date [35,36].

Briefly, Fisher introduced the concept of a p-values while Neyman and Pearson introduced the alternative hypothesis as complement to the null hypothesis, type I and type II errors and the power. There is an ongoing discussion about the differences of both concepts see, for example, References [37–39], which is in general very difficult to follow because these involve also philosophical interpretations of those theories. Unfortunately, these differences are not only of interest for historical reasons but lead to contaminations and misunderstandings of the modern formulation of statistical hypothesis testing because often arguments are taken out of context and properties differ among the different theories [40,41]. For this reason, we discuss some of those in the following.

1. Is the p-values the probability that the null hypothesis is true given the data?

   No, it is the probability of observing more extreme values than the test statistic, if the null hypothesis is true, that is, $P(x \geq |t| \; | H_0 \text{ is true})$ see Equation (11). Hence, one assumes already that H_0 is true for obtaining the p-values. Instead, the question aims to find $P(H_0|D)$.
2. Is the p-values the probability that the alternative hypothesis is true given the data?

   No, see question (1). This would be $P(H_1|D)$.
3. If the null hypothesis is rejected, is the p-values the probability of your rejection error?

   No, the rejection error is the type I error given by α.
4. Is the p-values the probability to observe our data sample given the null hypothesis is true?

   No, this would be the Likelihood.
5. If one repeats an experiments does one obtain the same p-values?

   No, because p-valuess do not provide information about the long run frequencies of repeated experiments as the type I or type II errors. Instead, they give the probability resulting from comparing the test statistic (as a function of the data) and the null hypothesis assumed to be true.
6. Does the p-values give the probability that the data were produced by random chance alone?

   No, despite the fact that the data were produced by H_0 assuming it is true. The p-values does not provide the probability for this.
7. Does the same p-values from two studies provide the same evidence against the null hypothesis?

   Yes, but only in the very rare case if everything in the two studies and the formulated hypotheses is identical. This includes also the sample sizes. In any other case, p-valuess are difficult to compare with each other and no conclusion can be drawn.

We think that many of the above confusions are a result from *verbal* interpretations of the theory by neglecting *mathematical* definitions of used entities. This is understandable since many people interested in the application of statistical hypothesis testing have not received formal training in the underlying probability theory. A related problem is that a hypothesis test is exactly set-up to answer one question and that is based on a data sample to reject a null hypothesis or not. There are certainly many more questions experimentalists would like to have answers for, however, a statistical hypothesis test is not designed for these. It is only possible to derive some related answers to questions that are closely related to the set-up of the hypothesis test. For this reason in general it is a good strategy to start answering any question in the context of statistical hypothesis testing by looking at the basic definition of the involved entities because only these are exact and provide unaltered interpretations.

8. The Future of Statistical Hypothesis Testing

Despite the fact that the core methodology of statistical hypothesis testing is dating back many decades questions regarding its interpretation and practical usage are to date under discussion [42–46]. This is due to the involvedness and complexity of the methodology demanding a thorough education because otherwise problems are implicated [47] and even unsound designs may be overlooked [12]. Furthermore, there are constantly new statistical hypothesis tests being developed that built upon the standard methodology, for example, by using novel test statistics [48–50]. Given the need to make sense of the increasing flood of data, we are currently facing in all areas of science and industry, statistical hypothesis testing provides a tool for binary decision making. Hence, it allows to convert data into decisions. Due to the need for scientific decision making a future without statistical hypothesis testing is hard to imagine.

9. Conclusions

In this paper we provided a primer on statistical hypothesis testing. Due to the difficulty of the problem, we were aiming at an accessible level of description and presented the bare backbone of the method. We avoided application domain specific formulations in order to make the knowledge transfer easier to different application areas in data science including biomedical science, economics, management, politics, marketing, medicine, psychology or social science [50–55].

Finally, we would like to note that in many practical applications one does not perform one but multiple hypothesis tests simultaneously. For instance, for identifying the differential expression of genes or the significant change of stock prices. In such a situation one needs to apply a multiple testing correction (MTC) for controlling the resulting errors [56–59]. This is a highly non-trivial and a complex topic for itself that can lead to erroneous outcomes if not properly addressed [60].

Author Contributions: F.E.-S. conceived the study. All authors contributed to the writing of the manuscript and approved the final version.

Funding: M.D. thanks the Austrian Science Funds for supporting this work (project P30031).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Helbing, D. The Automation of Society Is Next: How to Survive the Digital Revolution. 2015. Available online: https://ssrn.com/abstract=2694312 (accessed on 1 June 2019) .
2. Hacking, I. *Logic of Statistical Inference*; Cambridge University Press: Cambridge, UK, 2016.
3. Gigerenzer, G. The Superego, the Ego, and the id in Statistical Reasoning. In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, USA, 1993; pp. 311–339.
4. Fisher, R.A. *Statistical Methods for Research Workers*; Genesis Publishing Pvt Ltd.: Guildford, UK, 1925.
5. Fisher, R.A. The Arrangement of Field Experiments (1926). In *Breakthroughs in Statistics*; Springer: Berlin, Germany, 1992; pp. 82–91.
6. Fisher, R.A. The statistical method in psychical research. *Proc. Soc. Psych. Res.* **1929**, *39*, 189–192.
7. Neyman, J.; Pearson, E.S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* **1967**, *20*, 1–2.
8. Neyman, J.; Pearson, E.S. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philos. Trans. R. Soc. Lond.* **1933**, *231*, 289–337. [CrossRef]
9. Lehman, E. *Testing Statistical Hypotheses*; Springer: New York, NY, USA, 2005.
10. Dudoit, S.; Shaffer, J.; Boldrick, J. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* **2003**, *18*, 71–103. [CrossRef]
11. Tripathi, S.; Emmert-Streib, F. Assessment Method for a Power Analysis to Identify Differentially Expressed Pathways. *PLoS ONE* **2012**, *7*, e37510. [CrossRef] [PubMed]

12. Tripathi, S.; Glazko, G.; Emmert-Streib, F. Ensuring the statistical soundness of competitive gene set approaches: Gene filtering and genome-scale coverage are essential. *Nucleic Acids Res.* **2013**, *6*, e53354. [CrossRef] [PubMed]
13. Jiang, Z.; Gentleman, R. Extensions to gene set enrichment. *Bioinformatics* **2007**, *23*, 306–313. [CrossRef] [PubMed]
14. Emmert-Streib, F. The Chronic Fatigue Syndrome: A Comparative Pathway Analysis. *J. Comput. Biol.* **2007**, *14*, 961–972. [CrossRef]
15. Siroker, D.; Koomen, P. *A/B Testing: The Most Powerful Way to Turn Clicks into Customers*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
16. Mauri, L.; Hsieh, W.h.; Massaro, J.M.; Ho, K.K.; D'agostino, R.; Cutlip, D.E. Stent thrombosis in randomized clinical trials of drug-eluting stents. *N. Engl. J. Med.* **2007**, *356*, 1020–1029. [CrossRef] [PubMed]
17. Deuschl, G.; Schade-Brittinger, C.; Krack, P.; Volkmann, J.; Schäfer, H.; Bötzel, K.; Daniels, C.; Deutschländer, A.; Dillmann, U.; Eisner, W.; others. A randomized trial of deep-brain stimulation for Parkinson's disease. *N. Engl. J. Med.* **2006**, *355*, 896–908. [CrossRef] [PubMed]
18. Molina, I.; Gómez i Prat, J.; Salvador, F.; Treviño, B.; Sulleiro, E.; Serre, N.; Pou, D.; Roure, S.; Cabezos, J.; Valerio, L.; et al. Randomized trial of posaconazole and benznidazole for chronic Chagas' disease. *N. Engl. J. Med.* **2014**, *370*, 1899–1908. [CrossRef] [PubMed]
19. Shoptaw, S.; Yang, X.; Rotheram-Fuller, E.J.; Hsieh, Y.C.M.; Kintaudi, P.C.; Charuvastra, V.; Ling, W. Randomized placebo-controlled trial of baclofen for cocaine dependence: Preliminary effects for individuals with chronic patterns of cocaine use. *J. Clin. Psychiatry* **2003**, *64*, 1440–1448. [CrossRef] [PubMed]
20. Sedlmeier, P.; Eberth, J.; Schwarz, M.; Zimmermann, D.; Haarig, F.; Jaeger, S.; Kunze, S. The psychological effects of meditation: A meta-analysis. *Psychol. Bull.* **2012**, *138*, 1139. [CrossRef] [PubMed]
21. Casscells, W.; Schoenberger, A.; Graboys, T.B. Interpretation by Physicians of Clinical Laboratory Results. *N. Engl. J. Med.* **1978**, *299*, 999–1001. [CrossRef] [PubMed]
22. Ioannidis, J.P.A. Why Most Published Research Findings Are False. *PLoS Med.* **2005**, *2*. [CrossRef] [PubMed]
23. Banerjee, I.; Bhadury, T. Self-medication practice among undergraduate medical students in a tertiary care medical college, West Bengal. *Ind. Psychiatry J.* **2009**, *18*, 127–131. [CrossRef]
24. Taroni, F.; Biedermann, A.; Bozza, S. Statistical hypothesis testing and common misinterpretations: Should we abandon *p*-values in forensic science applications? *Forensic Sci. Int.* **2016**, *259*, e32 – e36. doi:10.1016/j.forsciint.2015.11.013. [CrossRef]
25. Emmert-Streib, F.; Dehmer, M. Defining Data Science by a Data-Driven Quantification of the Community. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 235–251. [CrossRef]
26. Sheskin, D.J. *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd ed.; RC Press: Boca Raton, FL, USA, 2004.
27. Chernick, M.R.; LaBudde, R.A. *An Introduction to Bootstrap Methods with Applications to R*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
28. Panagiotou, O.A.; Ioannidis, J.P.; Project, G.W.S. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* **2011**, *41*, 273–286. [CrossRef]
29. Murdoch, D.J.; Tsai, Y.L.; Adcock, J. *p*-valuess are random variables. *Am. Stat.* **2008**, *62*, 242–245. [CrossRef]
30. Emmert-Streib, F.; Moutari, S.; Dehmer, M. A comprehensive survey of error measures for evaluating binary decision making in data science. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, e1303. [CrossRef]
31. Breiman, L. *Statistics: With a View Toward Applications*; Houghton Mifflin Co.: Boston, MA, USA, 1973.
32. Baron, M. *Probability and Statistics for Computer Scientists*; Chapman and Hall/CRC: New York, NY, USA, 2013.
33. Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: New York, NY, USA, 1994.
34. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008; ISBN 3-900051-07-0.
35. Nix, T.W.; Barnette, J.J. The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Res. Sch.* **1998**, *5*, 3–14.
36. Szucs, D.; Ioannidis, J. When null hypothesis significance testing is unsuitable for research: A reassessment. *Front. Hum. Neurosci.* **2017**, *11*, 390. [CrossRef] [PubMed]
37. Biau, D.J.; Jolles, B.M.; Porcher, R. P value and the theory of hypothesis testing: An explanation for new researchers. *Clin. Orthop. Relat. Res.*® **2010**, *468*, 885–892. [CrossRef] [PubMed]

38. Lehmann, E.L. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *J. Am. stat. Assoc.* **1993**, *88*, 1242–1249. [CrossRef]
39. Perezgonzalez, J.D. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front. Psychol.* **2015**, *6*, 223. [CrossRef] [PubMed]
40. Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur. J. Epidemiol.* **2016**, *31*, 337–350. [CrossRef]
41. Goodman, S. A Dirty Dozen: Twelve *p*-values Misconceptions. In *Seminars in Hematology*; Elsevier: Amsterdam, The Netherlands, 2008; Volume 45, pp. 135–140.
42. Wasserstein, R.L.; Lazar, N.A. The ASA's statement on *p*-valuess: Context, process, and purpose. *Am. Stat.* **2016**, *70*, 129–133. [CrossRef]
43. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a World Beyond $p < 0.05$. *Am. Stat.* **2019**, *73*, 1–19. [CrossRef]
44. Ioannidis, J. Retiring significance: A free pass to bias. *Nature* **2019**, *567*, 461–461. [CrossRef]
45. Amrhein, V.; Greenland, S.; McShane, B. Scientists rise up against statistical significance. *Nature* **2019**, *567*, 305. [CrossRef]
46. Benjamin, D.J.; Berger, J.O. Three Recommendations for Improving the Use of *p*-valuess. *Am. Stat.* **2019**, *73*, 186–191. [CrossRef]
47. Gigerenzer, G.; Gaissmaier, W.; Kurz-Milcke, E.; Schwartz, L.M.; Woloshin, S. Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* **2007**, *8*, 53–96. [CrossRef] [PubMed]
48. Rahmatallah, Y.; Emmert-Streib, F.; Glazko, G. Gene Sets Net Correlations Analysis (GSNCA): A multivariate differential coexpression test for gene sets. *Bioinformatics* **2014**, *30*, 360–368. [CrossRef] [PubMed]
49. De Matos Simoes, R.; Emmert-Streib, F. Bagging statistical network inference from large-scale gene expression data. *PLoS ONE* **2012**, *7*, e33624. [CrossRef] [PubMed]
50. Rahmatallah, Y.; Zybailov, B.; Emmert-Streib, F.; Glazko, G. GSAR: Bioconductor package for Gene Set analysis in R. *BMC Bioinform.* **2017**, *18*, 61. [CrossRef]
51. Cortina, J.M.; Dunlap, W.P. On the logic and purpose of significance testing. *Psychol. Methods* **1997**, *2*, 161. [CrossRef]
52. Hubbard, R.; Parsa, R.A.; Luthy, M.R. The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917–1994. *Theory Psychol.* **1997**, *7*, 545–554. [CrossRef]
53. Emmert-Streib, F.; Dehmer. A Machine Learning Perspective on Personalized Medicine: An Automatized, Comprehensive Knowledge Base with Ontology for Pattern Recognition. *Mach. Learn. Knowl. Extr.* **2018**, *1*, 149–156. [CrossRef]
54. Nickerson, R.S. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychol. Methods* **2000**, *5*, 241. [CrossRef]
55. Sawyer, A.G.; Peter, J.P. The significance of statistical significance tests in marketing research. *J. Mark. Res.* **1983**, *20*, 122–133. [CrossRef]
56. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 125–133. [CrossRef]
57. Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*; Cambridge University Press: Cambridge, UK, 2010.
58. Emmert-Streib, F.; Dehmer, M. Large-Scale Simultaneous Inference with Hypothesis Testing: Multiple Testing Procedures in Practice. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 653–683. [CrossRef]
59. Farcomeni, A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* **2008**, *17*, 347–88. [CrossRef] [PubMed]
60. Bennett, C.M.; Baird, A.A.; Miller, M.B.; Wolford, G.L. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for proper multiple comparisons correction. *J. Serendipitous Unexpect. Results* **2011**, *1*, 1–5. [CrossRef]