International Journal of Business and Management Invention (IJBMI)
ISSN (Online): 2319-8028, ISSN (Print):2319-801X
www.ijbmi.org || Volume 11 Issue 7 Ser. I || July 2022 || PP 46-50

Understanding Statistical Data Testing: an overview

Mário Basto[1], Teresa Abreu[1], Ricardo Gonçalves[1], José M. Pereira[2]
Corresponding Author:Mário Basto
[1]*Department of Applied Sciences and Industrial Engineering, Higher School of Technology, Polytechnic Institute of Cávado and Ave, Barcelos, Portugal*
[2]*CICF - Research Center on Accounting and Taxation, Polytechnic Institute of Cávado and Ave, Barcelos, Portugal*

ABSTRACT:*In all business sectors, decision-making is crucial and frequently requires drawing conclusions from data. As a result, it is essential to employ statistical tools when extracting knowledge from data.However, there is no one size fits all approach to applied inference research. It is important thatdecision makers understand their options, including the advantages and disadvantages of eachapproach. Fisher's test of significance, Neyman-Pearson's test of acceptance, null hypothesis significance testing (NHST), and Bayesian approach, are the four major choices when it comes to hypothesis testing. Each of them is briefly discussed in this paper. This is intended to be a concise summary of these methodologies so that decision makers may better understand the ideas, use the right strategy in every circumstance, and critically assess the data's findings.*
KEY WORD*: Test of significance, Hypothesis Testing, Null Hypothesis Significance Testing (NHST), p-value, Bayesian Testing, Bayes Factor.*

Date of Submission: 03-07-2022 Date of Acceptance: 16-07-2022

I. INTRODUCTION

Several statistical solutions are at the disposal of the researcher to solve problems of inductive inference. From these options, the researcher chooses the one or the ones that best suit the research problem that he wants to address. There is no single solution to all problems [1]. Therefore, it is helpful to be aware of the different options available. This paper provides a short description of the most prevalent types of data testing, in particular, Fisher's test of significance, Neyman-Pearson's test of acceptance, null hypothesis significance testing (NHST), and Bayesian approach to data testing.

II. FISHER'S TEST OF SIGNIFICANCE

According to Fisher [2], every experiment may be conducted with the aim of disproving the null hypothesis H_0. Fisher created the so-called null hypothesis significance testing theory, a tool that should only be used for problems about which there is little or no knowledge. When one knows relatively little about a problem, this is the case [3]. The purpose is to use probability calculations to evaluate evidence.

As stated by Fisher's null hypothesis testing (Fisher's test of significance), only one hypothesis must be defined, the null hypothesis H_0, with no expressly stated alternative hypothesis H_1(albeit implicitly). It is called null hypothesis testing because the hypothesis must be nullified with research data, to be able to explain it [4]. It does not even have to be a null hypothesiswhich always equals zero(such as equal means among groups or zero correlation).

In contrast to the Newman-Pearson approach to data testing, Fisher's approach to significance testing allows all steps of the technique to be established a posteriori, after the data has been gathered [5,6].The process involves calculating the theoretical probability of obtaining data as least as extreme as the one observed, assuming the null hypothesis $H_0$is true. This probability is called the *p*-value. That is, *p*-value $= P(x + |H_0)$, where $x +$ denotes the data collected or even more extreme. The definition means that the same set of data can assign different *p*-values depending on the study sampling design[7,8]. It is also worth noting that, if the experiment is repeated, the *p*-value cannot be regarded as an error rate. It applies only to the actual data.

The purpose of this test is to obtain a statistically significant result. The significance of the result means that its *p*-value is low, which means that the dataare unlikely to have occurred by random variation alone, and therefore can be considered evidence against the null hypothesis. It is up to the researcher reporting the actual *p*-value to decide how little the *p*-value must be to be statistically significant. The *p*-value is frequently compared to a specified threshold. The *p*-value threshold used to reject the null hypothesis is known as the significance level. This significance level does not have to be fixed. It can be changed later on.

A significant result indicates that a rare occurrence happened or that H_0fails to explain the data. As a result, the lower the *p*-value, the stronger the evidence against the null hypothesis [5,6,9]. Nonetheless, because

a low *p*-value says nothing about the evidence against any other hypothesis, the *p*-value might overestimate the evidence against H_0. Furthermore, the *p*-value does not answer the real question of inductive inference: how credible is the research hypothesis (implicit alternative hypothesis) in light of the data? [10].Another disadvantage of the *p*-value is that it ignores the magnitude of the effect. A tiny effect in a large sample size study can have the same *p*-value as a substantial effect in a small sample size study [8].

Fisher's technique does not include statistical power or effect size (described in the next chapter), which is an important aspect to underline. Furthermore, the availability of one single hypothesis is the most fundamental testing limitation since it prevents comparison testing of competing hypotheses. As a result, symmetrical comparison of two or more hypotheses is impossible.

Following is a summary of the processes involved in Fisher's null hypothesis testing: one starts by constructing a statistical null hypothesis, which typically denotes the lack of an effect. The real *p*-value is then computed and provided so that a judgment can be made. The strength of the evidence against the null hypothesis is stated rather than being accepted or rejected [4]. Having strong evidence against the null hypothesis does not mean that the alternative is true. A replication of the study is required.

III. NEYMAN-PEARSON'S TEST OF ACCEPTANCE

The idea of Neyman and Pearson [11] is that the truth or falsehood of a hypothesis cannot be determined by any test based on a theory of probability on its own, however, one can look for guidelines to adhere to in order to make sure that the choice is not frequently wrong in the long run [11].

In contrast to Fisher's method, Neyman-Pearson's method focuses on choosing amongst competing hypotheses rather than eliminating a hypothesis [6]. In addition to the central hypothesis H_M(which is very similar to Fisher's null hypothesis and is also referred to as the null hypothesis), there is an alternative hypothesisH_A. Neyman-Pearson's approach utilizes long-run error probabilities as opposed to Pearson's method use of evidence to reject the null hypothesis [12].Some steps of the Neyman-Pearson approach must be established a priori [4,5,6].

When testing the hypotheses, there are two errors that can be made: the type I error, which involves rejecting the central or null hypothesis when it is true, and the type II error, which involves rejecting the alternative hypothesis when it is true. The probability of making a type I error is called alpha (α), and the probability of making a type II error is called beta (β). The power of the test $(1-\beta)$,is the probability of rejecting correctly the central or null hypothesis when it is false.

Unlike Fisher's method, Neyman-Pearson' data testing incorporates the concept of effect size [6]. An effect size is a measurement of anything of interest that is typically a standardized indicator of the strength of the effect, though it can also be expressed in the original units. It is a metric that is not dependent on sample size. Even so, the sample size can have an impact since larger samples produce more accurate estimates of the population effect size. The effect size must be established a priori. The effect size provides information on the probability of committing a type II error (β). The portion of the central hypothesis $H_M$that one does not wish the test to reject is represented by the minimal effect size. In other words, it includes values that are not of relevance to study but are desired to fall within the central hypothesis [6]. Values beyond this minimal effect size are those that are thought to be significant for research.The main distinction between the central hypothesis of Newman and Pearson and the null hypothesis of Fisher is that the former incorporates any value below the minimum effect size to be captured by the test (effect sizes are not part of Fisher's approach), and is one of two competing hypotheses to be tested.

The alpha level (α), which is the probability of making a type I error in the long run, is fixed a priori [4,6]. Long run refers to repeated similar tests over which one can anticipate rejecting the correct central or null hypothesis in approximately$\alpha \times 100\%$ of cases, but it does not apply to a single research.Keep in mind that the *p*-value for Fisher and the alpha level (α) for Neyman-Pearson are not the same thing. The focus of Neyman-Pearson's approach is on choosing which hypothesis to accept, not on the strength of the evidence against the null hypothesis. The alpha level must be decided upon beforehand and does not admit gradation.

Based on the probability distribution of the statistical test under the null hypothesis, the value of α assists in drawing a critical region or rejection region of the central hypothesis H_M. One accepts the alternative hypothesis $H_A$if the test value is inside the critical zone. One accepts the central hypothesis $H_M$if the test value is outside the critical zone, for a test's power $(1-\beta)$adequate; otherwise, no conclusion may be drawn [6].For the test to have a good power, the sample size *n* needs to be calculated beforehand. Note that in Neyman-Pearson's testing, the *p*-value has only one purpose: to serve as a marker for choosing among the hypotheses [6].

Neyman-Pearson's data testing can be summarized as follows: two statistical hypotheses, $H_M$and $H_A$are created. According to H_A, define the effect size that is considered important. Set the values for the errors $\alpha$and β and determine the sample size *n* in advance (by employing suitable software). Then the experience is carried outand the data is analyzed, and one of the hypotheses is accepted, without any assessment of the veracity of any of the hypotheses.

IV. NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST)

The most commonly used method for testing hypotheses is the so-called Null Hypothesis Significance Testing (NHST). It combines Fisher's and Neyman-Person's techniques, but it is not well defined and could lean more in one direction than the other. It emerged from these two opposing perspectives on the problem of how to employ data to test hypotheses and seeks to determine whether or not the alternative hypothesis $H_1$is most likely to be accurate.

NHST can be applied, for instance, while performing a significance test using Neyman-Pearson's principles, by adjusting the alpha level (α) (probability of making a type I error) to match the observed*p*-value. However, Fisher's method is the one that comes closest to the NHST despite the inclusion of an alternative hypothesis.Although they are not required in NHST, effect size and β(probability of making a type II error) must be considered when deciding the a priori sample size *n*to achieve high power $(1-\beta)$, even though statistical power is not typically taken into account.

NHST is often carried out as follows: the statistical null hypothesis of no effect is first established. The *p*-value is then calculated according to the alternative hypothesis, and the conventional threshold for rejecting the null hypothesis is usually 5% significance. The result is deemed significant and the alternative hypothesis is accepted if the *p*-value is less than the significance level of 5%.Otherwise, no conclusions can be drawn.

The information provided by the *p*-value is poor. Although it reveals nothing about the data under the alternative hypothesis H_1, a low *p*-value indicates that it is unusual under the null hypothesis H_0. Results are typically viewed as final in NHST since a significant result is frequently interpreted as proving the alternative hypothesis [6]. These conclusions may have negative effects [13]. Interpreting the *p*-value is overstated in many studies. A low *p*-value or significant finding does not necessarily indicate that the effect is pertinent or of practical significance. Significant results or *p*-values cannot be used to determine which hypothesis is true. Keep in mind that *p*-values do not provide information on the size or accuracy of the effect measured [14].The *p*-values must be reported along with power and effect size information or with confidence intervals since, unlike effect sizes, they depend on sample size.

V. BAYESIAN APPROACH

A crucial feature of Bayesian statistical approache is the inclusion of prior beliefs before collecting data. Prior distribution, which is required initially before data collection, is a probability distribution of parameter values. In Bayesian inference, these parameter values acquire a new level of credibility consistent with the data gathered; this new level of credibility is known as the posterior distribution.

NHST can only reject the null hypothesis. On the contrary, Bayesian testing can accept or reject the null hypothesis. By using Bayesian inference, parameter estimation is possible in addition to model comparison and hypothesis testing. In contrast to the other procedures, Bayesian inference involves assigning probabilities to parameters and models. In a nutshell, Bayesian statistic relies on Bayes' rule and probability theory, and therefore allows one to change prior beliefs in light of additional evidence, answering a crucial topic that is not addressed by other methods, namely, the probability of the hypotheses given the information that have been collected.

Unlike other procedures in which the *p*-value is calculated, the Bayesian approach uses the Bayes factor developed by Jeffreys [15]. The Bayes factor measures the extent to which data are more likely under one hypothesis than under another, allowing a quantitative comparison of the predictive performance of data for the null hypothesis $H_0$and that of the alternative hypothesis H_1[15,16]. For two hypotheses $H_0$and H_1, two equivalent Bayes factors arise from the ratio of probabilities (p may denote a probability or a probability density function, depending on the event):

$$BF_{10} = \frac{p(x|H_1)}{p(x|H_0)}$$

$$BF_{01} = \frac{p(x|H_0)}{p(x|H_1)} = \frac{1}{BF_{10}}$$

As a result, a Bayes factor BF_{01} greater than one denotes that the null hypothesis is more likely to explain the data than the alternative hypothesis, and a Bayes factor BF_{01}less than one denotes the opposite. Therefore, if the Bayes factor BF_{01} equals$y$, the data predict the null hypothesis y times more accurately than the alternative hypothesis, meaning that the likelihood of the data is y times higher under H_0than under H_1.

A disadvantage of the Bayes factor for an alternative composite hypothesis is that it is dependent on the prior distribution of the parameter under H_1. A two-sample *t*-test serves as an illustration, where H_0assigns a null effect size that denotes the null difference between two means in the population, $\mu_1 - \mu_2 = 0$, and the alternative hypothesis $H_1$is given by a range of values of$\mu_1 - \mu_2$, like$\mu_1 - \mu_2 \neq 0$, which presupposes that the effect is present and it is necessary to assign a distribution to it that accounts for the uncertainty regarding the effect size's actual true value in the population (given by the difference between the two means). In the event

that $H_0$is true, all of its probability is concentrated at one location, $\mu_1 - \mu_2 = 0$. When H_1 is true, $\mu_1 - \mu_2$ has an undetermined value, and a conditional probability distribution is established over all possible values.

However, the minimum Bayes factorBF_{01}(MBF), which is the lower bound ofBF_{01}, represents the strongest evidence against the null hypothesis, and does not rely on the prior [17]. It is the lowest across all the odds of the data likelihood of $H_0$to H_1. The alternative hypothesis' prior distribution is concentrated at the data's maximum likelihood estimate to produce it [17].The greatest amount of evidence is shown against the null hypothesis by the smallest Bayes factor.

As a unique feature, the Bayes' rule (unlike the *p*-value) allows the posterior probability of any hypothesis$P(H|x)$,to be computed (H is any of the hypotheses, $x$means data):

$$P(H_0|x) = \frac{P(x|H_0)}{P(x)} \times P(H_0)$$

$$P(H_1|x) = \frac{P(x|H_1)}{P(x)} \times P(H_1)$$

The posterior odds are therefore calculated as the product of the Bayes factor and the prior odds (posterior probabilities are then calculated for each hypotheses):

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{p(x|H_0)}{p(x|H_1)} \times \frac{P(H_0)}{P(H_1)}$$

VI. CONCLUSION

When there is only one hypothesis and one wants to assess the strength of the evidence against it using new data, Fisher's method may be appropriate. The Neynman-Pearson's technique, on the other hand, is recommended when one wants to make a decision in dichotomous terms. The NHST can be used for the same objective, but the power of the test and the size of the effect must be considered. In contrast to Fisher's technique and generally also in NHST, the *p*-value just acts as a proxy in Newmann-Pearson strategy to make a decision.

The *p*-value is based both on the data that have been observed and on more extreme data that have not been reported. In light of this, *p*-values may depend on the experiment's design and stopping criteria. The *p*-value for the same data may change if the design is modified.Furthermore, the *p*-value merely assesses the degree to which the data (including more extreme one) support the null hypothesis. It does not measure the predictive performance of the data for the alternative hypothesis.

The Bayesian technique, on the other hand, focuses on updating prior information about the nonzero effect size using the posterior distribution. The Bayes factor compares the prediction effectiveness of two competing models or hypotheses and provides the relative strength of each. Therefore, only estimation and interpretation are used instead of drawing binary judgments. The use of a prior, which is a subjective judgement, is the main criticism of Bayesian statistics [16,18]. However, Ly, Raj et al. [18]show that the evidence for a particular data set is limited and that even with an unrealistic prior distribution, it is not possible to uncover as much evidence as desired.Nevertheless, when the sample size *n*increases, the prior's effect fades [19]. Overall, as sample size increases, the prior choice becomes less important and the estimated impact size's precision rises [19]. The posterior distribution from one study can be used as a prior distribution for subsequent investigations, which is another benefit of using Bayesian analysis to analyze the data.

BIBLIOGRAPHY

[1]. Cascetta, Gigerenzer, G., Krauss, S., & Vitouch, P. (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. D. Kaplan (Ed.). The Sage handbook of quantitative methodology for the social sciences (pp. 391–408). Thousand Oaks, CA: Sage.

[2]. Fisher, R.A. (1937). Professor Karl Pearson and the Method of Moments. Annals of Eugenics, 7, 303-318.

[3]. Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). The empire of chance: How probability changed science and every day life. Cambridge, UK: Cambridge University Press.

[4]. Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33, 587–606.

[5]. Fisher, R. A. (1955). Statistical methods and scientific induction. Journal of the Royal Statistical Society. Series B, Statistical methodology, 17, 69–78.

[6]. Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology, 6, 1-11.

[7]. Kruschke, J.K. (2013). Bayesian estimation supersedes the t test (2013). Journal of Experimental Psychology: General, 142(2), 573-603.

[8]. Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p-value fallacy. Annals of Internal Medicine, 30(12), 995-1004.

[9]. Fisher, R. A. (1960). The Design of Experiments, 7th Edn. Edinburgh: Oliver and Boyd.

[10]. Page, R., & Satake, E. (2017). Beyond P values and Hypothesis Testing: Using the Minimum Bayes Factor to Teach Statistical Inference in Undergraduate Introductory Statistics Courses. Journal of Education and Learning, 6(4), 254-266.
[11]. Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society: A Mathematical, Physical and Engineering Sciences, 231, 289-337.
[12]. Perezgonzalez, J.D. (2014). A reconceptualization of significance testing. Theory & Psychology, 24, 852–859.
[13]. Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. Nature, 567, 305-307.
[14]. Cumming, G. (2014). The New Statistics: Why and How. Psychological Science, 25(1), 7–29.
[15]. Jeffreys, H. (1961). Theory of probability (3rd ed.). Oxford, UK: Oxford University Press.
[16]. Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. Psychonomic Bulletin & Review, 25(1), 5-34.
[17]. Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. Duke I&E Research Paper No. 2017-05. Available at SSRN: https://ssrn.com/ abstract=2893930.
[18]. Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q.F., & Wagenmakers, E.-J. (2018). Bayesian reanalyses from summary statistics: A guide for academic consumers. Advances in Methods and Practices in Psychological Science, 1, 367-374.
[19]. Ly, A. (2018). Teaching Bayesian Estimation with the Summary Stats Module. Retrieved at 21/05/2022. URL: https://jasp-stats.org/2018/04/11/teaching-bayesian-estimation-with-the-summary-stats-module/

Mário Basto, et. al. "Understanding Statistical Data Testing: an overview." *International Journal of Business and Management Invention (IJBMI)*, vol. 11(07), 2022, pp. 46-50. Journal DOI- 10.35629/8028