OPEN ACCESS

This is the English version.
The German version starts at p. 8.

Review Article

Fisher's significance test: A gentle introduction

Abstract

The p-value is often misunderstood and, for example, misinterpreted as a probability for the correctness of the null hypothesis. The aim of this article is to first explain the definition of the p-value. Determining the p-value requires knowledge of a probability function. How an appropriate statistical model is selected and how the p-value is determined using this model, the null hypothesis and the empirical data is explained using the t-distribution. When interpreting the p-value obtained in this way, two incompatible statistical schools of thought are confronted: the orthodox Neyman-Pearson hypothesis test, which amounts to a decision between the null hypothesis and a complementary alternative hypothesis, and Fisher's significance test, in which no alternative hypothesis is formulated and in which the smaller the p-value, the greater the evidence against the null hypothesis. The amount ends with some critical remarks about the handling of p-values.

Keywords: statistical models, statistical data interpretation, data analysis

Andreas Stang[1,2]
Bernd Kowall[1]

1 Institute of Medical Informatics, Biometry and Epidemiology; University Hospital of Essen, Germany

2 School of Public Health, Department of Epidemiology, Boston University, Boston, United States

Introduction

The p-value is often misunderstood and, for example, misinterpreted as a probability for the correctness of the null hypothesis. P-values play an important role in two schools of thought: Fisher's significance test and Neyman and Pearson's hypothesis test [1], [2]. While the significance test leads to a quantitative interpretation of the p-value, in which it is interpreted as a continuous measure of evidence against the null hypothesis, the p-value in the null hypothesis test merely serves a decision using predefined rules.

In 2016, the American Statistical Association (ASA) published a statement on the handling of p-values. Among other things it was stated: "The widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process" [3]. In 2019 Amrhein et al. published an article entitled "Retire statistical significance" in Nature in which they draw attention to the many pitfalls in the dichotomization of p-values into "significant" (usually $p \leq 0.05$) and "non-significant" (usually $p > 0.05$) and generally discourage this dichotomization of p-values, i.e. the categorization into two areas [4].

A dilemma in the application of the significance or hypothesis test remains the lack of understanding of what these methods can answer at all. The aim of this paper is to illustrate essential background information and the steps of the significance test by means of a fictive study in which two groups are compared with each other. Most biostatistics textbooks do not consistently provide this background information and steps of the significance test. The article is intended for people who can only vaguely describe what the procedure does.

Fundamental statistical concepts – standard deviation, sampling error, and standard error

Basic understanding – random sampling from a target population (population model)

The target population of a scientific question represents the totality of all observation units. If the target population is the resident population of the FRG, the total population in 2016 is 82.5 million. Interesting variables of this population could be mean values and scatters of characteristics (e.g. mean sleep latency, i.e. the average time from switching off the light in the bedroom to falling asleep in minutes). These characteristics of variables of the target population, which are usually unknown to us, are abbreviated with Greek letters in the sense of a statistical convention. For example, the Greek letter μ and σ are used for the mean value and the variance of a variable in the target population.

When conducting empirical studies, it is generally not possible to examine the whole target population. For this reason, only a sample from the target population is examined and information from the sample is used to make statements about the target population. The statistical inference of a sample to a target population represents an inductive conclusion and is referred to in statistics as inferential statistics.

When random samples are drawn from a target population, the so-called sampling error (sampling variability) occurs. Since only a part of the target population is examined, there is variability from sample to sample. This

can easily be illustrated by the toss of a fair coin. One would expect that 50% of all tosses would show head. This expected value, also called probability, is the prognosis of a relative frequency. If the coin were flipped 10 times, head could appear 4 times. Flipping the coin 10 times again would not necessarily result in 4 times head, but e.g. 6 times head. This variability is an expression of the sampling error. Thus there can be no certain conclusion from a sample to a target population. The law of large numbers states that with increasing study size the sampling error becomes smaller and smaller.

Variability versus uncertainty

If, for example, one undertakes a study on the basis of a sample of 30 adult women with sleep disorders aged 55–64 living in Germany with the aim of estimating the true mean value μ of the sleep latency of the target population, the sample provides a mean value $\bar{x}$ of e.g. 38 min and a corresponding empirical variance s^2, which is calculated according to the following formula:

$$Var(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Assuming a normal distribution of the variable sleep latency, a suitable statistical measure describing the variability in the sample would be the standard deviation (SD), which is the square root of the variance, in addition to the variance. The standard deviation s for the sample would be 8.5 min. If this study were repeated, in which a random sample of 30 adult women with sleep disorders aged 55–64, resident in Germany, is again obtained, the mean value would be for example 33 min and the standard deviation would be for example 8.4 min. The standard error of the mean (SE) is not a measure that quantifies the variability of the measured values within the sample, but rather the uncertainty of the estimate of the mean μ of the target population [5]. The standard error is calculated according to the following formula:

$$SE = \frac{s}{\sqrt{n}}$$

where *n* is the number of observations. It can be seen that the smaller the variability of the characteristic in the sample and the larger the sample, the smaller the SE becomes.

How does a statistical test work – the t-test as an example

Two-group comparison

In an example of two randomly sampled groups, we compare the effect of a new sleeping pill on sleep latency. The verum group includes 32 persons, the placebo group 30 persons (cf. Table 1). In both groups, sleep latency was determined after 7 days of treatment in the sleep laboratory (polysomnography). The null hypothesis is that the two groups do not differ with regard to sleep latency. Several tests have been suggested for such a group comparison.

Table 1: Results of the study on the new sleep pill to reduce sleep latency

	Placebo	Verum
Number of patients, n	30	32
$\bar{x}$ (min)	38	33
s (min)	8.5	8.4
SE (min)	1.6	1.5

$\bar{x}$ (min): average duration in min; s: standard deviation; SE: standard error of the mean

In Table 2, we briefly explain the permutation test that is historically important. The permutation test is rarely used nowadays because the computing effort may be huge. In our example, there are 4.5 times 10^{17} permutations. Therefore, in our case the t-test would be preferred which can be regarded as a good approximation of the permutation test and is most popular in the biomedical literature. A comparison of the mean values of the two samples shows that the mean sleep latency in the verum group is 5 min lower than in the placebo group. In both groups, sleep latency varied, as can be seen from the standard deviations. Both samples are associated with random error due to sampling error.

The question that arises here is whether the difference of 5 min is only an expression of a random error or whether this difference is an expression of an actual effect of the sleeping pill. In the first case, both samples would come from identical populations ($\mu_p=\mu_v$), in the second case, the two samples would come from different populations, i.e., populations with $\mu_p \neq \mu_v$. Figure 1 illustrates the problem: could it be that placebo and verum do not differ with respect to the true sleep latency averages, i.e. come from the same population with e.g. μ=38 min, and the two sample averages (33 min and 38 min) are merely an expression of the sampling error, similar to the coin toss of a fair coin? Or could it be that the new sleep pill actually has an effect on sleep latency so that the true mean values come from target populations with different mean values ($\mu_p \neq \mu_v$)?

Expectation of statistical variability of study results due to random error

A significance test can provide some, albeit imperfect, information on these central questions. To answer the above questions, the behavior of the mean difference due to the random error must first be determined, assuming that a null hypothesis H_0 were true. There is an infinite set of null hypotheses. In medicine, the nil hypothesis has prevailed, i.e. the null hypothesis of no association between treatment assignment (placebo or verum) and sleep latency (i.e. $\mu_p=\mu_v$). The Greek letters indicate that this null hypothesis refers to the target population. Under this hypothesis, mean differences that are not equal to

Table 2: Permutation test

Permutation test

- The 62 sleep latency values are arranged in all sequences (permutations) that are possible. For n=62 values to be assigned to two groups of 32 and 30 patients, there are overall $4.5 * 10^{17}$ possibilities to do that.
- For each permutation the group assignment is defined: the first 32 values are assigned to the verum group and the remaining 30 values to the placebo group.
- After each group assignment, the test statistics of interest, here the difference of the mean values, are calculated.
- Finally, the p-value is determined by dividing the number of test statistics for which the mean difference is ≥5 min (this was the observed mean difference of the study) by the number of all test statistics.

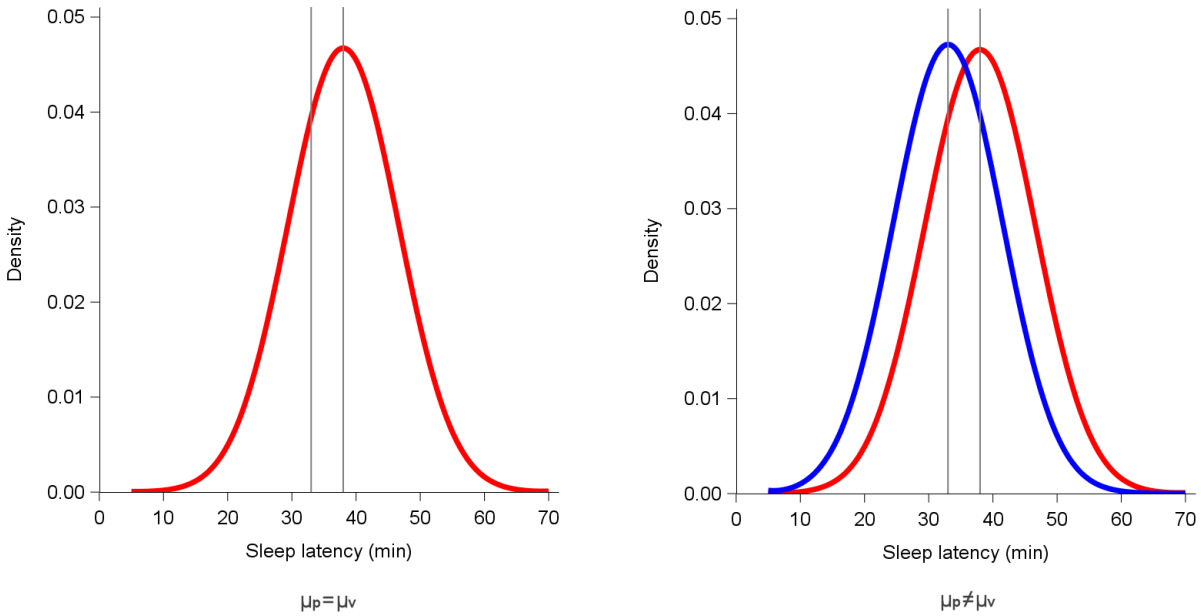

Figure 1: Normal distributions of the sleep latency in the target populations

zero are an expression of the random error. Similar to how extreme outcomes of experiments are rarely observed when tossing a fair coin (e.g. 10 tosses and it appears 10 times head), the difference of the means rarely takes extreme values under the null hypothesis.

But how many permuted arrangements of patients split into two groups do exist and how would differences of the means in these arrangements behave if the null hypothesis $\mu_p=\mu_v$ were true? The difficulty in answering this question lies in the fact that the behavior of the difference of the means under the null hypothesis depends on the variability of the sleep latency within the samples and the size of the samples.

So in order to predict how the differences of the means would behave if the null hypothesis were true, one has to take these two influencing variables into account. Here a kind of normalization is helpful, which will be illustrated by the following example. A difference of means of 3 seconds is observed for two groups of marathon runners (2 hours, 3 min, 40 seconds versus 2 hours, 3 min, 43 seconds) and for two groups of 400 meters runners (46 seconds versus 49 seconds). For similar groups of runners, the differences of 3 seconds have a different meaning. For marathon runners, the difference is very small in relation to the average total duration of the run, while it is relatively larger for 400 meters runners. The relation to the average running time is a kind of normalization. The choice of statistical test, which ensures such standardization, determines which test statistics is chosen. If, for example, the t-test is selected for independent samples, the corresponding test variable is the t-statistic, for the Chi-square test it is the Chi-square-statistic etc. The choice of the appropriate statistical test again depends on criteria, which are briefly explained in Table 3.

The t-statistic is defined as:

$$t = \frac{\text{Observed difference of means} - \text{Expexted difference of means}}{\text{Standard error of the observed difference of means}}$$

The expected difference of means in the t-statistic formula is the value assumed under the null hypothesis H_0. In the case of the nil hypothesis $\mu_p=\mu_v$ a difference of zero minutes is expected. This simplifies the t-statistics:

$$t = \frac{\text{Observed difference of means}}{\text{Standard error of the observed difference of means}}$$

Table 3: Criteria for test selection

The t-test for independent samples mentioned in the text is used to compare the mean values of two samples and is performed when several conditions are met: The variable for which the comparison is made must be interval scaled and normally distributed in both samples. The requirement of independence of the samples would not be met, in particular in the case of repeated measurements on the same sample.
Similarly as for other statistical tests, the model assumptions should be examined. In general, the following criteria play an important role in test selection:

1. The distribution of variables: so-called parametric tests, unlike non-parametric tests, require a certain distribution of variables. The t-test is therefore one of the parametric tests.
2. The dependence or independence of the samples: dependence of samples is mainly present in the case of repeated measurements. If measurements are made on the same sample before and after an intervention, the two measurement series are dependent.
3. The number of samples: there are tests with only one sample − for example, if one examines whether the IQ of a study population deviates from 100. If two or more groups are compared in terms of IQ, a two-sample or multi-sample test is required.

In the case of unequal variances, the standard error of the difference of the means is calculated according to the following formula:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

with
n_1: number of patients in group 1 (placebo)
n_2: number of patients in group 2 (verum)
s_1^2: variances of sleep latency in group 1
s_2^2: variances of sleep latency in group

The formula changes if the variances are equal (formula not shown). The standard error of the difference of the means depends on the variances of the variable (sleep latency) and the group sizes of the groups being compared. After determining the standard error, the t-statistic for two independent samples with unequal variances is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Independence means that the two patient groups are independent of each other and also that patients within the groups are independent of each other. For example, independence is violated if the outcome of a patient would contribute statistically to both patient groups. Similarly, independence would be violated if patients in the same group influenced each other in terms of outcomes of interest. Independence is also violated when a characteristic is collected from a group of patients several times over time (e.g. before and after treatment). The data of the sleep study now have the following t-value:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} = \frac{38 - 33}{\sqrt{\frac{8.5^2}{30} + \frac{8.4^2}{32}}} = 2.33$$

The t-value for the concrete study is therefore +2.33. This distribution can be determined by using the so-called degrees of freedom (df). The number of degrees of freedom is the number of values that can be freely varied without changing the mean values. If, for example, there are three numbers k, l and m and their sum is 100, it is clear that if two of the three numbers are known, the third number is automatically given. If k=20 and l=70, m must be 10. With 62 patients in the study one has $n_1-1+n_2-1=30-1+32-1=60$ degrees of freedom. If 60 values were freely selected, then one has no further choice for the last two observations.
With the help of the 60 degrees of freedom, the appropriate distribution can now be displayed under the assumption of the null hypothesis. The illustration of the formula for creating the t-distribution is omitted for didactic reasons (it is the ratio of the standard normal variable z and the square root of a chi-square value with n degrees of freedom divided by n). The t-distribution is symmetrical and bell-shaped like the normal distribution (Figure 2).

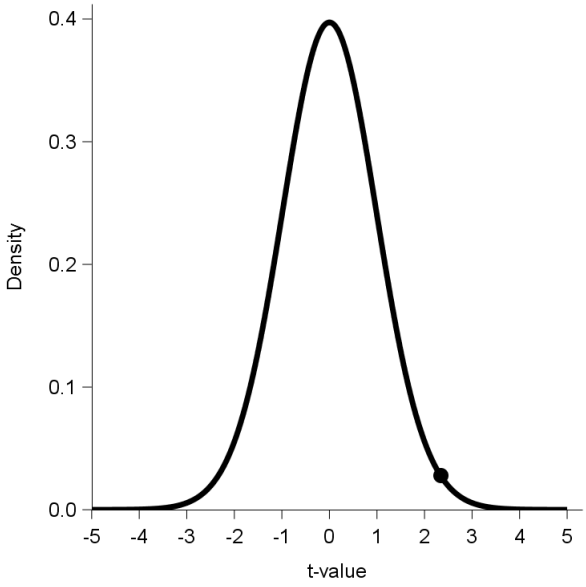

Figure 2: t-distribution with 60 degrees of freedom and marked result of the concrete study (t=2.33)

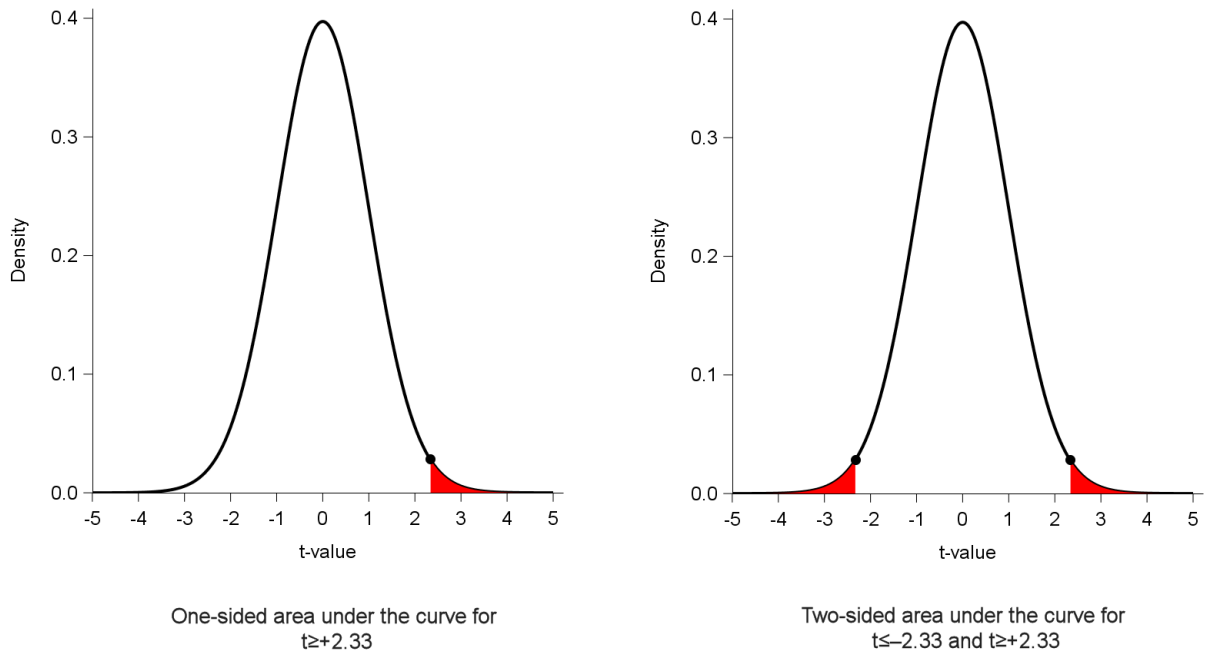

Figure 3: t-distribution with 60 degrees of freedom with marked areas under the curve for t≥+2.33 and t≤–2.33

The probability density function (PDF) supplies so-called density values depending on the t-values. In contrast to probabilities, which can only assume values between 0 and 1, densities can also assume values >1.

Interpretation of the t-value

A single density value of the PDF has no practical interpretation. The total area under the curve of the PDF is 1 so that (partial) areas under the probability density function have the interpretation of probabilities. In the context of the study, it is now possible to answer the question of how high the probability is that the t value assumes ≥+2.33 under the null hypothesis ($\mu_p=\mu_v$), i.e. t=0.

The cumulative distribution function (CDF) returns the probability that a t-value is smaller than or equal to a concrete value t_k. It is also possible to use the CDF to calculate the probability that t becomes $\geq t_k$ by subtracting the probability for t values $<t_k$ from the value of one. The formula for this function is omitted at this point, but can easily be found on the Internet [6]. In the case of the sleep study, t_k=+2.33. Figure 3 shows the area under the curve for t≥+2.33 for a one-sided view and for the areas under the curve for t≤–2.33 and t≥+2.33, a two-sided view.

The one-sided area has an amount of 0.01. This means that the probability that studies under the assumption of the null hypothesis ($\mu_p=\mu_v$) generate a t value of ≥+2.33 is 1%. On a two-sided basis, the probability that studies assuming the null hypothesis ($\mu_p=\mu_v$) generate a t value of ≤–2.33 or ≥+2.33 is 2%. The probability of 1% corresponds to the one-sided p-value, while the probability of 2% corresponds to the two-sided p-value.

The p-value – explanation and some caveats

Interpretation of the p-value

The p-value thus provides the probability (criterion 1) under a null hypothesis (criterion 2) of finding a result such as the present study result or observing study results that deviate even more from the null hypothesis (criterion 3). All three criteria are necessary criteria for the definition of the p-value.

It is important to note here that the p-value makes a statement about the behavior of a test statistic in presence of random error given the null hypothesis. At a p-value of 0.01, only 1% of the studies would generate a t-value of ≥+2.33 if the null hypothesis were true. Thus, the p-value also makes a statement about the outcomes of studies that were not observed (counterfactual element). Furthermore, it must be emphasized that the p-value was calculated under a condition: the condition that the null hypothesis H_0 were true, which is why the p-value is also referred to as a conditional probability. The null hypothesis was merely assumed, regardless of how large the truth content of this hypothesis is.

Fisher interpreted the p-value as a continuous measure of evidence against the null hypothesis. He said: "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances he rejects hypotheses; he rather gives mind to each particular case in the light of his evidence and his ideas" [7]. This means that, according to Fisher's school, the classification of a p-value is context-dependent and the application of a fixed threshold of typically 0.05 is not justified. The orthodox rejection of a null hypothesis at a pre-defined threshold of 0.05 comes from the competing school of

Neyman and Pearson, who introduced the hypothesis test as a decision-theoretical procedure.

What does a large p-value of e.g. 0.70 mean? Technically speaking, it means that the probability is 70% of the observed study outcome or of study outcomes deviating even more from the null hypothesis under the assumption of the null hypothesis. In practice, this means that the significance test provided little evidence against the tested null hypothesis or statistical model. However, it does not mean that the null hypothesis is true. The p-value is a function of the strength of effect (e.g. observed mean difference, here 5 min) and the study size (here 62 women). With a large p-value, a strong effect can actually be present, but the study size was very small. Typical errors in the definition of p-values are discussed below.

"The p-value is the probability that the null hypothesis is true." The p-value does not provide a statement about the probability of the truth of the null hypothesis, but the p-value was calculated under the assumption that the null hypothesis was true. Incidentally, the reference to even more extreme outcomes of the study (counterfactual element) is missing here.

"The p-value is the probability of type I error." This statement is incorrect because it mixes principles of the significance test (Fisher) with those of the hypothesis test (Neyman & Pearson). According to the school of Fisher, there is no a priori fixed level of significance (also called type I error). In contrast, according to Neyman & Pearson, the level of significance, called type I error, is fixed before the study started whereas the p-value is derived from the statistical model and the study data after the study has been done. According to Neyman & Pearson, the type I error remains as it is after the end of the study and the p-value is compared to the a priori fixed type I error for making a decision.

The type I error, also called α error, is determined according to Neyman and Pearson before the beginning of the study. At the end of the study, the p-value which is obtained from the null hypothesis, the statistical model (e.g. t-test) and the study data is compared with the α (most often 0.05). The statement that "a low p-value excludes chance as an explanation for an observed difference" proves a gross lack of understanding.

Almost correct sounding definitions of the p-value are for example: "The p-value is the probability to observe the present study result or even more extreme study results." In this definition, the central condition (criterion 2) of the p-value is missing: the calculation takes place under the assumption that the null hypothesis were true. The following incorrect definition is also popular: "The p-value is the probability of observing the present study result under the null hypothesis." Here criterion 3 is missing: the p-value also makes a statement about unobserved study results that deviate even more from the null hypothesis than the present study result.

In the significance test according to Fisher, there is no so-called type I error and type II error, there is no confidence interval, no alternative hypothesis and no concept for statistical power or sample size calculations. These phenomena originate from Neyman & Pearson and only become relevant when performing hypothesis tests, which are decision-theoretically only valid if all steps of the hypothesis test procedure are adhered to, which is why authors also speak of Neyman-Pearson orthodoxy [8]:

1. Definition of the null and alternative hypothesis before the start of the study.
2. Determination of type I and type II error before the start of the study.
3. Determination of test statistics before the start of the study.
4. Calculation of the required sample size before the start of the study.
5. Conduct the study in compliance with the required sample size
6. Calculation of the test statistics and comparison with a critical value of the test statistics or comparison of the p-value with the specified type I error (after the study).
7. Decision: if $p \leq \alpha$, the null hypothesis is rejected, if $p > \alpha$, the null hypothesis is not rejected (after the study).

If steps 1–7 are not complied with, the decision-theoretical procedure of hypothesis testing loses its validity. The decision (7^{th} step) must be consistently applied. If, for example, $\alpha=0.05$ was specified and $p=0.07$ came out at the end of the study, then according to Neyman & Pearson it cannot be said that there was a "significance trend" or something similar, but only that the null hypothesis was not rejected. Likewise p-values ≤ 0.05 are not sub-categorized into e.g. $p \leq 0.05$*, $p \leq 0.01$** and $p \leq 0.001$*** according to Neyman & Pearson.

Conditions necessary for the correct interpretation of the p-value

Many introductory textbooks of biostatistics merely introduce the theory of significance testing. This means that there are no sources of error other than random error. In the practice of empirical studies, however, this is an unrealistic assumption. Greenland et al. [9] rightly point out that in the case of a low p-value only a signal is given that something may be wrong with the so-called statistical model. The statistical model consists of three components: the chosen test statistics, the chosen null hypothesis and the empirical study data.

In addition to the hypothesis that the low p-value represents evidence against the null hypothesis, the following alternative explanations need to be considered, all of which are related to the statistical model and thus influence the p-value:

- An unsuitable test statistic was applied.
- Selection bias into the study or selection bias during follow-up of study subjects occurred.
- The comparison between two samples is confounded (mixing of effects).

- There is information bias in the measurement of the variables in the study.

If the p-value is low, we can only conclude that something is wrong with the statistical model. However, the p-value itself does not show what is wrong with the model. The inexperienced user of the significance test thinks of a low p-value only as an indication that the null hypothesis might be wrong. In addition to the contextual dependence of the meaning of low p-values explained by Fisher, the result of a significance test must always be seen in the light of the complete statistical model.

Summary

Fisher's significance test is a different procedure than the Neyman & Pearson hypothesis test, which is often ignored. While the significance test produces a p-value, which according to Fisher should be interpreted context-dependently as a continuous measure of evidence against the null hypothesis, the p-value serves as a decision criterion if the necessary steps of the hypothesis test are followed. The significance test leads to the p-value, whose definition must contain three criteria: probability, the use of the null hypothesis assumption, and the counterfactual element of the p-value. P-values can be small for various reasons and the evidence against the null hypothesis is one of several competing reasons in empirical studies.

Notes

Competing interests

The authors declare that they have no competing interests.

References

1. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. The empire of chance. How probability changed science and everyday life. Cambridge: Cambridge University Press; 1989.
2. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. PeerJ Preprints. 2018;6:e26857v4. DOI: 10.7287/peerj.preprints.26857v3
3. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70:129-33. DOI: 10.1080/00031305.2016.1154108
4. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305-307. DOI: 10.1038/d41586-019-00857-9
5. Cox DR. Principles of statistical inference. Cambridge: Cambridge University Press; 2006. DOI: 10.1017/CBO9780511813559
6. Student's t-distribution. In: Wikipedia. [accessed 2019 May 16]. Available from: https://en.wikipedia.org/wiki/Student%27s_t-distribution
7. Fisher RA. Statistical methods and scientific inference. Edinburgh: Oliver & Boyd; 1956.
8. Oakes MW. Statistical inference. Chichester: Wiley; 1986.
9. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016 Apr;31(4):337-50. DOI: 10.1007/s10654-016-0149-3
10. Manly BFJ. Randomization, bootstrap and Monte Carlo methods in biology. London: Chapman & Hall; 1996. Randomization; p. 3-7.
11. Feinstein AR. Principles of medical statistics. Boca Raton: Chapman & Hall/CRC; 2002. Testing stochastic hypotheses; p. 190-1.

Corresponding author:
Prof. Dr. med. Andreas Stang, MPH
Center of Clinical Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Hufelandstr. 55, 45147 Essen, Germany, Phone: +49 201-723-77-289, Fax: +49 201-723-77-333
andreas.stang@uk-essen.de

Please cite as
Stang A, Kowall B. Fisher's significance test: A gentle introduction. GMS Med Inform Biom Epidemiol. 2020;16(1):Doc03.
DOI: 10.3205/mibe000206, URN: urn:nbn:de:0183-mibe0002065

This article is freely available from
https://www.egms.de/en/journals/mibe/2020-16/mibe000206.shtml

Published: *2020-05-11*

Fishers Signifikanztest: Eine sanfte Einführung

Zusammenfassung

Der p-Wert wird häufig missverstanden und beispielsweise als Wahrscheinlichkeit für die Richtigkeit der Nullhypothese fehlinterpretiert. Ziel des vorliegenden Beitrags ist es, zunächst die Definition des p-Werts zu erläutern. Die Ermittlung des p-Werts erfordert die Kenntnis einer Wahrscheinlichkeitsfunktion. Wie ein geeignetes statistisches Modell ausgewählt wird und anhand dieses Modells, der Nullhypothese und der empirischen Daten der p-Wert bestimmt wird, wird an der t-Verteilung erklärt. Bei der Interpretation des so erhaltenen p-Werts stehen sich zwei nicht vereinbare statistische Denkschulen gegenüber: Der orthodoxe Neyman-Pearson Hypothesentest, der auf eine Entscheidung zwischen der Nullhypothese und einer komplementären Alternativhypothese hinausläuft, und Fishers Signifikanztest, bei dem keine Alternativhypothese formuliert wird und in der die Evidenz gegen die Nullhypothese umso größer ist, je kleiner der p-Wert ist. Der Beitrag endet mit einigen kritischen Bemerkungen zum Umgang mit p-Werten.

Andreas Stang[1,2]
Bernd Kowall[1]

1 Institut für Medizinische Informatik, Biometrie und Epidemiologie, Universitätsklinikum Essen, Deutschland

2 School of Public Health, Department of Epidemiology, Boston University, Boston, Vereinigte Staaten

Einleitung

Der p-Wert wird oft missverstanden und z.B. als Wahrscheinlichkeit für die Richtigkeit der Nullhypothese missinterpretiert. P-Werte spielen in zwei Denkschulen eine wichtige Rolle: Dem Signifikanztest nach Fisher und dem Hypothesentest nach Neyman und Pearson [1], [2]. Während der Signifikanztest zu einer quantitativen Interpretation des p-Wertes führt, in der er als ein kontinuierliches Maß für die Evidenz gegen die Nullhypothese interpretiert wird, dient der p-Wert im Nullhypothesentest lediglich einer Entscheidung anhand vordefinierter Regeln. Im Jahr 2016 veröffentlichte die American Statistical Association (ASA) eine Erklärung über die Handhabung von p-Werten. Darin wurde unter anderem erklärt: „Die weit verbreitete Verwendung von statistischer Signifikanz' (im Allgemeinen als $p \leq 0{,}05$ interpretiert) als Lizenz für die Behauptung eines wissenschaftlichen Befundes (oder einer impliziten Wahrheit) führt zu einer erheblichen Verzerrung des wissenschaftlichen Prozesses" [3]. Im Jahr 2019 veröffentlichten Amrhein et al. in der Fachzeitschrift Nature einen Artikel mit dem Titel „Retire statistical significance", in dem sie auf die vielen Fallstricke bei der Dichotomisierung von p-Werten in „signifikant" (üblicherweise $p \leq 0{,}05$) und „nicht-signifikant" (üblicherweise $p > 0{,}05$) aufmerksam machen und generell von dieser Dichotomisierung von p-Werten, d.h. der Einteilung in zwei Bereiche, abraten [4].

Ein Dilemma bei der Anwendung des Signifikanz- oder Hypothesentests bleibt das mangelnde Verständnis dafür, was diese Methoden überhaupt beantworten können. Das Ziel dieser Arbeit ist es, wesentliche Hintergrundinformationen und die Schritte des Signifikanztests anhand einer fiktiven Studie zu veranschaulichen, in der zwei Gruppen miteinander verglichen werden. Die meisten Biostatistik-Lehrbücher liefern diese Hintergrundinformationen und die Schritte des Signifikanztests nicht konsistent. Der Artikel richtet sich an Personen, die nur vage beschreiben können, was das Verfahren bewirkt.

Statistische Grundbegriffe – Standardabweichung, Stichprobenfehler und Standardfehler

Grundlegendes Verständnis – Zufallsstichproben aus einer Zielpopulation (Bevölkerungsmodell)

Die Zielpopulation einer wissenschaftlichen Frage stellt die Gesamtheit aller Beobachtungseinheiten dar. Wenn die Zielpopulation die Wohnbevölkerung der BRD ist, beträgt die Gesamtbevölkerung im Jahr 2016 82,5 Millionen. Interessante Variablen dieser Grundgesamtheit könnten Mittelwerte und Streuungen von Merkmalen sein (z.B. die mittlere Schlaflatenz, d.h. die durchschnittliche Zeit vom Ausschalten des Lichts im Schlafzimmer bis zum Einschlafen in Minuten). Diese Merkmale von Variablen der Zielpopulation, die uns in der Regel unbekannt sind, werden im Sinne einer statistischen Konvention mit griechischen Buchstaben abgekürzt. Beispielsweise werden die griechischen Buchstaben μ und s für den Mittelwert und die Varianz einer Variablen der Zielpopulation verwendet.

Bei der Durchführung empirischer Studien ist es im Allgemeinen nicht möglich, die gesamte Zielpopulation zu untersuchen. Aus diesem Grund wird nur eine Stichprobe aus der Zielpopulation untersucht und die Informationen aus der Stichprobe werden verwendet, um Aussagen über die Zielpopulation zu treffen. Der statistische Rückschluss einer Stichprobe auf eine Zielpopulation stellt eine induktive Schlussfolgerung dar und wird in der Statistik als Inferenzstatistik bezeichnet.

Wenn aus einer Zielpopulation Zufallsstichproben gezogen werden, tritt der so genannte Stichprobenfehler (Stichprobenvariabilität) auf. Da nur ein Teil der Zielpopulation untersucht wird, gibt es eine Variabilität von Stichprobe zu Stichprobe. Dies kann leicht durch den Wurf einer ungezinkten Münze veranschaulicht werden. Man würde erwarten, dass 50% aller Würfe Kopf zeigen würden. Dieser Erwartungswert, auch Wahrscheinlichkeit genannt, ist die Prognose einer relativen Häufigkeit. Wenn die Münze 10-mal geworfen würde, könnte Kopf 4-mal erscheinen. Würde man die Münze noch einmal 10-mal werfen, so würde nicht unbedingt Kopf 4-mal, sondern z.B. 6-mal auftreten. Diese Variabilität ist Ausdruck des Stichprobenfehlers. Es kann also keine sichere Schlussfolgerung aus einer Stichprobe auf eine Zielpopulation gezogen werden. Das Gesetz der großen Zahlen besagt, dass mit zunehmender Studiengröße der Stichprobenfehler immer kleiner wird.

Variabilität versus Unsicherheit

Führt man z.B. eine Studie auf der Basis einer Stichprobe von 30 erwachsenen Frauen mit Schlafstörungen im Alter von 55–64 Jahren, die in Deutschland leben, durch, um den wahren Mittelwert μ der Schlaflatenz der Zielpopulation abzuschätzen, so liefert die Stichprobe einen Mittelwert $\bar{x}$ von z.B. 38 min und eine entsprechende empirische Varianz s^2, die nach folgender Formel berechnet wird:

$$Var(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Unter der Annahme einer Normalverteilung der Variable Schlaflatenz wäre ein geeignetes statistisches Maß, das die Variabilität in der Stichprobe beschreibt, neben der Varianz die Standardabweichung (SD), die die Quadratwurzel der Varianz ist. Die Standardabweichung s für die Stichprobe würde 8,5 min betragen. Würde diese Studie wiederholt, bei der wiederum eine Zufallsstichprobe von 30 erwachsenen Frauen mit Schlafstörungen im Alter von 55–64 Jahren, die in Deutschland wohnen, gewonnen wird, so würde der Mittelwert z.B. 33 min und die Standardabweichung z.B. 8,4 min betragen. Der Standardfehler des Mittelwertes (SE) ist kein Maß, das die Variabilität der Messwerte innerhalb der Stichprobe quantifiziert, sondern vielmehr die Unsicherheit der Schätzung des Mittelwertes μ der Zielpopulation [5]. Der Standardfehler wird nach der folgenden Formel berechnet:

$$SE = \frac{s}{\sqrt{n}}$$

wobei n die Anzahl der Beobachtungen ist. Es ist zu erkennen, dass der Standardfehler umso kleiner wird, je kleiner die Variabilität des Merkmals in der Stichprobe und je größer die Stichprobe ist.

Wie funktioniert ein statistischer Test – der t-Test als Beispiel

Zwei-Gruppen-Vergleich

In einem Beispiel von zwei zufällig ausgewählten Gruppen vergleichen wir die Wirkung eines neuen Schlafmittels auf die Schlaflatenz. Die Verumgruppe umfasst 32 Personen, die Placebogruppe 30 Personen (vgl. Tabelle 1). In beiden Gruppen wurde die Schlaflatenz nach 7 Tagen Behandlung im Schlaflabor (Polysomnographie) bestimmt. Die Nullhypothese ist, dass sich die beiden Gruppen hinsichtlich der Schlaflatenz nicht unterscheiden. Es wurden mehrere Tests für einen solchen Gruppenvergleich vorgeschlagen.

Tabelle 1: Ergebnisse der Studie zum Einfluss eines neuen Schlafmedikaments auf die Schlaflatenz

	Placebo	Verum
Anzahl Patienten, n	30	32
$\bar{x}$ (min)	38	33
s (min)	8,5	8,4
SE (min)	1,6	1,5

$\bar{x}$ (min): durchschnittliche Dauer in Minuten; s: Standardabweichung; SE: Standardfehler des Mittelwerts

In Tabelle 2 erläutern wir kurz den Permutationstest, der historisch wichtig ist. Der Permutationstest wird heutzutage nur noch selten verwendet, da der Rechenaufwand sehr groß sein kann. In unserem Beispiel gibt es 4,5 mal 10^{17} Permutationen. Daher wäre in unserem Fall der t-Test zu bevorzugen, der als gute Annäherung an den Permutationstest angesehen werden kann und in der biomedizinischen Literatur am beliebtesten ist.

Ein Vergleich der Mittelwerte der beiden Stichproben zeigt, dass die mittlere Schlaflatenz in der Verumgruppe 5 min kleiner ist als in der Placebogruppe. In beiden Gruppen variierte die Schlaflatenz, wie aus den Standardabweichungen ersichtlich ist. Beide Stichproben sind aufgrund von Stichprobenfehlern mit einem Zufallsfehler verbunden.

Die Frage, die sich hier stellt, ist, ob die Differenz von 5 min nur Ausdruck eines zufälligen Fehlers ist oder ob diese Differenz Ausdruck einer tatsächlichen Wirkung des Schlafmittels ist. Im ersten Fall würden beide Stichproben aus identischen Populationen stammen ($\mu_p=\mu_v$), im zweiten Fall würden die beiden Stichproben aus unterschiedlichen Populationen stammen, d.h. aus Populationen mit $\mu_p \neq \mu_v$. Abbildung 1 veranschaulicht das Problem:

Tabelle 2: Permutationstest

Permutationstest

- Die 62 Schlaflatenzwerte sind in allen möglichen Sequenzen (Permutationen) angeordnet. Für n=62 Werte, die zwei Gruppen von 32 und 30 Patienten zugeordnet werden sollen, gibt es insgesamt $4{,}5 \cdot 10^{17}$ Möglichkeiten.
- Für jede Permutation wird die Gruppenzuordnung definiert: die ersten 32 Werte werden der Verumgruppe und die restlichen 30 Werte der Placebogruppe zugeordnet.
- Nach jeder Gruppenzuweisung werden die interessierenden Teststatistiken, hier die Differenz der Mittelwerte, berechnet.
- Schließlich wird der p-Wert bestimmt, indem die Anzahl der Teststatistiken, für die die Mittelwertdifferenz ≥5 min beträgt (dies war die beobachtete Mittelwertdifferenz der Studie), durch die Anzahl aller Teststatistiken geteilt wird.

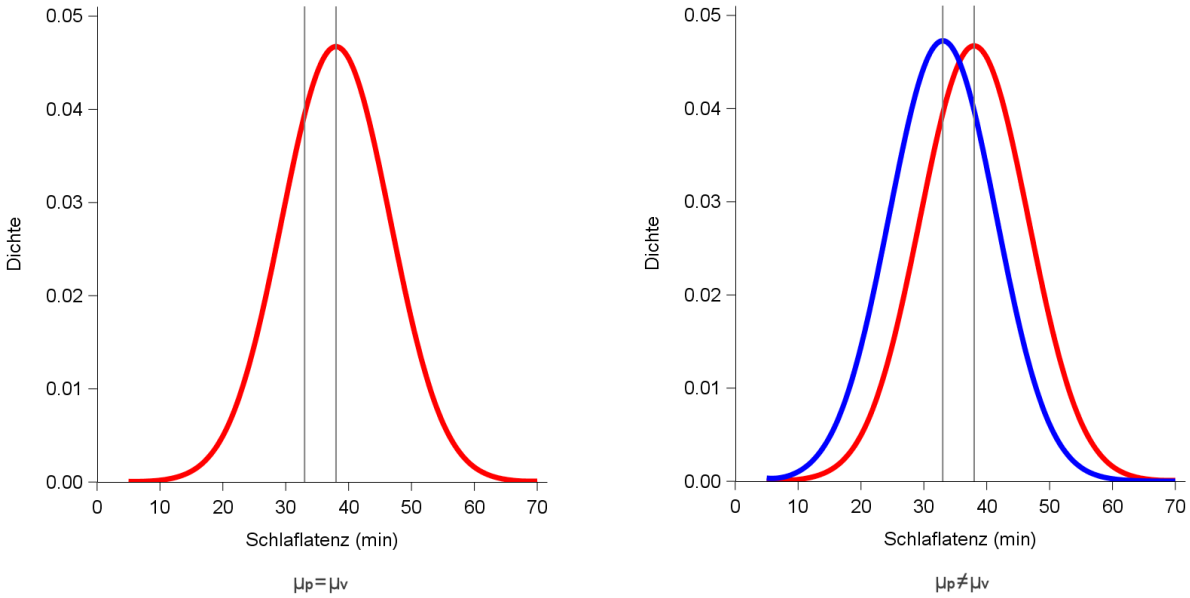

Abbildung 1: Normalverteilungen der Schlaflatenz in der Zielpopulation

Könnte es sein, dass sich Placebo und Verum in Bezug auf die wahren Schlaflatenz-Durchschnitte nicht unterscheiden, d.h. aus der gleichen Population mit z.B. μ=38 min stammen, und die beiden Stichproben-Durchschnitte (33 min und 38 min) lediglich ein Ausdruck des Stichprobenfehlers sind, ähnlich wie beim Münzwurf einer ungezinkten Münze? Oder könnte es sein, dass das neue Schlafmittel tatsächlich einen Einfluss auf die Schlaflatenz hat, so dass die wahren Mittelwerte aus Zielpopulationen mit unterschiedlichen Mittelwerten stammen ($\mu_p \neq \mu_v$)?

Erwartung der statistischen Variabilität von Studienergebnissen aufgrund eines Zufallsfehlers

Ein Signifikanztest kann gewisse, wenn auch unvollständige Informationen zu diesen zentralen Fragen liefern. Zur Beantwortung der obigen Fragen muss zunächst das Verhalten der Differenz der Mittelwerte aufgrund des Zufallsfehlers bestimmt werden, wobei angenommen wird, dass eine Nullhypothese H_0 wahr wäre. Es gibt eine unendliche Menge von Nullhypothesen. In der Medizin hat sich die Nil-Hypothese durchgesetzt, d.h. die Nullhypothese, dass es keinen Zusammenhang zwischen der Behandlungszuweisung (Placebo oder Verum) und der Schlaflatenz gibt (d.h. $\mu_p=\mu_v$). Die griechischen Buchstaben zeigen an, dass sich diese Nullhypothese auf die Zielpopulation bezieht. Unter dieser Hypothese sind Mittelwertunterschiede, die nicht gleich Null sind, ein Ausdruck des Zufallsfehlers. Ähnlich wie extreme Ergebnisse von Experimenten selten beobachtet werden, wenn eine ungezinkte Münze geworfen wird (z.B. 10 Würfe und es erscheint 10-mal Kopf), nimmt die Differenz der Mittelwerte unter der Nullhypothese selten extreme Werte an.

Aber wie viele permutierte Anordnungen von Patienten, die in zwei Gruppen aufgeteilt sind, gibt es und wie würden sich die Unterschiede der Mittel in diesen Arrangements verhalten, wenn die Nullhypothese $\mu_p=\mu_v$ wahr wäre? Die Schwierigkeit bei der Beantwortung dieser Frage liegt darin, dass das Verhalten der Mittelwertunterschiede unter der Nullhypothese von der Variabilität der Schlaflatenz innerhalb der Stichproben und der Größe der Stichproben abhängt.

Tabelle 3: Kriterien für die Testauswahl

Der im Text erwähnte t-Test für unabhängige Stichproben dient dem Vergleich der Mittelwerte zweier Stichproben und wird durchgeführt, wenn mehrere Bedingungen erfüllt sind: Die Variable, für die der Vergleich durchgeführt wird, muss intervallskaliert und in beiden Stichproben normalverteilt sein. Die Anforderung der Unabhängigkeit der Stichproben wäre nicht erfüllt, insbesondere im Falle wiederholter Messungen an derselben Stichprobe.
Ähnlich wie bei anderen statistischen Tests sollten die Modellannahmen geprüft werden. Im Allgemeinen spielen die folgenden Kriterien bei der Testauswahl eine wichtige Rolle:

1. Die Verteilung der Variablen: So genannte parametrische Tests erfordern im Gegensatz zu nichtparametrischen Tests eine bestimmte Verteilung der Variablen. Der t-Test ist daher einer der parametrischen Tests.
2. Die Abhängigkeit oder Unabhängigkeit der Stichproben: Die Abhängigkeit der Stichproben ist hauptsächlich bei wiederholten Messungen vorhanden. Wenn Messungen an derselben Stichprobe vor und nach einem Eingriff durchgeführt werden, sind die beiden Messreihen abhängig.
3. Die Anzahl der Stichproben. Es gibt Tests mit nur einer Stichprobe − zum Beispiel, wenn man untersucht, ob der IQ einer Studienpopulation von 100 abweicht. Wenn zwei oder mehr Gruppen hinsichtlich des IQs verglichen werden, ist ein Zwei- oder Mehrstichproben-Test erforderlich.

Um also vorherzusagen, wie sich die Unterschiede der Mittelwerte verhalten würden, wenn die Nullhypothese wahr wäre, muss man diese beiden Einflussgrößen berücksichtigen. Hier ist eine Art Normalisierung hilfreich, die durch das folgende Beispiel veranschaulicht werden soll. Ein Mittelwertunterschied von 3 Sekunden wird für zwei Gruppen von Marathonläufern (2 Stunden, 3 Minuten, 40 Sekunden gegenüber 2 Stunden, 3 Minuten, 43 Sekunden) und für zwei Gruppen von 400-Meter-Läufern (46 Sekunden gegenüber 49 Sekunden) beobachtet. Bei ähnlichen Läufer-Gruppen haben die Unterschiede von 3 Sekunden eine unterschiedliche Bedeutung. Bei Marathonläufern ist der Unterschied im Verhältnis zur durchschnittlichen Gesamtdauer des Laufs sehr gering, während er bei 400-Meter-Läufern relativ groß ist. Das Verhältnis zur durchschnittlichen Laufdauer ist eine Art Normalisierung. Die Wahl des statistischen Tests, der eine solche Normierung gewährleistet, bestimmt, welche Teststatistik gewählt wird. Wenn z.B. der t-Test für unabhängige Stichproben gewählt wird, ist die entsprechende Testvariable die t-Statistik, für den Chi-Quadrat-Test die Chi-Quadrat-Statistik usw. Die Wahl des geeigneten statistischen Tests hängt wiederum von Kriterien ab, die in Tabelle 3 kurz erläutert werden.

Die t-Statistik ist definiert als:

$$t = \frac{\text{Beobachtete Differenz der Mittelwerte} - \text{Erwartete Differenz der Mittelwerte}}{\text{Standardfehler der beobachteten Differenz der Mittelwerte}}$$

Die erwartete Differenz der Mittelwerte in der Formel der t-Statistik ist der unter der Nullhypothese H_0 angenommene Wert. Im Falle der Nullhypothese μp=μv wird eine Differenz von null Minuten erwartet. Dies vereinfacht die t-Statistik:

$$t = \frac{\text{Beobachtete Differenz der Mittelwerte}}{\text{Standardfehler der beobachteten Differenz der Mittelwerte}}$$

Bei ungleichen Varianzen wird der Standardfehler der Differenz der Mittelwerte nach folgender Formel berechnet:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

mit
n_1: Anzahl von Patienten in Gruppe 1 (Placebo)
n_2: Anzahl von Patienten in Gruppe 2 (Verum)
s_1^2: Varianz der Schlaflatenz in Gruppe 1
s_2^2: Varianz der Schlaflatenz in Gruppe 2

Die Formel ändert sich, wenn die Varianzen gleich sind (Formel nicht dargestellt). Der Standardfehler der Differenz der Mittelwerte hängt von den Varianzen der Variablen (Schlaflatenz) und den Gruppengrößen der zu vergleichenden Gruppen ab. Nach der Bestimmung des Standardfehlers ergibt sich die t-Statistik für zwei unabhängige Stichproben mit ungleichen Varianzen:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Unabhängigkeit bedeutet, dass die beiden Patientengruppen voneinander unabhängig sind und auch dass die Patienten innerhalb der Gruppen unabhängig voneinander sind. Die Unabhängigkeit wird beispielsweise verletzt, wenn das Ergebnis eines Patienten statistisch gesehen zu beiden Patientengruppen beitragen würde. Ebenso wird die Unabhängigkeit verletzt, wenn Patienten derselben Gruppe sich gegenseitig in Bezug auf die Ergebnisse von Interesse beeinflussen würden. Die Unabhängigkeit ist auch verletzt, wenn ein Merkmal von einer Gruppe von Patienten im Laufe der Zeit mehrfach erhoben wird (z.B. vor und nach der Behandlung). Die Daten der Schlafstudie haben nun folgenden t-Wert:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} = \frac{38 - 33}{\sqrt{\frac{8{,}5^2}{30} + \frac{8{,}4^2}{32}}} = 2{,}33$$

Der t-Wert für die konkrete Studie beträgt daher +2,33. Diese Verteilung kann mit Hilfe der sogenannten Freiheitsgrade (df) bestimmt werden. Die Anzahl der Freiheitsgrade

ist die Anzahl der Werte, die ohne Veränderung der Mittelwerte frei variiert werden können. Wenn es z.B. drei Zahlen k, l und m gibt und ihre Summe 100 ist, ist klar, dass, wenn zwei der drei Zahlen bekannt sind, automatisch die dritte Zahl gegeben ist. Wenn k=20 und l=70 ist, muss m 10 sein. Bei 62 Patienten in der Studie hat man $n_1-1+n_2-1=30-1+32-1=60$ Freiheitsgrade. Wurden 60 Werte frei gewählt, so hat man für die letzten beiden Beobachtungen keine weitere Wahl.

Mit Hilfe der 60 Freiheitsgrade, kann nun die geeignete Verteilung unter der Annahme der Nullhypothese dargestellt werden. Auf die Darstellung der Formel zur Erstellung der t-Verteilung wird aus didaktischen Gründen verzichtet (es ist das Verhältnis der Standard-Normalvariable z und der Quadratwurzel eines Chi-Quadrat-Wertes mit n Freiheitsgraden geteilt durch n). Die t-Verteilung ist symmetrisch und glockenförmig wie die Normalverteilung (Abbildung 2).

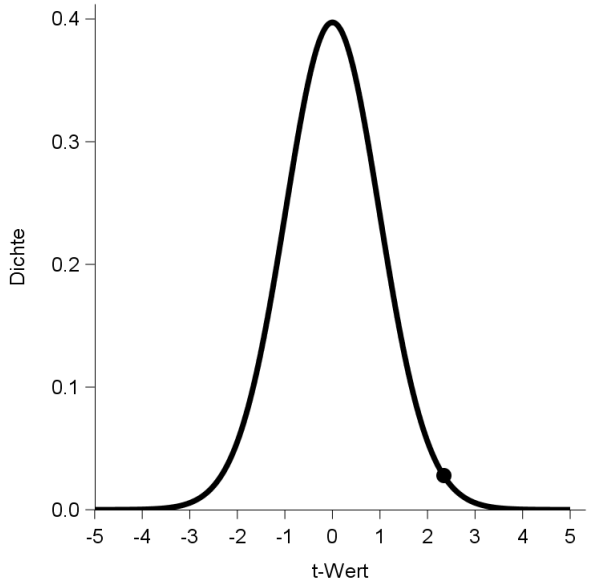

Abbildung 2: t-Verteilung mit 60 Freiheitsgraden und markiertes Studienergebnis (t=2,33)

Die Wahrscheinlichkeitsdichtefunktion (PDF) liefert in Abhängigkeit von den t-Werten sogenannte Dichtewerte. Im Gegensatz zu den Wahrscheinlichkeiten, die nur Werte zwischen 0 und 1 annehmen können, können Dichten auch Werte >1 annehmen.

Interpretation des t-Wertes

Ein einziger Dichtewert der PDF hat keine praktische Bedeutung. Die Gesamtfläche unter der Kurve der PDF ist 1, so dass (Teil-)Flächen unter der Wahrscheinlichkeitsdichtefunktion die Interpretation von Wahrscheinlichkeiten haben. Im Rahmen der Studie ist es nun möglich, die Frage zu beantworten, wie hoch die Wahrscheinlichkeit ist, dass der t-Wert ≥+2,33 unter der Nullhypothese ($\mu_p=\mu_v$) annimmt, d.h. t=0.

Die kumulative Verteilungsfunktion (CDF) liefert die Wahrscheinlichkeit, dass ein t-Wert kleiner oder gleich einem konkreten Wert t_k ist. Es ist auch möglich, die CDF zu verwenden, um die Wahrscheinlichkeit zu berechnen, dass $t\geq t_k$ wird, indem die Wahrscheinlichkeit für t-Werte $<t_k$ vom Wert 1 subtrahiert wird. Die Formel für diese Funktion wird an dieser Stelle weggelassen, kann aber im Internet leicht gefunden werden [6]. Im Fall der Schlafstudie ist $t_k\geq+2,33$. Abbildung 3 zeigt die Fläche unter der Verteilung für t≥+2,33 bei einseitiger Betrachtung und für die Flächen unter der Verteilung für t≤−2,33 und t≥+2,33 bei zweiseitiger Betrachtung.

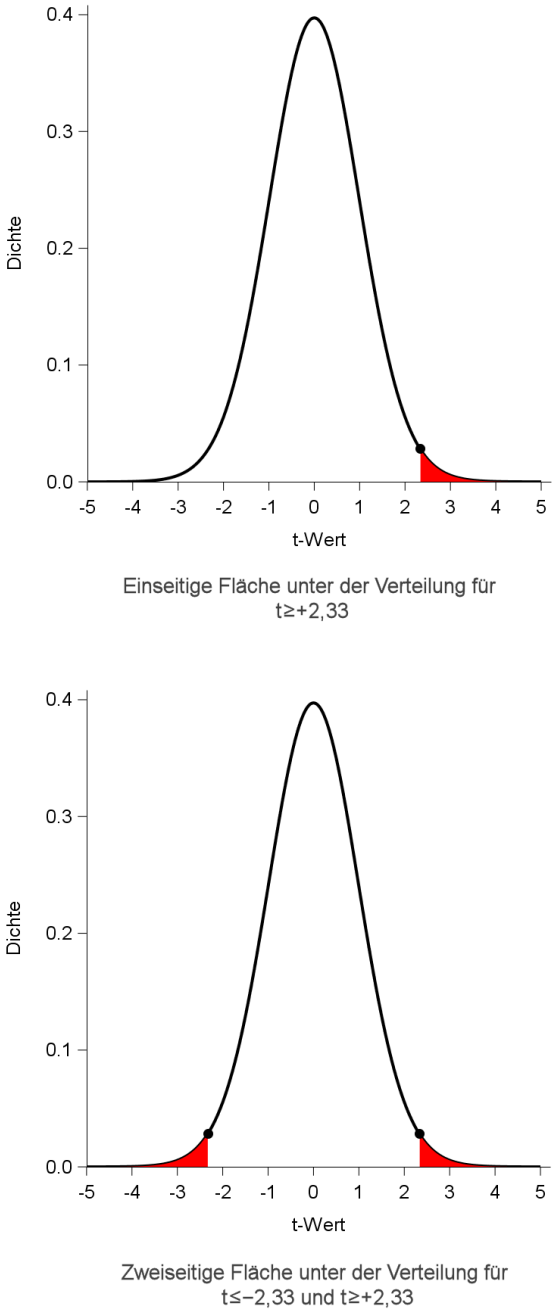

Abbildung 3: t-Verteilung mit 60 Freiheitsgraden und markierten Flächen unter der Verteilung für t≥+2,33 und t≤−2,33

Der einseitige Bereich hat einen Betrag von 0,01. Das bedeutet, dass die Wahrscheinlichkeit, dass Studien unter der Annahme der Nullhypothese ($\mu_p=\mu_v$) einen t-Wert von ≥+2,33 erzeugen, 1% beträgt. Bei zweiseitiger Betrachtung beträgt die Wahrscheinlichkeit, dass Studien unter der Annahme der Nullhypothese ($\mu_p=\mu_v$) einen t-Wert von ≤–2,33 oder ≥+2,33 erzeugen, 2%. Die Wahrscheinlichkeit von 1% entspricht dem einseitigen p-Wert, während die Wahrscheinlichkeit von 2% dem zweiseitigen p-Wert entspricht.

Der p-Wert – Erläuterung und einige Caveats

Interpretation des p-Wertes

Der p-Wert gibt somit die Wahrscheinlichkeit (Kriterium 1) unter einer Nullhypothese (Kriterium 2) an, ein Ergebnis wie das vorliegende Studienergebnis zu beobachten oder Studienergebnisse zu beobachten, die noch stärker von der Nullhypothese (Kriterium 3) abweichen. Alle drei Kriterien sind notwendige Kriterien für die Definition des p-Wertes.

Wichtig ist hier, dass der p-Wert eine Aussage über das Verhalten einer Teststatistik bei Vorliegen eines zufälligen Fehlers unter der Nullhypothese macht. Bei einem p-Wert von 0,01 würde nur 1% der Studien einen t-Wert von ≥+2,33 erzeugen, wenn die Nullhypothese wahr wäre. Der p-Wert macht also auch eine Aussage über Ergebnisse von Studien, die nicht beobachtet wurden (kontrafaktisches Element). Ferner muss betont werden, dass der p-Wert unter der Bedingung berechnet wurde, dass die Nullhypothese H_0 wahr wäre, weshalb der p-Wert auch als bedingte Wahrscheinlichkeit bezeichnet wird. Die Nullhypothese wurde lediglich angenommen, unabhängig davon, wie groß der Wahrheitsgehalt dieser Hypothese ist.

Fisher interpretierte den p-Wert als ein kontinuierliches Maß für die Evidenz gegen die Nullhypothese. Er sagte: „Kein Wissenschaftler hat ein festgelegtes Signifikanzniveau, auf dem er von Jahr zu Jahr und unter allen Umständen Hypothesen ablehnt; er macht sich vielmehr zu jedem einzelnen Fall Gedanken im Lichte der Evidenz und seiner Ideen“ [7]. Das bedeutet, dass nach Fishers Schule die Einstufung eines p-Wertes kontextabhängig ist und die Anwendung eines festen Schwellenwertes von typischerweise 0,05 nicht gerechtfertigt ist. Die orthodoxe Ablehnung einer Nullhypothese bei einem vordefinierten Schwellenwert von 0,05 stammt von der konkurrierenden Schule von Neyman und Pearson, die den Hypothesentest als entscheidungstheoretisches Verfahren einführten.

Was bedeutet ein großer p-Wert von z.B. 0,70? Technisch gesehen bedeutet er, dass die Wahrscheinlichkeit 70% beträgt, das beobachtete Studienergebnis oder Studienergebnisse, die noch stärker von der Nullhypothese abweichen, zu beobachten, unter der Annahme die Nullhypothese sei wahr. In der Praxis bedeutet das, dass der Signifikanztest wenig Evidenz gegen die getestete Nullhypothese oder das statistische Modell liefert. Es bedeutet jedoch nicht, dass die Nullhypothese wahr ist. Der p-Wert ist eine Funktion der Stärke des Effekts (z.B. beobachteter Mittelwertunterschied, hier 5 min) und der Studiengröße (hier 62 Frauen). Bei einem großen p-Wert kann tatsächlich ein starker Effekt vorhanden sein, aber die Studiengröße war sehr klein. Typische Fehler bei der Definition von p-Werten werden im Folgenden diskutiert.

„Der p-Wert ist die Wahrscheinlichkeit, dass die Nullhypothese wahr ist.“ Der p-Wert macht keine Aussage über die Wahrscheinlichkeit der Wahrheit der Nullhypothese, jedoch wurde der p-Wert unter der Annahme berechnet, dass die Nullhypothese wahr ist. Übrigens fehlt hier der Hinweis auf noch extremere Ergebnisse der Studie (kontrafaktisches Element).

„Der p-Wert ist die Wahrscheinlichkeit eines Typ-I-Fehlers.“ Diese Aussage ist falsch, weil sie die Prinzipien des Signifikanztests (Fisher) mit denen des Hypothesentests (Neyman & Pearson) vermischt. Nach der Schule von Fisher gibt es kein a priori festgelegtes Signifikanzniveau (auch Typ-I-Fehler genannt). Im Gegensatz dazu wird nach Neyman & Pearson das Signifikanzniveau, auch Typ-I-Fehler genannt, vor Beginn der Studie festgelegt, während der p-Wert aus dem statistischen Modell und den Studiendaten nach Durchführung der Studie abgeleitet werden. Nach Neyman & Pearson bleibt der Typ-I-Fehler nach dem Ende der Studie unverändert und der p-Wert wird mit dem a priori festgelegten Typ-I-Fehler verglichen, um eine Entscheidung zu treffen.

Der Typ-I-Fehler, auch α-Fehler genannt, wird nach Neyman und Pearson vor Beginn der Studie bestimmt. Am Ende der Studie wird der p-Wert, der sich aus der Nullhypothese, dem statistischen Modell (z.B. t-Test) und den Studiendaten ergibt, mit dem α (meist 0,05) verglichen. Die Aussage, dass „ein niedriger p-Wert den Zufall als Erklärung für einen beobachteten Unterschied ausschließt“, beweist einen groben Mangel an Verständnis.

Nahezu korrekt klingende Definitionen des p-Wertes sind zum Beispiel: „Der p-Wert ist die Wahrscheinlichkeit, das vorliegende Studienergebnis oder noch extremere Studienergebnisse zu beobachten“. In dieser Definition fehlt die zentrale Bedingung (Kriterium 2) des p-Wertes: Die Berechnung erfolgt unter der Annahme, dass die Nullhypothese zutrifft. Auch die folgende falsche Definition ist beliebt: „Der p-Wert ist die Wahrscheinlichkeit, das vorliegende Studienergebnis unter der Nullhypothese zu beobachten.“ Hier fehlt Kriterium 3: Der p-Wert macht auch eine Aussage über unbeobachtete Studienergebnisse, die noch stärker von der Nullhypothese abweichen als das vorliegende Studienergebnis.

Beim Signifikanztest nach Fisher gibt es keinen so genannten Typ-I-Fehler und Typ-II-Fehler, es gibt kein Konfidenzintervall, keine Alternativhypothese und kein Konzept für statistische Macht (Power) oder Stichprobengrößenberechnungen. Diese Phänomene gehen auf Neyman & Pearson zurück und werden erst bei der Durchführung von Hypothesentests relevant, die entscheidungstheoretisch nur dann gültig sind, wenn alle Schritte des Hypothesentestverfahrens eingehalten werden, weshalb die

Autoren auch von Neyman-Pearson-Orthodoxie sprechen [8]:

1. Definition der Nullhypothese und Alternativhypothese vor Beginn der Studie
2. Festlegung des Typ-I-Fehlers und Typ-II-Fehlers vor Beginn der Studie
3. Festlegung der Teststatistik vor Beginn der Studie
4. Berechnung der erforderlichen Stichprobengrößen vor Beginn der Studie
5. Durchführung der Studie unter Einhaltung der erforderlichen Stichprobengrößen
6. Berechnung der Teststatistik und Vergleich mit dem kritischen Wert der Teststatistik oder Vergleich des p-Wertes mit dem vorab definierten Typ-I-Fehler nach Durchführung der Studie
7. Entscheidung: Wenn $p \leq \alpha$, wird die Nullhypothese abgelehnt, wenn $p > \alpha$, wird die Nullhypothese nicht abgelehnt (nach Durchführung der Studie).

Wenn die Schritte 1–7 nicht eingehalten werden, verliert das entscheidungstheoretische Verfahren des Hypothesentestens seine Gültigkeit. Die Entscheidungsregel (7. Schritt) muss konsequent angewendet werden. Wenn z.B. $\alpha=0{,}05$ angegeben wurde und $p=0{,}07$ am Ende der Studie herauskam, dann kann nach Neyman & Pearson nicht gesagt werden, dass es einen „Signifikanztrend“ oder etwas Ähnliches gab, sondern nur, dass die Nullhypothese nicht abgelehnt wurde. Auch werden p-Werte $\leq 0{,}05$ nach Neyman & Pearson nicht in z.B. $p \leq 0{,}05$*, $p \leq 0{,}01$** und $p \leq 0{,}001$*** weiter unterteilt.

Bedingungen, die für die korrekte Interpretation des p-Wertes notwendig sind

Viele einführende Lehrbücher der Biostatistik führen lediglich die Theorie der Signifikanztests ein. Das bedeutet, dass es außer dem Zufallsfehler keine weiteren Fehlerquellen gibt. In der Praxis der empirischen Studien ist dies jedoch eine unrealistische Annahme. Greenland et al. [9] weisen zu Recht darauf hin, dass im Falle eines niedrigen p-Wertes nur ein Signal gegeben wird, dass mit dem sogenannten statistischen Modell etwas nicht in Ordnung sein könnte. Das statistische Modell besteht aus drei Komponenten: Der gewählten Teststatistik, der gewählten Nullhypothese und den empirischen Studiendaten.
Zusätzlich zu der Hypothese, dass der niedrige p-Wert Evidenz gegen die Nullhypothese darstellt, müssen die folgenden alternativen Erklärungen in Betracht gezogen werden, die alle mit dem statistischen Modell zusammenhängen und somit den p-Wert beeinflussen:

- Es wurde eine ungeeignete Teststatistik angewandt.
- Es kam zu einem Selektionsbias in die Studie oder zu einem Selektionsbias bei der Nachbeobachtung der Probanden.
- Der Vergleich zwischen zwei Stichproben ist konfundiert (Vermengung von Effekten).
- Es gibt einen Informationsbias bei der Messung der Variablen in der Studie.

Wenn der p-Wert niedrig ist, können wir nur den Schluss ziehen, dass etwas mit dem statistischen Modell nicht stimmt. Der p-Wert selbst zeigt jedoch nicht, was mit dem Modell nicht stimmt. Der unerfahrene Benutzer des Signifikanztests betrachtet einen niedrigen p-Wert nur als einen Hinweis darauf, dass die Nullhypothese falsch sein könnte. Zusätzlich zu der von Fisher erklärten kontextuellen Abhängigkeit der Bedeutung niedriger p-Werte muss das Ergebnis eines Signifikanztests immer im Licht des vollständigen statistischen Modells gesehen werden.

Fazit

Fishers Signifikanztest ist ein anderes Verfahren als der Hypothesentest von Neyman & Pearson, was oft ignoriert wird. Während der Signifikanztest einen p-Wert erzeugt, der nach Fisher kontextabhängig als ein kontinuierliches Maß für die Evidenz gegen die Nullhypothese interpretiert werden sollte, dient der p-Wert als Entscheidungskriterium, wenn die notwendigen Schritte des Hypothesentests befolgt werden. Der Signifikanztest führt zum p-Wert, dessen Definition drei Kriterien enthalten muss: Die Wahrscheinlichkeit, die Verwendung der Nullhypothesen-Annahme und das kontrafaktische Element des p-Wertes. P-Werte können aus verschiedenen Gründen klein sein, und die Evidenz gegen die Nullhypothese ist einer von mehreren konkurrierenden Gründen in empirischen Studien.

Anmerkungen

Interessenkonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel haben.

Literatur

1. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. The empire of chance. How probability changed science and everyday life. Cambridge: Cambridge University Press; 1989.
2. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. PeerJ Preprints. 2018;6:e26857v4. DOI: 10.7287/peerj.preprints.26857v3
3. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70:129-33. DOI: 10.1080/00031305.2016.1154108
4. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305-307. DOI: 10.1038/d41586-019-00857-9
5. Cox DR. Principles of statistical inference. Cambridge: Cambridge University Press; 2006. DOI: 10.1017/CBO9780511813559

6. Student's t-distribution. In: Wikipedia. [accessed 2019 May 16]. Available from: https://en.wikipedia.org/wiki/Student%27s_t-distribution

7. Fisher RA. Statistical methods and scientific inference. Edinburgh: Oliver & Boyd; 1956.

8. Oakes MW. Statistical inference. Chichester: Wiley; 1986.

9. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016 Apr;31(4):337-50. DOI: 10.1007/s10654-016-0149-3

10. Manly BFJ. Randomization, bootstrap and Monte Carlo methods in biology. London: Chapman & Hall; 1996. Randomization; p. 3-7.

11. Feinstein AR. Principles of medical statistics. Boca Raton: Chapman & Hall/CRC; 2002. Testing stochastic hypotheses; p. 190-1.

Korrespondenzadresse:
Prof. Dr. med. Andreas Stang, MPH
Institut für Medizinische Informatik, Biometrie und Epidemiologie, Universitätsklinikum Essen, Hufelandstr. 55, 45147 Essen, Deutschland, Tel.: 0201-723-77-289, Fax: 0201-723-77-333
andreas.stang@uk-essen.de

Bitte zitieren als
Stang A, Kowall B. Fisher's significance test: A gentle introduction. GMS Med Inform Biom Epidemiol. 2020;16(1):Doc03. DOI: 10.3205/mibe000206, URN: urn:nbn:de:0183-mibe0002065

Artikel online frei zugänglich unter
https://www.egms.de/en/journals/mibe/2020-16/mibe000206.shtml

Veröffentlicht: *11.05.2020*