Advanced Mathematical Models & Applications
Vol.8, No.1, 2023, pp.5-13

# EXAMPLE-BASED OVERVIEW TO STATISTICAL INFERENCE[*]

**Mário Basto[1†], Teresa Abreu[1], Ricardo Gonçalves[1], José M. Pereira[2]**

[1]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 statistical reasoning, hypothesis testing is of particular importance for decision-making. Based on the *p-value*, the null hypothesis significance testing (NHST) is the most common approach. Bayesian analysis, which makes use of the Bayes factor, is an alternative, but less popular method. The goal of this work is to present a brief comparison of the frequentist perspective based on the *p-value* (NHST) and the Bayesian perspective based on the Bayes factor. A few examples are given to illustrate some of the differences between each concept and its own characteristics. Readers are intended to be able to perceive some of those differences, their advantages and disadvantages, as well as their limitations, so that both techniques can be applied with discretion and criticism, enabling better decisions to be made in a range of circumstances.

**Keywords**: Hypothesis testing, NHST, p-value, Bayes factor, Bayes theorem, decision-making.
**AMS Subject Classification:** 62F03, 62G10, 62F15.

[†]**Corresponding author:** Mário, Basto, Higher School of Technology Polytechnic Institute of Cávado and Ave, Campus do IPCA, Barcelos, Tel.: +351-253802260, e-mail: *mbasto@ipca.pt*

*Received: 1 December 2022; Revised: 31 January 2023; Accepted: 13 February 2023;*
*Published: 13 April 2023.*

## 1 Introduction

Correct data-driven decision-making is critical in a wide range of professional and scientific fields. Hypothesis testing is one of the statistical procedures used to achieve this goal. However, there is no single solution that fits all cases (Gigerenzer et al., 2004). Often, descriptive statistics and exploratory data analysis are enough, and there is no need for hypothesis testing. But when resourcing on it, one has to be able to interpret the results correctly. Many people use the frequentist technique based on the *p-value* to test hypotheses, whereas others prefer the Bayesian approach based on the Bayes factor. Understanding the strengths and weaknesses of each approach allows one to make more accurate decisions.

The purpose of this paper is not to provide a comprehensive treatment of the subject or a tutorial on it, but rather to provide some examples to highlight some advantages and limitations of the null hypothesis significance testing (NHST) and the Bayesian approach to data testing, that will, hopefully, encourage researchers and professionals to question some of the results reached through their analysis, in order to ensure that proper conclusions are drawn based on the data.

A brief review of the NHST and the Bayesian approach is described herein. There are a number of other publications available that provide a more thorough review and discussion

[*] **How to cite (APA)**: Basto, M., Abreu, T., Gonçalves, R. & Pereira, J.M. (2023). Example-based overview to statistical inference. *Advanced Mathematical Models & Applications*, *8*(1), 5-13.

(Perezgonzalez, 2015; Malone & Coyne, 2020; Goodman, 1999a,b; Bayarri & Berger, 2004; Gibson, 2021; Lakens, 2021; Schmalz et al., 2021; Muff et al., 2022; Wong et al., 2022).

## 1.1 Null Hypothesis Significance Testing (NHST)

The NHST is the most widely used method for testing hypotheses, and it is a hybrid of Fisher and Neyman-Pearson procedures. However, it is not well defined, and it may favor one approach over the other (Perezgonzalez, 2015). The conclusions are based on the *p-value*, which is the probability of getting data at least as extreme as the one seen if the null hypothesis $H_0$, is true. That is, $p\text{-}value = P(x + |H_0)$, where $x+$ denotes the data observed or more extreme data. The effect size (practical effect) and statistical power (probability of rejecting correctly the null hypothesis $H_0$) are optional and not required (although recommended) in NHST (Perezgonzalez, 2015).

NHST is frequently carried out after establishing a statistical null hypothesis of no effect. The null hypothesis, $H_0$, is then accepted or rejected by comparing the *p-value* with a threshold significance level, which is typically set at 5%. The result is considered significant if the *p-value* is less than the significance level, and the null hypothesis, $H_0$, is not accepted. A low *p-value* means that the data observed or more extreme data is uncommon under $H_0$. As described next, some drawbacks of the *p-value* are highlighted by the definition.

The *p-value* is a function of the randomly observed data under $H_0$. It is a random variable whose distribution under $H_0$ is uniform, and therefore cannot give evidence for $H_0$ or against $H_1$ (Johansson, 2011). It tells nothing about the data under $H_1$, and it doesn't tell one how likely either hypotheses is accurate (Goodman, 1999a). The results of a single experiment cannot be used to calculate the probability that a conclusion is correct. Additionally, a low *p-value* or a significant finding does not imply that the effect is helpful or noteworthy. The *p-value* does not provide information on the effect size or the precision of the effect (Cumming, 2011, 2014). Finally, it assesses evidence by accounting for data that are not observed (more extreme data).

## 1.2 Bayesian approach

The fundamental addition to the Bayesian statistical approach is the inclusion of prior beliefs before collecting the data. Initially, there must be a distribution of credibility of the parameter values, called the prior distribution. This is the most common disadvantage pointed to this framework. Bayesian inference comprises the reassignment of the credibility of the parameter values, consistent with the data collected, called the posterior distribution.

Evidence is assessed using the Bayes factor ($BF_{01}$), which was developed by Jeffreys (1961). It allows for the quantification of the relative comparison of the predictive performance of the data $x$ for the null hypothesis $H_0$ against that for the alternative hypothesis $H_1$. This alternative metric of the *p-value* is computed as the ratio of two marginal likelihoods, the likelihoods of the data under each of the two hypotheses. That is, $BF_{01} = f(x|H_0)/f(x|H_1)$. In other words, the Bayes factor quantifies the degree to which the data are more likely under one hypothesis versus another (Jeffreys, 1961; Goodman, 1999b; Ly et al., 2016; Etz & Vandekerckhove, 2018). According to Lee & Wagenmakers (2014), the Bayes factor $BF_{01}$ provides moderate evidence for $H_0$ if its value is above 3 and strong evidence if it is above 10, or moderate evidence for $H_1$ if its value is below 1/3 and strong evidence if it is below 1/10. In summary, the Bayesian factor can quantify evidence for both the null and alternative hypotheses, whereas the NHST's *p-value* can only quantify evidence against the null hypothesis.

On the other hand, the Bayes factor, for an alternative composite hypothesis, is particularly sensitive to the prior distribution. This is an issue, especially when there is little prior knowledge of the model. However, the lower bound for the Bayes factor $BF_{01}$, which is the minimum across all the odds of the data likelihood of $H_0$ to $H_1$ (minimum Bayes factor, MBF), is independent of the prior (Harvey, 2017) and evaluates the strongest evidence against the null hypothesis. It is calculated by concentrating the prior distribution of the alternative hypothesis at the maximum likelihood estimate of the data.

Unlike NHST, the Bayesian approach allows for the computation of the posterior probability of any of the hypotheses, $P(H|x)$, where $H$ is the hypothesis and $x$ is the data collected.

In a nutshell, Bayesian statistics use the Bayes rule to allow one to change prior beliefs in light of new information (data), so answering a crucial topic not addressed by the other approach: the probability of the hypotheses given the observed data. In addition, unlike the *p-value*, Bayesian inference never uses the more extreme data (data not observed).

## 2 Material and Methods

The following seven examples were chosen in order to meet the goals of this paper.

### 2.1 Example 1

Consider the following hypotheses:

$H_0 : John\ does\ not\ have\ disease\ A$
$H_1 : John\ has\ disease\ A$

Consider the information gathered by a diagnostic test, where:

$sensitivity = P(test\ positive|John\ has\ disease\ A) = 0.99$, and
$specificity = P(test\ negative|John\ does\ not\ have\ disease\ A) = 0.97$.

Furthermore, it is known that disease A has a prevalence of 1% for the risk group to which John belongs. The test turned out to be positive. The *p-value* is then given by:

$p\text{-}value = P(test\ positive|John\ does\ not\ have\ disease\ A) = 1 - specificity = 0.03$.

It is worth noting that the *p-value* ignores the test's sensitivity. This low *p-value* (below the generally used threshold of 5%) indicates that the sample data does not support the null hypothesis. As a result, evidence exists against $H_0$. That is all there is to it. It would be a mistake to conclude that the alternative hypothesis $H_1$, which states that John has disease A, is true. The *p-value* does not provide information about the veracity or probability of the hypotheses.

The Bayes factor suggests strong evidence for $H_1$, pointing in the same direction as the *p-value*:

$$BF_{01} = \frac{P(test\ positive|John\ does\ not\ have\ diseaseA)}{P(test\ positive|John\ has\ disease\ A)} = \frac{1-specificity}{sensitivity} = \frac{0.03}{0.99} = \frac{1}{33}.$$

The Bayesian technique, unlike the *p-value*, allows one to turn the evidence provided by the Bayes factor into a probability for the hypotheses. Calculating the posterior odds is as follows:

$$Odds(John\ does\ not\ have\ disease\ A|test\ positive) =$$
$$= BF_{01} \times Odds(John\ does\ not\ have\ disease\ A) = \frac{1}{33} \times \frac{0.99}{0.01} = 3.$$

Calculating the posterior probability of $H_0$ and $H_1$:

$$P(John\ does\ not\ have\ diseaseA|test\ positive) = \frac{Odds}{1+Odds} = \frac{3}{1+3} = 0.75;$$

$$P(John\ has\ disease\ A|test\ positive) = 1 - 0.75 = 0.25.$$

These same probabilities could be achieved by applying Bayes' theorem (Stone, 2013).

As can be seen, the probability that John has disease A increased from 1% to 25% after receiving the positive test result. However, despite a positive test, John has a higher probability of not having disease A. Despite the evidence against the null hypothesis, the rejection of the null hypothesis is not able to overcome the low prior probability of the alternative hypothesis.

If the test turns out to be negative, the *p-value* is one, indicating that there is no evidence that John has disease A, but it says nothing about the evidence that John does not have disease A:

$$p\text{-}value = P(test\ negative\ or\ positive|John\ does\ not\ have\ disease\ A) = 1.$$

The Bayes factor, on the other hand, reveals that the null hypothesis that John does not have disease A is quite strong:

$$BF_{01} = \frac{P(test\ negative|John\ does\ not\ have\ disease\ A)}{P(test\ negative|John\ has\ disease\ A)} = \frac{specificity}{1-sensitivity} = \frac{0.97}{0.01} = 97.$$

Furthermore, the posterior odds and posterior probability support this conclusion:

$$Odds(John\ does\ not\ have\ disease\ A|test\ negative) =$$
$$= BF_{01} \times Odds(John\ does\ not\ have\ disease\ A) = 97 \times \frac{0.99}{0.01} = 9603;$$

$$P(John\ does\ not\ have\ disease\ A|test\ negative) = \frac{Odds}{1+Odds} = \frac{9603}{1+9603} = 0.999896.$$

The *p-value* only provides information for the evidence against $H_0$, while the Bayes factor provides information for both $H_0$ and $H_1$. If the test turns out to be negative, the evidence is against $H_1$ and in favor of $H_0$, according to the Bayes approach.

## 2.2 Example 2

Consider the same hypotheses:

$$H_0 : John\ does\ not\ have\ disease\ A$$
$$H_1 : John\ has\ disease\ A$$

Assume that a diagnostic test is ineffective, meaning that the result is 95% negative and 5% positive, regardless of whether or not the patient has disease A. The *p-value* for a positive test in this scenario is:

$$p\text{-}value = P(test\ positive|John\ does\ not\ have\ disease\ A) = 0.05.$$

For a level of 5%, this is a significant result. As the test does not provide information about disease A, it does not make sense to state that the evidence points to John having disease A. The Bayes factor, on the other hand, is one, indicating that no evidence exists for or against any of the hypotheses:

$$BF_{01} = \frac{P(test\ positive|John\ does\ not\ have\ disease\ A)}{P(test\ positive|John\ has\ disease\ A)} = \frac{0.05}{0.05} = 1.$$

## 2.3 Example 3

Consider the following hypotheses for determining whether a coin is fair:

$H_0 : The\ coin\ is\ fair$
$H_1 : The\ coin\ is\ not\ fair$

A researcher flipped the coin 165 times to choose between the hypotheses. The random variable $X$, that represents the number of heads in 165 tosses, follows a binomial distribution under $H_0$, with a probability $\pi = 0.5$ for a single flip being the head. There were 65 heads and 100 tails in the experience. Computing the *p-value*:

$p\text{-}value = 2 \times P(\bar{X} \leq 65|\pi = 0.5) = 0.0079$

Because the *p-value* is 0.0079, a statistically significant result for a cut-off level of 0.01 and 0.05, the null hypothesis is rejected, and the coin is classified as not fair. The evidence is overwhelming against $H_0$, which states the coin is fair.

Applying the Bayesian binomial test implemented in the free software JASP (0.16.2), and assuming a uniform prior distribution, $BF_{01} = 0.250$, that is, the data are approximately 4 times more likely under $H_1$ than under $H_0$. The evidence supporting the alternative hypothesis is at best moderate. Assuming a beta distribution for the prior with parameters $\alpha = \beta = 10$, $BF_{01} = 0.108$, indicating that the data are roughly 9.3 times more likely under $H_1$ than under $H_0$, providing moderate to strong evidence in support of $H_1$. The prior used in a Bayesian analysis do have an impact on the evidence. It is essential to select an appropriate prior. However, as can be seen, even a very low *p-value* does not always imply strong evidence against the null hypothesis.

## 2.4 Example 4

Consider the following hypotheses for the mean of a normal population with a known standard deviation of 5.5:

$H_0 : \mu = 20$
$H_1 : \mu = 26$

A total of 40 individuals were gathered for the study. The sample mean was $\bar{x} = 23$. Computing the unilateral *p-value*:

$p\text{-}value = P(\bar{X} \geq 23|\mu = 20) = 0.00028.$

The *p-value* obtained of 0.00028 denotes very strong evidence against $H_0$, and therefore $\mu = 26$ must be the correct population mean value. The sample mean, on the other hand, is exactly halfway between the two hypothetical values defined in the hypotheses. Because the z statistic's distribution is symmetric, if the hypotheses are swapped, the *p-value* will remain the same, implying that $\mu = 20$ must be the correct value for the mean. This appears to make no sense whatsoever. Because the data is only ever compared to $H_0$ and never to $H_1$, this occurs.

On the other hand, the Bayes factor equals one for the original hypotheses (and for the swapped one), revealing no definite evidence for any of the hypotheses (being $f$ the normal probability density function):

$BF_{01} = \frac{f(23|\mu=20)}{f(23|\mu=26)} = 1.$

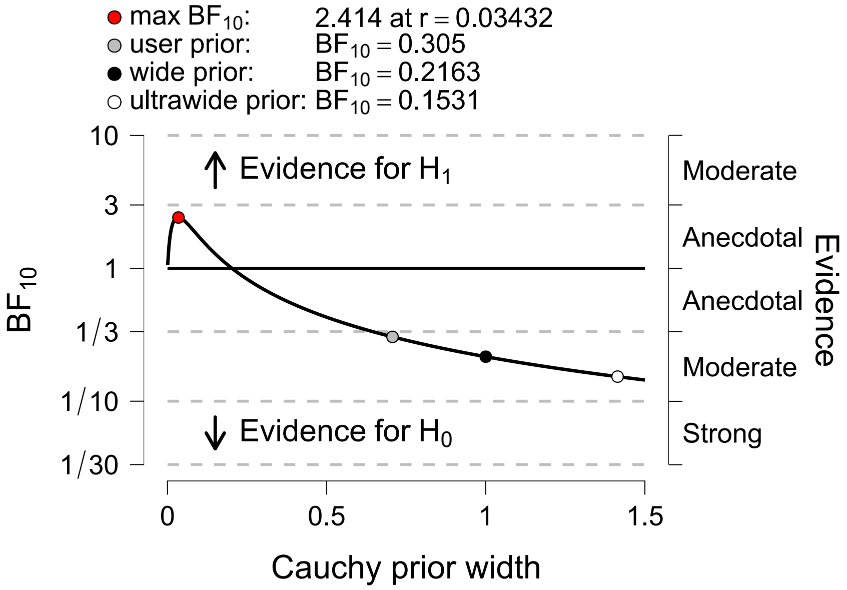

**Figure 1:** Factor Robustness Check in JASP 0.16.2. The inverse of Bayes factor $BF_{01}$, Bayes factor $BF_{10}$, is depicted.

## 2.5 Example 5

Let one take a look at these hypotheses for the mean of a normal population:

$$H_0 : \mu = 20$$
$$H_1 : \mu \neq 20.$$

A sample of $n = 2000$ was collected. The sample standard deviation was $s = 20$, and the sample mean $\bar{x} = 21$. The t statistic is 2.236, and the associated *p-value* is 0.025, which is a significant result at the 5% level:

$$p\text{-}value = 2 \times P(\bar{X} \geq 21 | \mu = 20) = 0.025.$$

By checking, in the free JASP software, 'Factor Robustness Check', the Bayes factor for different values of the Cauchy prior width shows that the evidence is at most moderate in favor of $H_0$ and only for minimal values of the Cauchy prior width shows evidence anecdotal against $H_0$ (Figure 1). The conclusions drawn differ significantly from those obtained using the *p-value*.

## 2.6 Example 6

Consider the following hypotheses for the proportion of a binomial population:

$$H_0 : \pi = 0.05$$
$$H_1 : \pi < 0.05$$

Consider the 5% significance threshold. A sample of $n = 50$ was collected. There was no success recorded, hence the $p - value$ is the lowest possible. The $p - value$ is given by $P(\hat{p} = 0)$ where $\hat{p}$ follows a binomial distribution with $n = 50$ and $\pi = 0.05$. Therefore, $p - value = 0.077$. It does not meet the 5% significance threshold and due to the fact that its value is the minimum achievable, the result can not be significant for this level. The sample size is too small to achieve a significant result, that is, the outcome is always known beforehand. With this small sample size, the test is worthless at the 5% significance level, because it cannot rule out the null hypothesis.

Applying the Bayesian binomial test implemented in the free software JASP and assuming a uniform prior distribution, the value achieved by the Bayes factor, $BF_{01} = 0.212$, indicates that the data is approximately 4.7 (1/0.212) times more likely under $H_1$ than under $H_0$. Contrary to what occurred with the $p - value$, the small sample size does not dismiss any of the hypotheses in advance.

### 2.7 Example 7

This last example was taken from Stone (2013). A patient displays spots on the body and face similar to those observed in smallpox and chickenpox. A test is carried out:

$H_0$ : *the patient has chickenpox*
$H_1$ : *the patient has smallpox*

Following Stone (2013), $P(spots|chickenpox) = 0.8$ and $P(spots|smallpox) = 0.9$, which means that the likelihood is higher for smallpox than for chickenpox.

Since *p-value* $= P(spots|chickenpox) = 0.8$, there is no evidence that the patient is infected with smallpox.

When the hypotheses are switched, the *p-value* is 0.9, and the conclusion is that there is no evidence that the patient has chickenpox.

There are no test conclusions in this case for the NHST test.

The same conclusions are achieved with the Bayes factor, which is given by:

$BF_{01} = \frac{P(spots|chickenpox)}{P(spots|smallpox)} = \frac{0.8}{0.9} = 0.889$
and, when the hypotheses are switched, by $BF_{01} = 1.125$

Because the Bayes factor is close to one, no choice is made between the hypotheses also.

However, when the real probability of the patient having each of the diseases is calculated, the picture changes. Given the prior probabilities:

$P(smallpox) = 0.0011$ and $P(chickenpox) = 0.1$, the posterior probabilities of the patient having each of the diseases are (they can be obtained using the prior odds and the Bayes factor or using the Bayes' theorem (Stone, 2013), as in the example of subsection 2.1):

$P(smallpox|spots) = 0.012$ and $P(chickenpox|spots) = 0.988$.

These values indicate that the patient has a very low chance of having smallpox and a huge one of having chickenpox.

## 3 Discussion and Conclusion

Because a low *p-value* says nothing about evidence against $H_1$, it can exaggerate the evidence against the null hypothesis $H_0$. The *p-value* also uses data that is not observed (more extreme data). It also has the disadvantage of not accounting for the magnitude of the effect. A small effect in a large sample size study and a large effect in a small sample size study can both have the same *p-value* (Goodman, 1999a). Finally, the *p-value* does not answer the main question of inductive inference: how probable is the research hypothesis $H_1$ (the implicit alternative hypothesis) in light of the data? (Page and Satake, 2003).

Prior parameters and effect sizes can be adjusted for any amount of fresh data in Bayesian analysis to give posterior uncertainty. Bayes factor compares and contrasts the predictive ability of two competing hypotheses, providing relative evidence for both. The main disadvantage is that prior assumptions and probabilities are required.

Both approaches, however, can be valuable and should be utilized with the necessary knowledge of their benefits and downsides, so that both procedures are used critically, allowing better decisions to be made.

## 4 Acknowledgement

The authors would like to thank the editor and the reviewers for their valuable suggestions that improved the quality of this work.

## References

Bayarri, M.J., Berger, J.O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, *19*, 58-80.

Cumming, G. (2011). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* Routledge, Taylor & Francis Group.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

Etz, A., Vandekerckhove, J. (2018). Introduction to bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5-34.

Gibson, E.W. (2021). The Role of p-Values in Judging the Strength of Evidence and Realistic Replication Expectations. *Statistics in Biopharmaceutical Research*, *13*, 6-18.

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In: D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences*, Thousand Oaks, CA:Sage, 391-408.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals of Internal Medicine*, *130*, 995-1004.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 2: The bayes factor. *Annals of Internal Medicine*, *130*, 1005-1013.

Harvey, C.R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, *72*, 1399-1440.

JASP 0.16.2 (2022). Jasp (version 0.16.2)[computer software]. `https://jasp-stats.org/`

Jeffreys, H. (1961). *Theory of probability.* Oxford University Press.

Johansson, T. (2011). Hail the impossible: p-values, evidence, and likelihood. *Scandinavian Journal of Psychology*, *52*, 113-125.

Lakens, D. (2021). The Practical Alternative to the p-Value. Is the Correctly Used p-Value. *Perspectives on Psychological Science*, *16*, 639-648.

Lee, M.D., Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Ly, A., Verhagen, J. & Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19-32.

Malone, H.E., Coyne, I. (2020). Complementing the P-value from null-hypothesis significance testing with a Bayes factor from null-hypothesis Bayesian testing. *Nurse Researcher, 28*(4).

Muff, S., Nilsen, E.B., O'Hara, R.B., & Nater, C.R. (2022). Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*, *37*, 203-210.

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*, 254-266.

Perezgonzalez, J.D. (2015). Fisher, neyman-pearson or nhst? a tutorial for teaching data testing. *Frontiers in Psychology*, *6*, 223.

Schmalz, X., Manresa, J.B. & Zhang, L. (2021). What is a Bayes factor?. *Psychological Methods*. ahead of print.

Stone, J.V. (2013). *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Sebtel Press.

Wong, T.K., Kiers, H. & Tendeiro, J. (2022). On the Potential Mismatch Between the Function of the Bayes Factor and Researchers' Expectations. *Collabra: Psychology*, *8*, 36357.