



## UvA-DARE (Digital Academic Repository)

### The Bayesian Approach to Audit Evidence

*Quantifying Statistical Evidence Using the Bayes Factor*

Derks, Koen; de Swart, Jacques; Wagenmakers, Eric Jan; Wetzels, Ruud

#### DOI

[10.2308/AJPT-2021-086](https://doi.org/10.2308/AJPT-2021-086)

#### Publication date

2025

#### Document Version

Final published version

#### Published in

Auditing

#### License

Article 25fa Dutch Copyright Act (<https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment>)

[Link to publication](#)

#### Citation for published version (APA):

Derks, K., de Swart, J., Wagenmakers, E. J., & Wetzels, R. (2025). The Bayesian Approach to Audit Evidence: Quantifying Statistical Evidence Using the Bayes Factor. *Auditing*, 44(1), 55-71. <https://doi.org/10.2308/AJPT-2021-086>

#### General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

#### Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

# The Bayesian Approach to Audit Evidence: Quantifying Statistical Evidence Using the Bayes Factor

**Koen Derks**

*Nyenrode Business University*

**Jacques de Swart**

*Nyenrode Business University*

*PwC Advisory*

**Eric-Jan Wagenmakers**

*University of Amsterdam*

**Ruud Wetzels**

*Nyenrode Business University*

*PwC Advisory*

**SUMMARY:** Statistical methods play an important role in auditors' analyses of their clients' data. A key component of the statistical approach to auditing is assessing the strength of evidence for or against a hypothesis. We argue that the frequentist statistical methods often used by auditors cannot provide the statistical evidence that audit standards advocate. In this article, we discuss an alternative approach that can provide this evidence: Bayesian inference. First, we explore the philosophical differences between frequentist and Bayesian inference. Second, we discuss misconceptions in the interpretation of frequentist statistical evidence. Finally, we show (as an alternative to the frequentist p-value) how the Bayes factor allows the auditor to obtain and interpret statistical evidence in line with audit standards. Thus, we contribute to audit theory and practice by showing how Bayesian inference can quantify audit evidence.

**Data Availability:** The data supporting the findings in this article are available in the OSF repository at <https://doi.org/10.17605/OSF.IO/WTN9G>.<sup>1</sup>

**Keywords:** audit evidence; analytical procedures; Bayes factor; substantive testing.

---

The authors of this publication have no conflicts of interest related to this research.

Koen Derks, Nyenrode Business University, Center of Accounting, Auditing & Control, Breukelen, The Netherlands; Jacques de Swart, Nyenrode Business University, Center of Accounting, Auditing & Control, Breukelen, The Netherlands, and PwC Advisory, Amsterdam, The Netherlands; Eric-Jan Wagenmakers, University of Amsterdam, Faculty of Social and Behavioural Sciences, Department of Psychology, Amsterdam, The Netherlands; Ruud Wetzels, Nyenrode Business University, Center of Accounting, Auditing & Control, Breukelen, The Netherlands, and PwC Advisory, Amsterdam, The Netherlands.

Editor's note: Accepted by Mark Cecchini, under the Senior Editorship of Jayanthi Krishnan.

*Submitted: July 2021*  
*Accepted: April 2024*  
*Early Access: May 2024*

---

<sup>1</sup> Supplemental materials accompanying this article are included in the OSF repository.

## I. INTRODUCTION

Auditors play a key role in preserving the integrity of the financial reporting of companies, (nonprofit) organizations, and governments. The objective of the auditor is to provide an opinion on whether the auditee's financial statements present a fair depiction of its financial condition and are formulated in compliance with a generally accepted financial reporting framework (IAASB 2021, ISA 200, paragraph 3; AICPA 2021, AU-C 200, paragraph .04; PCAOB 2020, AS 1001, paragraph .01). Audit standards mandate that this opinion is based on "reasonable assurance" (IAASB 2021, ISA 200, paragraph 5; AICPA 2021, AU-C 200, paragraph .06; PCAOB 2020, AS 1001, paragraph .02), which is obtained by collecting "sufficient appropriate" audit evidence (IAASB 2021, ISA 200, paragraph 5; AICPA 2021, AU-C 200, paragraph .06; PCAOB 2020, AS 1105, paragraph .04). The audit standards advocate that audit evidence consists of any information that can support or contradict management's assertions in the financial statements (IAASB 2021, ISA 500, paragraph A5; AICPA 2021, AU-C 500, paragraph 6; PCAOB 2020, AS 1105, paragraph 2). We argue that the frequentist statistical methods that are standard in practice (Stewart 2012) cannot provide the audit evidence that audit standards advocate.

In this article, we introduce a statistical approach that can quantify this evidence in line with auditing standards: Bayesian inference. The Bayesian approach primarily focuses on quantifying statistical evidence using the Bayes factor (Kass and Raftery 1995), and it has therefore been advocated as a suitable statistical framework for audit practice (Johnstone 2018; Stewart 2013). Because the Bayes factor can provide the statistical evidence that the audit standards advocate, it can enhance the way that auditors currently analyze and evaluate statistical evidence. The main contributions of this article are to explain how the Bayes factor enables the auditor to intuitively quantify statistical audit evidence, to demonstrate its advantages over standard frequentist methodology, to show how it can be applied in practice using a variety of relevant audit examples, and to enable practical use by supplying easy-to-use, free, and open-source software to calculate the Bayes factor.

Two types of audit evidence can be distinguished: nonstatistical and statistical. Nonstatistical audit evidence is collected from supervision, inquiry, or correspondence with the auditee (Bennett and Hatfield 2013; Yin 2020). Statistical audit evidence is collected from statistical analyses (Gillett and Srivastava 2000; van den Acker 2000), such as internal control testing (Li, Raman, Sun, and Yang 2020), analytical procedures (Appelbaum, Kogan, and Vasarhelyi 2018; Daroca and Holder 1985), or audit sampling (AICPA 2019; Dowling and Leech 2007). Regardless of the type of statistical analysis, the auditor requires a data sample to perform statistical inference on a certain characteristic of the auditee.

Where statistical sampling is applied, probability theory is required (IAASB 2021, ISA 530, paragraph 5g). There are two main schools of probability theory: frequentist and Bayesian (Wagenmakers, Lee, Lodewyckx, and Iverson 2008). In current practice, frequentism is the dominant methodology to obtain statistical audit evidence. Audit guides implicitly nudge auditors toward a frequentist approach to evaluate their samples using confidence intervals or p-values (AICPA 2019; Stewart 2012). However, frequentism has several well-known drawbacks relating to the ability of auditors to formulate an appropriate audit opinion (i.e., effectiveness) and the effort required to establish a sufficient basis for that opinion (i.e., efficiency). Most importantly, a frequentist hypothesis test does not give the auditor what the audit standards advocate: statistical evidence that can support or contradict their conclusions. In particular, its main decision-making tool, the p-value, strictly provides only indirect support against the conclusions made by the auditor. More specifically, the frequentist p-value has two main disadvantages. First, it is not possible to quantify evidence that can support the auditor's conclusions. Second, it is inefficient because the auditor is not allowed to stop gathering data when a certain evidential threshold is reached (e.g., when the p-value is smaller than 0.05) (Wagenmakers 2007). For these reasons, many scholars have raised concerns about the efficiency, transparency, and applicability of the frequentist methodology to audit practice (Beck, Solomon, and Tomassini 1985; Hubbard and Lindsay 2008; Johnstone 1986, 1990; Kim, Ahmed, and Ji 2018; Kinney 1975; Scott 1973).

In contrast to the frequentist methodology, Bayesian hypothesis testing using the Bayes factor as a measure of statistical evidence is more in line with audit practice due to three reasons. First, it can quantify evidence for and against hypotheses (Wagenmakers 2007). Second, it enables the incorporation of expert knowledge or other preexisting information into the statistical analysis (Corless 1972). Finally, it allows for sequential adding of information (Rouder 2014).

In a Bayesian approach, the rationale is that there is evidence for the null hypothesis over the alternative hypothesis if the data are more likely to occur under the null hypothesis (and *vice versa*). Comparing the likelihood of the data under two hypotheses allows for a direct assessment of their relative evidence. This enables auditors to obtain statistical evidence that can support or contradict their conclusions, which is often desired but impossible using a frequentist or nonstatistical approach.

Another key aspect of Bayesian inference is that the auditor needs to specify a so-called prior distribution. The prior distribution represents the preexisting information about an unknown parameter or hypothesis, which in turn is combined with the probability distribution of new data (i.e., the likelihood) to yield a posterior distribution. This philosophy fits well in an audit context because an audit is inherently a continuous process (Leslie 1984), and audit evidence is considered “cumulative in nature” (IAASB 2021, ISA 200, paragraph A30; AICPA 2021, AU-C 500, paragraph A3). However, the use of Bayesian statistics in audit practice is scarce, which is unfortunate considering the practical benefits that it can offer to auditors.

The rest of this paper is organized as follows. In Section II, we provide a theoretical introduction to statistical audit evidence and compare statistical audit evidence from the frequentist and Bayesian points of view. In Section III, we discuss the practical implications of the Bayesian approach for audit practice. In Section IV thereafter, to illustrate how the auditor can quantify evidence from a wide range of activities, we use the Bayes factor in an analysis of three typical audit questions. Section V presents our concluding comments.

## II. TWO APPROACHES TO STATISTICAL AUDIT EVIDENCE

Audit evidence is the subject of auditing standards ISA 500 (IAASB, international firms), AU-C 500 (AICPA, private firms), and AS 1105 (PCAOB, public firms). ISA 500 describes audit evidence as “[i]nformation used by the auditor in arriving at the conclusions on which the auditor’s opinion is based” (IAASB 2021, paragraph 5c) and states that audit evidence “comprises both information that supports and corroborates management’s assertions, and any information that contradicts such assertions” (IAASB 2021, paragraph A.1). Both AU-C 500 (AICPA 2021, paragraph 6) and AS 1105 (PCAOB 2020, paragraph 2) describe audit evidence in a similar manner. Clearly, auditing standards advocate that audit evidence should be able to support or contradict the auditor’s conclusions about management’s assertions in the financial statements.

### Hypothesis Testing in Audit Sampling

Audit sampling enables the auditor to obtain audit evidence to provide a reasonable basis on which to draw conclusions about the misstatement in the entire population (IAASB 2021, ISA 530, paragraph 5a). There are two approaches to conducting audit sampling: nonstatistical and statistical. In nonstatistical sampling, the auditor uses professional judgment to select and evaluate a representative sample from the population. An example could be an auditor deciding to review 50 invoices from representative suppliers in the sample, finding out that one of them contains a misstatement, and concluding that, on average, 2 percent of the invoices in the population are misstated. On the other hand, statistical sampling uses probability theory to select and evaluate a representative sample from the population. The use of probability theory enables the auditor to design an efficient sample, calculate the sampling risk involved, and project the sample outcomes to the population while considering uncertainty. For instance, based on their expectation of the misstatement rate and the desired sampling risk, the auditor can calculate that a minimum sample size of 50 invoices is required. If they select the 50 invoices randomly and find a similar rate of misstatement as in the nonstatistical sample, they conclude that, on average, 2 percent of the invoices in the population are misstated. However, the use of probability theory also enables the auditor to conclude that there is a 95 percent probability that the misstatement in the population falls between 0 and 10 percent. Hence, although both nonstatistical and statistical sampling can project the sample outcomes to the population, statistical sampling provides the additional benefit of estimating upper and lower bounds for the misstatement in the population.

To perform statistical sampling, the auditor needs to formulate a statistical hypothesis and perform a statistical hypothesis test. Because the auditor only inspects a sample of the population, the hypothesis cannot be evaluated with absolute certainty. However, because the auditor is required to obtain a reasonable assurance, they must evaluate the hypothesis to a level of certainty and must determine how much information is required to reach this level. Audit standards prescribe that the information from a sample is sufficient when it has reduced the sampling risk to an acceptably low level (IAASB 2021, ISA 530, paragraph 7; AICPA 2021, AU-C 530, paragraph .07; PCAOB 2020, AS 2315, paragraph .19). There are two types of sampling risk that can lead to an incorrect conclusion about financial misstatements (Elliott and Rogers 1972). First, a type I error ( $\alpha$ ) is the risk of incorrectly concluding that the population contains material misstatement. Second, a type II error ( $\beta$ ) is the risk of incorrectly concluding that the population does not contain material misstatement. According to auditing standards, auditors are primarily concerned with reducing the second type of risk,  $\beta$ , because it affects their effectiveness and their ability to provide an appropriate audit opinion (IAASB 2021, ISA 530, paragraph 5c(i); AICPA 2021, AU-C 530, paragraph .05a).

If the auditor engages in statistical sampling, they can quantify the sampling risk (IAASB 2021, ISA 530, paragraph 5 g(ii); AICPA 2021, AU-C 530, paragraph .05; PCAOB 2020, AS 2315, paragraph .46). Because the audit standards do not prescribe how this should be done, the auditor is free to choose which philosophy of probability is applied, a frequentist or a Bayesian philosophy. To illustrate the differences between these two philosophies, consider the following example. An auditor must assess the financial statements of a publicly traded company and wants to determine if no more than 3 percent of the auditee's purchase orders contain an incorrect signature. The auditor's one-sided null hypothesis stating that the misstatement in the purchase orders does not exceed the maximum tolerable error rate of 3 percent can be formulated as  $H_0: \theta \leq 0.03$ , whereas the one-sided alternative hypothesis reads  $H_1: \theta > 0.03$ .<sup>2</sup> In these hypotheses,  $\theta$  represents the proportion of purchase orders with an incorrect signature.

Because the purchase orders have not been audited before, the true value of  $\theta$  is unknown. Therefore, the auditor would like to determine the credibility of the hypothesis  $H_0$  by selecting several purchase orders and determining the correctness of the signature. After inspecting a sample of 99 purchase orders from the auditee's archives, the auditor finds that none of these purchase orders contain an incorrect signature.

### Frequentist Null Hypothesis Testing

The traditional method of analyzing these data is frequentist null hypothesis significance testing (NHST) (AICPA 2019; Elliott and Rogers 1972; Stewart 2012). In NHST, the auditor formalizes a null hypothesis that represents the minimal value of  $\theta$  at which their alternative hypothesis is not supported (Fisher 1934), which in the example is the hypothesis  $H_0: \theta \leq 0.03$ . Unfortunately, NHST only allows for quantifying evidence against the null hypothesis  $H_0$ . This means that the auditor cannot quantify evidence that can support the null hypothesis and cannot quantify evidence that supports the alternative hypothesis if an alternative hypothesis is defined.

When testing the null hypothesis, the auditor assumes that  $H_0$  is true and sets out to gather data to evaluate  $H_0$ . The rationale behind NHST is that increasingly stronger evidence will be obtained against  $H_0$  (the population is free of material misstatement) when the data become increasingly implausible under  $H_0$ . If sufficient evidence is obtained that contradicts  $H_0$ , it can be rejected with reasonable assurance. NHST allows the auditor to quantify the evidence against  $H_0$  using the p-value. This value is the probability of seeing the observed sample outcome or more extreme sample outcomes, assuming that  $H_0$  is true. Hence, the frequentist p-value enables the auditor to assess the risk of incorrectly rejecting  $H_0$ , which in this case is the  $\alpha$  risk. However, it is also possible to derive the  $\beta$  risk. To explain how these two risks can be computed in NHST, we will first show the calculation of the  $\alpha$  risk and subsequently show the calculation of the  $\beta$  risk.

To perform the statistical inference, the data from the sample need to be connected to the null hypothesis about  $\theta$  by means of a probability distribution (Lehmann and Romano 2006). For example, if the auditor assumes that the purchase orders in the sample are all independent observations and that the only parameter that exerts influence on the data is the probability of misstatement  $\theta$ , a binomial distribution can be applied (Johnstone 1990; Sorensen 1969).<sup>3</sup>

$$\text{Probability of } k \text{ incorrect signatures in } n \text{ purchase orders} = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \quad (1)$$

To determine when  $H_0: \theta \leq 0.03$  should be rejected for a given sample size, the auditor must calculate the maximum number of incorrect signatures that can be observed while the risk of incorrectly rejecting  $H_0$  is still sufficiently low. Suppose that the auditor has determined  $\alpha$ —the risk of incorrectly concluding that the population contains material misstatement—to be 5 percent. This threshold is referred to as the significance level. In this case, the rule for rejection of  $H_0$  is  $k \geq 7$  because if it is true that  $\theta = 0.03$ , then the probability of finding seven or more incorrect signatures in the sample of 99 purchase orders equals 2.98 percent (see Equation (1)). This probability is lower than the required significance level of 5 percent. This procedure for rejecting  $H_0$  can also be regarded as rejecting  $H_0$  if the p-value associated with the observed value of  $k$  is less than or equal to the significance level  $\alpha$ .

To continue the running example, we will calculate the p-value. In this case, calculating the p-value is trivial because the data are certain to occur under  $H_0$ . The p-value is the probability of finding  $k = 0$  to  $k = 99$  incorrect signatures in the sample, given that the purchase orders contain 3 percent misstatements, and equals 1 (Equation (2)). This means

<sup>2</sup> A different way to test these hypotheses is to include the value of the maximum tolerable error rate in the alternative hypothesis (see the file "Two approaches to NHST.pdf" in Derks, de Swart, Wagenmakers, and Wetzels (2025) for further details). We include the maximum tolerable error rate in the null hypothesis because this is the dominant approach in the auditing literature; see, for example, Elliott and Rogers (1972); Johnstone (1994); Martel-Escobar, Vázquez-Polo, and Hernández-Bastida (2018); and Edmonds, Miller, and Savage (2019).

<sup>3</sup> AICPA (2019) prescribes the binomial distribution for attribute sampling and the Poisson distribution for monetary unit sampling. However, the binomial distribution can also be used for monetary unit sampling.

that the data are certain to occur under  $H_0$ . Because the calculated p-value of 1 is higher than the significance level  $\alpha = 0.05$ , the auditor cannot reject the null hypothesis.

$$p = \sum_{k=0}^{99} \binom{99}{k} 0.03^k (1 - 0.03)^{99-k} = 1 \quad (2)$$

A default conclusion for a p-value larger than the significance level  $\alpha$  is to not reject—and thus maintain—the null hypothesis. However, the finding that the data contain no evidence against the null hypothesis does not imply that they contain evidence that supports the null hypothesis (Altman and Bland 1995; Goodman 2008; Keyzers, Gazzola, and Wagenmakers 2020). Because the p-value is solely a measure of evidence against the null hypothesis, it fails to address to which extent the sample provides support for the null hypothesis. Thus, based on the p-value, the auditor cannot quantify statistical evidence that can support their conclusion that no more than 3 percent of the auditee’s purchase orders contain an incorrect signature.

On the other hand, a p-value lower than the significance level  $\alpha$  generally leads the auditor to reject the null hypothesis and to accept the alternative hypothesis. However, in NHST, the p-value only concerns  $H_0$  and not an alternative hypothesis. Therefore, if the auditor uses this procedure to conclude that the data support the opposing hypothesis,  $H_1$ , they fall into a statistical trap (Berger and Sellke 1987; Berkson 1942; Wagenmakers 2007). The (im)plausibility of  $H_1$  is not considered in the computation of the p-value, and because the computation of the p-value is solely based on the evaluation of the data in light of  $H_0$ , it provides only an indirect argument for  $H_1$ . Hence, based on the p-value, the auditor cannot quantify statistical evidence that can support the conclusion that more than 3 percent of the auditee’s purchase orders contain an incorrect signature. Furthermore, they cannot statistically contradict the conclusion that no more than 3 percent of the auditee’s purchase orders contain an incorrect signature.

The auditor can also calculate the sampling risk  $\beta$ —the risk of incorrectly concluding that the population does not contain material misstatement. To calculate  $\beta$  for an alternative point hypothesis about the population misstatement, the auditor needs to make an assumption about  $\theta$ . Suppose that the auditor assumes that the population misstatement is equal to  $\theta = 0.04$ , which is slightly higher than the maximum tolerable error rate. When assuming the truth of this hypothesis, the risk of failing to reject  $H_0$  is the probability of finding an outcome that would yield a p-value above 5 percent (e.g.,  $k = 7$  misstatements would give  $p < 0.05$  and thus lead to rejection of  $H_0$ ). Hence, the sampling risk  $\beta$  can be calculated as the probability of finding  $k = 0$  to  $k = 6$  incorrect signatures in the sample under the binomial ( $k | n = 99, \theta = 0.04$ ) distribution and equals  $\beta = 0.90$ . This calculation shows that if 4 percent of the purchase orders in the population contain an incorrect signature, there is a 90 percent probability of incorrectly deciding that the null hypothesis should not be rejected. The reason that this probability is high is because the assumed population misstatement is close to the upper bound of the null hypothesis. The sampling risk  $\beta$  would be lower if the sample size would be higher or if the assumed population misstatement would be higher.

Because the auditing standards state that audit evidence consists of information that can support or contradict the auditor’s conclusions, the inability of the p-value to provide support for the null and alternative hypotheses makes it arguably unsuited for quantifying statistical evidence in an audit context. Unfortunately, this is not limited to audit sampling but applies to any procedure in which the auditor quantifies statistical audit evidence using the p-value. In the next section, we show that, by using a Bayesian hypothesis test, the auditor can quantify statistical evidence as advocated by the auditing standards via the Bayes factor.

### Bayesian Hypothesis Testing

In contrast to NHST, where the auditor’s evidence is solely based on the model for  $H_0$ , a Bayesian hypothesis test takes both hypotheses  $H_1: \theta > 0.03$  and  $H_0: \theta \leq 0.03$  into account. The driving force behind Bayesian inference is Bayes’ theorem (Jeffreys 1939), which stipulates how existing information about an event  $A$  can be updated using information from a new event  $B$  (Equation (3)).

$$p(A|B) = p(A) \times \frac{p(B|A)}{p(B)} \quad (3)$$

On a conceptual level, Bayes’ theorem embodies the fundamental principle in the audit that audit evidence is “cumulative in nature” (IAASB 2021, ISA 200, paragraph A.30; AICPA 2021, AU-C 500, paragraph A3) and that the auditor can therefore aggregate audit evidence over the audit.

In line with accumulating evidence, it is important to state the current level of information before an analysis is performed. This is reflected in the prior distribution, which needs to be defined for every aspect of the statistical model that

is to be estimated. In audit sampling, the prior distribution can be used to incorporate preexisting information about the possible values of the population misstatement  $\theta$  into the statistical model (Corless 1972). For example, the auditor's risk assessments relating to inherent risk and control risk can be incorporated into the prior distribution (Derks, de Swart, van Batenburg, Wagenmakers, and Wetzels 2021; Stewart 2013).

To use Bayes' theorem for hypothesis testing, the auditor must also quantify their preexisting information about the plausibility of the two competing hypotheses using so-called prior probabilities. The prior probability  $p(H_i)$  incorporates the auditor's preexisting information about the probability of the hypothesis  $H_i$  before seeing the data. This allows information for or against a given hypothesis to be taken into account before a sample is planned and selected. The ratio of prior probabilities is called the prior odds and is an indication of the relative plausibility of the hypotheses before analyzing the intended sample.

When performing the audit, new data  $y$  are observed, and the auditor aims to update the prior probability  $p(H_i)$  of the hypothesis  $H_i$  to a posterior probability  $p(H_i | y)$ . This is done via Bayes' rule, which allows the auditor to update their prior knowledge about the hypothesis  $p(H_i)$  with the evidence in the data for or against this hypothesis  $p(y | H_i)$ , resulting in the posterior probability of this hypothesis (see Equation (4)).

$$\underbrace{p(H_i | y)}_{\text{Posterior probability}} = \underbrace{p(H_i)}_{\text{Prior probability}} \times \underbrace{\frac{p(y | H_i)}{p(y)}}_{\text{Evidence}} \quad (4)$$

The posterior probability  $p(H_i | y)$  represents the probability that the hypothesis  $H_i$  is true, conditioned on the prior distribution and the sample data. For example, a posterior probability  $p(H_0 | y) = 0.95$  implies that, given the prior distribution and the sample data, there is a 95 percent probability of correctly deciding that  $H_0$  is true. Hence, the posterior probabilities can be intuitively related to the sampling risks  $\alpha$  and  $\beta$ . That is, when accepting  $H_0$ , the posterior probability  $p(H_1 | y)$  can be interpreted as the  $\beta$  risk. *Vice versa*, when rejecting  $H_0$  and accepting  $H_1$ , the posterior probability  $p(H_0 | y)$  can be interpreted as the  $\alpha$  risk.

However, because the auditor is interested in comparing the evidence for two hypotheses, they can use Bayes' theorem to obtain the ratio of posterior probabilities for  $H_0$  and  $H_1$ : the posterior odds. The posterior odds can be denoted as the product of the prior odds and the relative evidence for the hypotheses (Equation (5)).

$$\underbrace{\frac{p(H_0 | y)}{p(H_1 | y)}}_{\text{Posterior odds}} = \underbrace{\frac{p(H_0)}{p(H_1)}}_{\text{Prior odds}} \times \underbrace{\frac{p(y | H_0)}{p(y | H_1)}}_{\text{Relative evidence}} \quad (5)$$

Because the posterior odds depend on the prior odds as well as the information from the sample, and because it can be difficult to define the prior odds, it is common practice to quantify the relative evidence in the sample using the ratio of evidence for the two hypotheses. This ratio is called the Bayes factor, and it quantifies the change in prior to posterior odds brought about by the data (Kass and Raftery 1995). The Bayes factor is a direct comparison of the evidence for both hypotheses in the sample (Equation (6)).

$$BF_{01}(y) = \frac{p(y | H_0)}{p(y | H_1)} \quad (6)$$

Because the Bayes factor is a ratio, it can quantify evidence in both directions. For example, a Bayes factor in favor of  $H_0$  of 7 ( $BF_{01} = 7$ ) indicates that the sample outcomes are seven times more likely to occur under the null hypothesis  $H_0$  than under the alternative hypothesis  $H_1$ . Furthermore, because the following holds:  $BF_{10} = \frac{1}{BF_{01}} = \frac{1}{7}$ ,  $BF_{01} = 7$  also indicates that the sample outcomes are seven times less likely to occur under the alternative hypothesis  $H_1$  than under the null hypothesis  $H_0$ . This property of the Bayes factor fits well with the audit standards' description of audit evidence because it enables the auditor to quantify evidence that can support their hypotheses as well as evidence that can contradict their hypotheses.<sup>4</sup>

<sup>4</sup> Because of the ease of interpretation of the Bayes factor, it is rapidly being adopted in many areas of business and science, such as psychology (Heck et al. 2023; Ly, Verhagen, and Wagenmakers 2016), sociology (Bollen, Ray, Zavisca, and Harden 2012; Lynch and Bartlett 2019), and economics (Cipriani, Costantini, and Guarino 2012; Richard and Vecer 2021). Furthermore, Bayes factor calculations have been made very easy in many standard situations, such as the (partial) correlation test (Wetzels and Wagenmakers 2012), the t-test (Rouder et al. 2009; Wetzels et al. 2009), or the ANOVA (Rouder et al. 2012; Wetzels et al. 2012) and implementation in easy-to-use software such as JASP (Love et al. 2019).

In a similar fashion to the p-value, there exist pragmatic decision rules for which the Bayes factor represents sufficient evidence. A collection of labels has been proposed and reiterated in numerous academic articles (Jeffreys 1961; Wetzels et al. 2011; van Doorn et al. 2021). Table 1 displays these evidential thresholds, which auditors can use to interpret the strength of evidence provided by the Bayes factor. For instance,  $BF_{01} = 7$  implies moderate evidence in favor of  $H_0$  and at the same time implies moderate evidence against  $H_1$ . Similarly,  $BF_{01} = 20$  implies strong evidence in favor of  $H_0$  and at the same time strong evidence against  $H_1$ . Note that, although p-values and Bayes factors mostly agree about which hypothesis is supported by the data, they often disagree about the strength of this support (Wetzels et al. 2011).

To continue the example, a Bayesian auditor starts their audit sampling procedure by specifying a prior distribution  $p(\theta)$  that reflects their preexisting information about the misstatement in the population  $\theta$ . The two hypotheses  $H_0: \theta \leq 0.03$  and  $H_1: \theta > 0.03$  are defined as the range of the prior distribution that corresponds to the hypotheses' restrictions with respect to  $\theta$ . This means that the prior probability for  $H_0$  corresponds to the total probability under the prior distribution in the range  $[0, 0.03]$ . Similarly, the prior probability for  $H_1$  corresponds to the total probability under the prior distribution in the range  $[0.03, 1]$ .

If the auditor has assessed inherent risk and internal control risk according to the Audit Risk Model, they can incorporate this information into the prior distribution. For illustrative purposes, it is convenient to specify a uniform beta ( $\alpha = 1, \beta = 1$ ) prior distribution that represents negligible information about  $\theta$  (Stewart 2013). Suppose the auditor has assessed both inherent risk and internal control risk as "medium," which, according to their audit guide, translates into a reduction in the sample size of  $\Delta n = 33$ . This reduction can be seen as unseen samples that are assumed to be correct, which in turn can be incorporated in the prior distribution by setting the  $\beta$  parameter to  $1 + \Delta n = 34$  (Derks et al. 2021; Steele 1992). For the beta(1, 34) distribution, the prior odds in favor of  $H_0$  are  $\frac{0.645}{0.355} = 1.82$ .

After seeing the data consisting of  $n$  purchase orders of which  $k$  contain an incorrect signature, the prior distribution is updated by the binomial likelihood to a posterior distribution  $p(\theta | n, k)$  according to Bayes' theorem (Equation (7)).

$$\underbrace{p(\theta | n, k)}_{\text{Posterior}} = \underbrace{p(\theta)}_{\text{Prior}} \times \underbrace{\frac{p(k | n, \theta)}{p(y = k, n)}}_{\text{Evidence}} \tag{7}$$

In this example, the posterior distribution is the beta( $1 + 0 = 1, 34 + 99 = 133$ ) distribution. Like the prior distribution, the posterior distribution induces a probability for the occurrence of the hypotheses. The posterior odds in favor of  $H_0$  induced by the posterior distribution are  $\frac{0.983}{0.017} = 57.82$ . The posterior probability  $p(H_0 | y) = 0.983$  implies that there is a 98.3 percent probability that no more than 3 percent of the auditee's purchase orders contain an incorrect signature. Hence, when accepting  $H_0$ , there is a 98.3 percent probability that the auditor correctly judges that the population does not contain material misstatement and a 1.7 percent probability that the auditor incorrectly judges that the population does not contain material misstatement. The latter probability is sufficiently low to find  $H_0$  credible.

Because the prior odds and the posterior odds are known, the Bayes factor can be calculated by dividing the two. Thus, the Bayes factor in this example is  $BF_{01} = \frac{57.82}{1.82} \approx 31$ , which implies that the data are about 31 times more likely to occur under  $H_0$  than under  $H_1$ . This Bayes factor implies very strong evidence in favor of  $H_0$  (see Table 1).

**Comparison of Frequentist and Bayesian Conclusions**

Note that a frequentist analysis ( $p = 1 > \alpha$ ) only facilitates a statement about the (im)plausibility of the data (or data more extreme) under  $H_0$  and enables a conclusion of the auditor that  $H_0$  cannot be rejected. As mentioned in the previous section, using this p-value, the auditor cannot say that there is evidence in favor of  $H_0$ . The Bayes factor differs from the p-value in that it can quantify evidence directly in favor of  $H_0$  and that it provides an intuitive interpretation of the strength of evidence. That is, the Bayes factor  $BF_{01} \approx 31$  shows that  $H_0$  is more likely than  $H_1$  and that there is strong evidence in favor of  $H_0$  that no more than 3 percent of the auditee's purchase orders contain an incorrect signature. Moreover, because in a Bayesian procedure the auditor takes into account both  $H_0$  and  $H_1$ , this calculation shows that they are able to intuitively assess both the  $\alpha$  and  $\beta$  risk.

In sum, because the Bayes factor can quantify audit evidence in both directions, it is more in line with the philosophy of evidence described in the audit standards than the p-value. However, the Bayes factor is not only an attractive alternative to the p-value because of its intuitive theoretical interpretation, it also addresses some of the practical limitations that the p-value has. In the next section, we describe these limitations of the p-value in more detail and explain why the Bayes factor does not suffer from these limitations.

**TABLE 1**  
**Bayes Factor Labels Proposed by Jeffreys (1961)**

$BF_{01} = \frac{1}{BF_{10}}$	<u>Strength of Evidence</u>
$< \frac{1}{100}$	Extreme evidence for $H_1$
$\frac{1}{100} - \frac{1}{30}$	Very strong evidence for $H_1$
$\frac{1}{30} - \frac{1}{10}$	Strong evidence for $H_1$
$\frac{1}{10} - \frac{1}{3}$	Moderate evidence for $H_1$
$\frac{1}{3} - 1$	Anecdotal evidence for $H_1$
1	No evidence for $H_0$ or $H_1$
1–3	Anecdotal evidence for $H_0$
3 – 10	Moderate evidence for $H_0$
10 – 30	Strong evidence for $H_0$
30 – 100	Very strong evidence for $H_0$
>100	Extreme evidence for $H_0$

The subscript of the Bayes factor ( $BF$ ) indicates for which hypothesis the Bayes factor quantifies relative evidence. For instance,  $BF_{01} = 10$  indicates that the data are ten times more likely under the null hypothesis than under the alternative hypothesis, and  $BF_{10} = 10$  indicates that the data are ten times more likely under the alternative hypothesis than under the null hypothesis.

### III. PRACTICAL IMPLICATIONS

In this section, we illustrate that the use of frequentist NHST using the p-value limits auditors in their effectiveness and efficiency in quantifying audit evidence. Next, we show that the Bayes factor does not suffer from these limitations and that it is therefore an attractive alternative to the p-value. To illustrate which improvements the Bayes factor brings, we focus on the practical implications of two undesirable properties of the p-value: the p-value cannot provide evidence for the null hypothesis, and it does not allow for sequential sampling (Wasserstein and Lazar 2016; Rouder 2014; Wagenmakers 2007; Wagenmakers, Gronau, and Vandekerckhove 2019).

First, there are many scenarios (other than the sampling scenario in the previous section) where supporting  $H_0$  is the goal of the statistical analysis. Suppose that the auditor wants to support the null hypothesis that an auditee's inventory is valued fairly, wants to confirm the auditee's accounts receivable, or wants to confirm the auditee's sales transactions. As we have discussed in Section II, relying on the p-value makes supporting this null hypothesis impossible. However, by reporting a Bayes factor, the auditor can quantify evidence directly in favor of  $H_0$ , thereby removing this limitation. This makes it possible to compare the evidence in the data for both hypotheses, thus allowing the auditor to support or contradict their conclusion. In Section IV, we discuss two more examples in which the auditor wants to support  $H_0$ , one in which an auditor wants to support the conclusion that the data in the auditee's financial statements are subject to Benford's law (example 1) and another in which an auditor wants to support the conclusion that all taxable persons are treated fairly (example 3). We show in more detail how in these commonly occurring scenarios the p-value does not fit well with the audit question at hand because it cannot quantify support for the null hypothesis.<sup>5</sup>

Second, the p-value can lead to an inefficient audit when the auditor already has enough evidence to support a particular hypothesis, but due to the nature of frequentist hypothesis testing, they still need to perform the remainder of the planned work (Berger and Wolpert 1988; Lindley 1993). Suppose an auditor wants to obtain evidence to support the conclusion that a certain population contains misstatements lower than a certain threshold  $t$ . They define the null hypothesis as  $H_0: \theta \geq t$ . In this case, the auditor wants to sample until they can reject  $H_0$ . They have planned a sample size such that, when no misstatements are found, they can reject  $H_0$  with a sampling risk  $\alpha$  of 5 percent. As it turns out, the sample contains a single misstatement, which means that the auditor cannot reject  $H_0$ . If the auditor still wants to be able to reject  $H_0$  using the p-value, they will need to plan an extension for their sample. Because there is an increase in

<sup>5</sup> Example 2 illustrates a situation in which the p-value does not fit well with the audit question at hand because it cannot quantify support for the alternative hypothesis.

the sampling risk  $\alpha$  after looking at the data (Armitage, McPherson, and Rowe 1969; Wagenmakers 2007), one possible way to proceed is to plan a follow-up sample in which they adjust the maximum p-value. However, this practice generally leads to a substantial increase in the sample size. For example, most audit guides recommend inspecting an additional number of items equal to the initial sample (AICPA 2019, Appendix B). In this case, that would mean increasing the sample size from 99 to at least 198. This limits the auditor in their efficiency because, at this point, they are performing more work than necessary. The Bayes factor does not suffer from this limitation because a Bayesian analysis is not dependent on a sampling plan. That means that the auditor is allowed to monitor the evidence for a particular hypothesis and to stop data collection when enough evidence is obtained (Rouder 2014; Wagenmakers et al. 2019). From a Bayesian point of view “It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience” (Edwards, Lindman, and Savage 1963, 193). Therefore, the auditor can extend their sample from 99 to 156 (the sample size that they would have gotten when initially planning for one misstatement in the sample) or to any other  $n$  depending on the desired strength of evidence (i.e., the desired Bayes factor). In addition to being more efficient, a Bayesian sample size extension is arguably more intuitive and easier to explain for the auditor than a frequentist one.

#### IV. APPLYING THE BAYES FACTOR IN A MODERN AUDIT

To facilitate the use of the Bayes factor and illustrate the benefits of a Bayesian approach to audit evidence, we will now apply the Bayes factor to three typical audit questions.<sup>6</sup> In the first example, we apply Benford’s law to a financial dataset. In the second example, we analyze historical data from an auditee’s sales revenue to obtain evidence for a seasonal effect. In the final example, we analyze an auditee’s classification algorithm to determine if it exhibits bias toward certain groups of people.

##### Example 1: Assessing Benford’s Law

Benford’s law (Benford 1938) has been advocated as a simple and effective method for auditors to uncover potential data manipulation in financial statements (Durtschi, Hillison, and Pacini 2004), enterprise resource planning systems (Ma’arif, Mohd Satar, Abdul Jalal, and Samah 2020), or official information released by authorities (Wei and Vellwock 2020). Simply put, Benford’s law states that in naturally occurring collections of numbers, the leading digit is likely to be small. More concretely, a set of numbers is said to satisfy Benford’s law if the leading digits  $d \in \{1, \dots, 9\}$  occur with probability  $p(d) = \log_{10}\left(1 + \frac{1}{d}\right)$ .

Benford’s law can be used as an analytical procedure in an early stage of the audit (Nigrini and Mittermaier 1997). For example, small deviations from Benford’s law may suggest that the data have passed a reasonableness test, whereas large deviations may be a sign of possible data manipulation (Drake and Nigrini 2000). The goal of the analysis in this example is to determine how much evidence the data provide for the statement that the leading digits in a population of items follow Benford’s law.

##### Data

The data for this example come from the financial statements of the Sino Forest Corporation’s 2010 report (Nigrini 2012). For illustrative purposes, we will only analyze the leading digits of the recorded book values, but this procedure can be generalized to include the first two, or last, digits. The second and third columns of Table 2 display the observed (relative) frequencies of the leading digits in the data.

##### Frequentist Analysis

In the NHST framework, the auditor wants to test the null hypothesis that the first digits are distributed according to Benford’s law. An example application of this procedure is described by Varma and Khan (2012), who used Benford’s law to identify potential fraud in a similar population. The null hypothesis  $H_0: p(d) = \log_{10}\left(1 + \frac{1}{d}\right)$  is assessed by means of the p-value. The last column of Table 2 shows the expected relative frequencies  $p_d$  for the leading digits under Benford’s law.

<sup>6</sup> Please see the file “Bayes factor calculations.pdf” in Derks et al. (2025) for details about the derivations of various statistics and the calculations of the Bayes factors in this section.

**TABLE 2**  
**Observed and Expected (Relative) Frequencies**

Leading Digit ( $d$ )	Frequency	Relative Frequency	Expected Relative Frequency ( $p_d$ )
1	231	29.92%	30.10%
2	124	16.06%	17.61%
3	97	12.56%	12.49%
4	70	9.07%	9.69%
5	64	8.29%	7.92%
6	54	6.99%	6.69%
7	40	5.18%	5.80%
8	54	6.99%	5.12%
9	38	4.92%	4.58%

The last column displays the expected relative frequency  $p_d$  under Benford's law, where each digit  $d$  occurs with relative frequency  $p_d = \log_{10}(1 + \frac{1}{d})$ .

Using a Chi-square test ( $\chi^2 = 7.65$ ,  $df = 8$ ), the p-value for these data is 0.47. The interpretation of this p-value is that, assuming that the first digits are distributed according to Benford's law, there is a 47 percent probability that the auditor would have found the observed (or more extremely deviating) distribution of first digits in the dataset. In a standard fashion, the conclusion would be to not reject, and thus maintain,  $H_0$ .

### Bayesian Analysis

In a Bayesian analysis of Benford's law (Good 1967; Sarafoglou et al. 2023),  $H_0$  is compared, by means of the Bayes factor, with the alternative hypothesis  $H_1$ , which states that the first digits are not distributed according to Benford's law (i.e., the digit probabilities are free to vary). The prior probabilities for the hypotheses are set to be equal:  $p(H_0) = p(H_1) = 0.5$ . The prior distribution for  $H_1$  is assumed to be a Dirichlet( $\alpha_1, \alpha_2, \dots, \alpha_9$ ) distribution with all  $\alpha$  parameters set to 1. Note that the parameter  $\alpha_d$  of the Dirichlet distribution reflects the prior count for the digit  $d$  and can be adjusted to incorporate prior information into  $H_1$ .

The corresponding Bayes factor in favor of the null hypothesis is  $BF_{01} = 6,899,678$ , which implies that the data are 6,899,678 times more likely (extreme evidence) to have occurred under the hypothesis that the first digits are distributed according to Benford's law than under the hypothesis that they are not. Because the prior probabilities are set to be equal, the Bayes factor equals the posterior odds, which implies that the posterior probability for the null hypothesis can be deduced as  $p(H_0 | y) = \frac{BF_{01} \times p(H_0)}{BF_{01} \times p(H_0) + [1 - p(H_0)]} = 0.99$ . This means that, when accepting  $H_0$ , there is a 99 percent probability that the auditor correctly accepts  $H_0$  and a 1 percent probability that the auditor incorrectly accepts  $H_0$ .

### Comparison of Frequentist and Bayesian Conclusions

The p-value of 0.47 leads the auditor to not reject  $H_0$ . However, based on this low p-value, the auditor cannot say that the data contain evidence that supports the conclusion that the auditee's data follow Benford's law. The Bayes factor  $BF_{10} = 6,899,678$  facilitates the conclusion that the data show extreme evidence in favor of the conclusion that the auditee's data follow Benford's law.

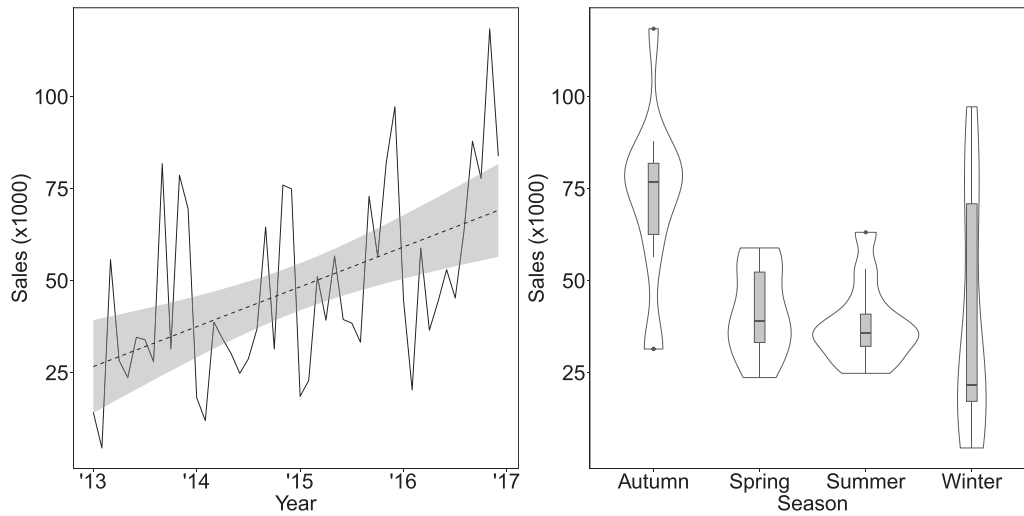
### Example 2: Uncovering Seasonal Patterns

We now turn to a situation where the auditor uses historical data in an analytical procedure. In particular, the auditor is concerned with the question of how much evidence there is that the sales of the auditee are influenced by seasonal factors. For example, yearly sales may be increasing, but sales might be higher in autumn than in other seasons. Hence, the auditor is interested in whether the sales of the auditee are subject to seasonal effects. In addition to a seasonality effect, the auditor wants to know to what extent the data support a difference in sales between each season.

### Data

The data for this example consist of the monthly sales of the auditee for the years 2013–2016 ( $n = 48$ ). These data are plotted over time and are categorized by season in Figure 1. The sales trend, depicted in the left panel of Figure 1,

**FIGURE 1**  
**Monthly Sales of the Auditee**



The sales trend over time, depicted in the left panel along with a 95 percent confidence interval, reveals an upward trajectory. In the right panel, box-violin plots categorize the sales by season, indicating a distinction between sales in the autumn and sales during the other three seasons. The lower and upper ends of the gray boxes correspond to the first and third quartiles of the sales, respectively, and the black line inside the box represents the median of the sales. The black lines extend from the first and third quartiles to the smallest and largest values no further than 1.5 times the interquartile range. The white shapes represent the distribution of the sales in each season.

reveals an upward trajectory. In the right panel of [Figure 1](#), box-violin plots classify sales by season, highlighting a noticeable difference between autumn sales and those in the remaining three seasons.

### ***Frequentist Analysis***

In the frequentist analysis, the null hypothesis  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$  is assessed via ANOVA by means of the p-value. To determine between which seasons there are differences in sales, *post hoc* tests are performed and assessed using Tukey's p-value, corrected for multiple comparisons.

The results of the ANOVA indicate a significant effect for season ( $R^2 = 0.30$ ,  $df = 3$ ,  $F = 6.42$ ,  $p < 0.01$ ). The interpretation of this p-value is that, assuming that there is no seasonal effect, there is less than 1 percent probability that the auditor would find the observed (or more extremely deviating) outcomes in the dataset. In addition, Tukey's *post hoc* tests indicate a significant difference in autumn sales compared with the other seasons. That is, the p-values for these comparisons are all smaller than 0.01 (the top three rows of column  $p_{\text{Tukey}}$  in [Table 3](#)).

### ***Bayesian Analysis***

In a Bayesian analysis, the null model of no effect is compared, by means of the Bayes factor, with an alternative model that incorporates the season as a predictive variable ([Rouder, Morey, Speckman, and Province 2012](#); [Wetzels, Grasman, and Wagenmakers 2012](#)). *Post hoc* tests between seasons are evaluated using the Bayes factor based on the default Bayesian t-test using a  $\text{Cauchy}(0, r = 1/\sqrt{2})$  prior ([Rouder, Speckman, Sun, Morey, and Iverson 2009](#); [Wetzels, Raaijmakers, Jakab, and Wagenmakers 2009](#)). In these *post hoc* tests, the posterior odds were adjusted for multiple comparisons by setting the prior probability of the null hypothesis at 0.5 for all comparisons ([Westfall, Johnson, and Utts 1997](#)).

The Bayes factor for the model that includes a seasonal effect over the model that does not is  $BF_{10} \approx 34$ . This Bayes factor implies that the observed data are 34 times more likely to have occurred under the hypothesis of a seasonal effect than under the hypothesis of no seasonal effect, which implies strong evidence in favor of a seasonal effect (see [Table 1](#)). To answer the question of how much more likely it is that the autumn sales differ from those in other seasons, the auditor must inspect the Bayes factors  $BF_{10}$  obtained from the individual comparisons in the last column of [Table 3](#). These

**TABLE 3**  
**Post Hoc Comparisons between Seasons**

		<u>Mean Difference</u>	<u>SE</u>	<u>t</u>	<u>p<sub>Tukey</sub></u>	<u>Prior Odds</u>	<u>Posterior Odds</u>	<u>BF<sub>10</sub></u>
Autumn	Spring	30,168.261	8,876.126	3.399	0.008	0.414	16.053	38.755
	Summer	33,339.340	8,876.126	3.756	0.003	0.414	41.542	100.291
	Winter	31,554.932	8,876.126	3.555	0.005	0.414	1.775	4.285
Spring	Summer	3,171.079	8,876.126	0.357	0.984	0.414	0.183	0.442
	Winter	1,386.672	8,876.126	0.156	0.999	0.414	0.156	0.376
Summer	Winter	-1,784.408	8,876.126	-0.201	0.997	0.414	0.156	0.378

The p-value obtained from Tukey's test for multiple comparisons (column  $p_{\text{Tukey}}$ ) indicates a significant difference in the sales in autumn compared with any other season. The Bayes factor (column  $BF_{10}$ ) indicates strong evidence to support this hypothesis for the comparison with the spring and the summer and moderate evidence in favor of this hypothesis for the comparison with the winter.

Bayes factors are roughly 39 and 100 for the spring and the summer, respectively, and thus indicate strong evidence for the statement that the autumn sales differ from those in spring and summer. However, the Bayes factor for the comparison with the winter is about 4, which indicates that the data contain only moderate evidence for a difference between the autumn sales and the winter sales.

#### Comparison of Frequentist and Bayesian Conclusions

The fact that  $p < 0.01$  leads the auditor to reject  $H_0$ . However, based on this low p-value, the auditor cannot say that there is evidence that supports a seasonal effect in the data. In contrast to the p-value, the auditor can use the Bayes factor of  $BF_{10} \approx 34$  to conclude that the data contain strong evidence in favor of a seasonal effect.

#### Example 3: Determining Algorithmic Fairness

As a final example, we consider an increasingly relevant issue in the context of algorithmic auditing and artificial intelligence (AI). With the rapid growth of information systems that collect and mine customer data, an increasing portion of auditees' business decisions is being guided by AI. On April 21, 2021, the European Commission presented a proposal for a regulation concerning AI, the AI Act for short (European Commission 2021). One major focus of the AI Act is the classification of various types of AI systems according to the risks involved. One of the risks that has special attention is that application of AI might lead to unfair treatment and discrimination. Hence, it becomes increasingly important to verify that business decisions made with the aid of these algorithms are fair (Kearns, Neel, Roth, and Wu 2018).

For example, predictive algorithms must avoid exhibiting discriminatory biases toward features such as gender, race, or age. Suppose that the auditor works with an auditee in the banking industry that uses an algorithm to predict whether customers are going to default on a loan. Naturally, it is highly undesirable that, given that a customer is going to pay their loan, they are more likely to get classified by the algorithm as possibly defaulting on that loan because of their ethnicity. The following analytical procedure aims to test this algorithmic fairness with respect to ethnicity.

To illustrate this procedure, we focus on a single relatively simple criterion of algorithmic fairness. This criterion requires equality of false positive or false negative rates across all subgroups in the data (Hardt, Price, and Srebro 2016). In the context of this example, a false positive implies that a customer gets wrongly marked as a possible defaulter. A false negative, on the other hand, would mean that a customer will likely default on their loan, but no action will be taken by the bank as this customer is not identified by the algorithm. The algorithm may display bias if the false positive rate is higher for some ethnic groups than for others. Statistically, this implies that the false positive rate should be equal across ethnic groups (i.e., the algorithm's prediction is independent of a customer's ethnicity). To find out how much evidence there is for this hypothesis, we describe one possible analysis that the auditor can use.

#### Data

We use a fictional benchmark dataset ( $n = 10,000$ ) from the field of credit risk prediction. The data contain information about a customer's ethnicity, a target variable that indicates defaulting behavior, and other financial information about the customer. The auditor applies the auditee's algorithm to this dataset to obtain the confusion matrix in Table 4.

**TABLE 4**  
**Confusion Matrix**

Group	Observed	Predicted		Total
		Defaulted	Paid	
Asian	Paid	97	826	923
	Defaulted	11	66	77
	Total	108	892	1,000
African	Paid	167	1,678	1,845
	Defaulted	12	143	155
	Total	179	1,821	2,000
Hispanic	Paid	195	1,648	1,843
	Defaulted	17	140	157
	Total	212	1,788	2,000
Caucasian	Paid	477	4,137	4,614
	Defaulted	55	331	386
	Total	532	4,468	5,000
Total	Paid	936	8,289	9,225
	Defaulted	95	680	775
	Total	1,031	8,969	10,000

The confusion matrix summarizes the quality of the predictions made by the auditee’s algorithm regarding customer defaults versus payments categorized by ethnic group (first column). The row totals show the totals for the observed classes in the benchmark data (default or pay), and the column totals show the predictions of the auditee’s algorithm. The four remaining cells show the number of true positives (bottom left), true negatives (top right), false positives (top left), and false negatives (bottom right).

The confusion matrix summarizes the quality of the predictions made by the auditee’s algorithm regarding customer defaults versus payments, categorized by ethnic group (first column). For each ethnic group, there are four potential combinations of predicted and observed classes. The algorithm can make a correct prediction in two ways: it can correctly predict a defaulting customer (a true positive) or it can correctly predict a paying customer (a true negative). However, the algorithm can also make an incorrect prediction in two ways: it can incorrectly predict a defaulting customer (a false positive) or it can incorrectly predict a paying customer (a false negative).

Table 4 shows the confusion matrices for all ethnic groups. The row totals show the totals for the observed classes in the benchmark data (default or pay), and the column totals show the predictions of the auditee’s algorithm. The four remaining cells show the number of true positives (bottom left), true negatives (top right), false positives (top left), and false negatives (bottom right). For example, within the group of Asian customers, the predictions of the auditee’s algorithm contained 97 false positives, 11 true positives, 826 true negatives, and 66 false negatives. The auditor calculates the false positive rate  $p_i$  for each ethnic group as the number of false positives divided by the number of false positives plus the number of true negatives (i.e., the number of accurately predicted paying customers).

**Frequentist Analysis**

In the frequentist analysis, the null hypothesis of independence  $H_0: p_1 = p_2 = p_3 = p_4$  is tested using a Chi-square test and is assessed by means of the p-value. The false positive rates  $p_1, p_2, p_3,$  and  $p_4$  are 0.1, 0.09, 0.11, and 0.1, respectively. Using a Chi-square test ( $\chi^2 = 3.13, df = 3$ ), the p-value is 0.37. The interpretation of this p-value is that, assuming that the false positive rate is equal across all ethnic groups, there is a 37 percent probability of observing these (or more extremely deviating) false positive rates.

**Bayesian Analysis**

In the Bayesian analysis, the null hypothesis of independence will be tested against the alternative hypothesis that the false positive rates are dependent on the ethnic group (Gunel and Dickey 1974; Jamil et al. 2017). The prior distribution for the alternative hypothesis is a Dirichlet( $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ ) distribution with all  $\alpha$  parameters set to 1.

The Bayes factor in favor of  $H_0$  is  $BF_{01} = 11,078$ , which implies that the data are 11,078 times more likely to have occurred under the hypothesis that the false positive rates are equal across ethnic groups than under the hypothesis that they are not. Using this Bayes factor, the auditor can quantify evidence in favor of the null hypothesis and support the statement that the false positive rates are equal across groups.

### *Comparison of Frequentist and Bayesian Conclusions*

The p-value of 0.37 leads the auditor to not reject  $H_0$ . However, based on this p-value, the auditor cannot say that the data show evidence that supports algorithmic fairness. In contrast to the p-value, the auditor can use the Bayes factor of  $BF_{01} = 11,078$  to conclude that the data contain extreme evidence in favor of algorithmic fairness.

## V. CONCLUDING COMMENTS

Audit evidence plays a crucial role for auditors in providing an appropriate opinion about the fairness of the auditee's financial statements. However, the frequentist method by which statistical audit evidence is currently often quantified in audits has raised legitimate concerns over the years. In this article, we have emphasized that a frequentist hypothesis test does not produce the type of evidence that the audit standards advocate and that the p-value does not fit well with the nature of audit questions. We have shown that a Bayesian hypothesis test can produce, in certain situations, a more fitting type of evidence for the auditor's conclusions about the financial statements and that it does not suffer from the same limitations as the p-value when it comes to effectiveness and efficiency. Because the Bayes factor can quantify evidence in both directions, the Bayesian approach to audit evidence is more in line with the audit standards than a frequentist hypothesis test. We therefore propose the Bayes factor as an addition to the auditor's statistical toolbox. Because the auditing standards explicitly call for evidence that can support or contradict the auditor's conclusions, we expect that the Bayes factor will enhance the way that auditors are able to quantify and evaluate statistical evidence.

Moreover, Bayesian inference provides auditors with the tools to aggregate audit evidence and therefore to statistically accumulate audit evidence over the course of an audit. This makes the Bayes factor a good fit for today's audit practice because it can facilitate the growing use of complex data analytics by the auditor and the auditee. As data will become more complex, and statistical analyses will become more prevalent, the auditor will require an intuitive framework to integrate, quantify, and interpret the information from these procedures. This will be the case especially if they are to meet the constant demand for a more efficient audit. Because the Bayesian framework provides the flexibility to incorporate many types of prior information into the statistical analysis, we believe that it will be more useful for the auditor in the long term than the current frequentist methods.

However, despite our arguments in favor of Bayesian hypothesis testing using the Bayes factor, it is not always practical to use this approach in favor of a frequentist hypothesis test. As discussed in [Section II](#), Bayes' theorem uses prior information in the form of the prior probabilities and the prior distribution on the parameters to perform inference. This means that, to get to the Bayes factor in practice, the auditor must think about how they incorporate preexisting information into the statistical model. Consequently, a frequentist hypothesis test can sometimes be more beneficial to an auditor if translating preexisting information into a prior distribution is difficult, expensive, or time-consuming. Although the Bayesian approach comes with advantages, such as being able to quantify evidence in favor and against the auditor's conclusions or the ability to engage in sequential testing without penalty, the auditor needs to decide whether the benefits of the Bayesian approach outweigh the costs of justifying that approach.

The sensitivity of the Bayes factor to the prior distribution is an avenue for further research. For many Bayesian hypothesis tests, default prior distributions exist and have been evaluated in a wide variety of settings ([Rouder et al. 2009](#); [Wetzels and Wagenmakers 2012](#); [Wetzels et al. 2012](#)). However, no default prior distribution exists or has been evaluated specifically in the context of audit sampling. Moreover, it remains to be investigated how auditors use and interpret Bayesian evidence in practice and if it increases the ease of interpretation of statistical results for auditors.

The examples shown in this manuscript show a selection of data-rich audit scenarios that the Bayes factor can be applied in, but, in principle, any statistical analysis can be performed in a Bayesian fashion. Most Bayesian analyses are easily accessible in a standardized format through open-source software packages (e.g., R) or graphical user interfaces (e.g., JASP). We have performed all statistical analyses in this manuscript using JASP (version 0.18.3; [Love et al. 2019](#)) and have included reproducible examples in [Derks et al. \(2025\)](#). Our proposition for a way forward is that next to their frequentist analyses, auditors perform Bayesian equivalents of these analyses to become acquainted with these techniques and to be able to compare the two measures of evidence (p-values and Bayes factors) in practice.

To conclude, we suggest the use of Bayesian inference in the audit because it fits well with the goals of the auditor. First, the Bayes factor embodies the audit standards' description of audit evidence and provides the auditor with a

measure of statistical evidence that can support or contradict their conclusions. Second, the theoretical foundations underlying the Bayesian framework have long been argued to be beneficial for the audit because they enable the auditor to quantify and aggregate evidence over the audit using the prior and posterior probabilities. In sum, Bayesian inference provides a fitting answer to the problems that today's auditors face.

## REFERENCES

- AICPA. 2019. *Audit Sampling: Audit Guide*. New York, NY: AICPA.
- AICPA. 2021. *Clarified Statements on Auditing Standards*. New York, NY: AICPA.
- Altman, D. G., and J. M. Bland. 1995. Statistics notes: Absence of evidence is not evidence of absence. *BMJ* 311 (7003): 485–485. <https://doi.org/10.1136/bmj.311.7003.485>
- Appelbaum, D. A., A. Kogan, and M. A. Vasarhelyi. 2018. Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics. *Journal of Accounting Literature* 40 (1): 83–101. <https://doi.org/10.1016/j.aclit.2018.01.001>
- Armitage, P., C. K. McPherson, and B. C. Rowe. 1969. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)* 132 (2): 235–244. <https://doi.org/10.2307/2343787>
- Beck, P. J., I. Solomon, and L. A. Tomassini. 1985. Subjective prior probability distributions and audit risk. *Journal of Accounting Research* 23 (1): 37–56. <https://doi.org/10.2307/2490906>
- Benford, F. 1938. The law of anomalous numbers. *Proceedings of the American Philosophical Society* 78 (4): 551–572.
- Bennett, G. B., and R. C. Hatfield. 2013. The effect of the social mismatch between staff auditors and client management on the collection of audit evidence. *The Accounting Review* 88 (1): 31–50. <https://doi.org/10.2308/accr-50286>
- Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association* 82 (397): 112–122. <https://doi.org/10.2307/2289131>
- Berger, J. O., and R. L. Wolpert. 1988. *The Likelihood Principle*, 2nd edition. Hayward, CA: Institute of Mathematical Statistics.
- Berkson, J. 1942. Tests of significance considered as evidence. *Journal of the American Statistical Association* 37 (219): 325–335. <https://doi.org/10.1080/01621459.1942.10501760>
- Bollen, K. A., S. Ray, J. Zavisca, and J. J. Harden. 2012. A comparison of Bayes factor approximation methods including two new methods. *Sociological Methods & Research* 41 (2): 294–324. <https://doi.org/10.1177/0049124112452393>
- Cipriani, M., R. Costantini, and A. Guarino. 2012. A Bayesian approach to experimental analysis: Trading in a laboratory financial market. *Review of Economic Design* 16 (2–3): 175–191. <https://doi.org/10.1007/s10058-012-0124-8>
- Corless, J. C. 1972. Assessing prior distributions for applying Bayesian statistics in auditing. *The Accounting Review* 47 (3): 556–566.
- Daroca, F. P., and W. W. Holder. 1985. The use of analytical procedures in review and audit engagements. *Auditing: A Journal of Practice and Theory* 4 (2): 80–92.
- Derks, K., J. de Swart, E.-J. Wagenmakers, and R. Wetzels. 2025. The Bayesian approach to audit evidence: Quantifying statistical evidence using the Bayes factor. OSF. <https://osf.io/wtn9g/>
- Derks, K., J. de Swart, P. van Batenburg, E.-J. Wagenmakers, and R. Wetzels. 2021. Priors in a Bayesian audit: How integration of existing information into the prior distribution can improve audit transparency and efficiency. *International Journal of Auditing* 25 (3): 621–636. <https://doi.org/10.1111/ijau.12240>
- Dowling, C., and S. Leech. 2007. Audit support systems and decision aids: Current practice and opportunities for future research. *International Journal of Accounting Information Systems* 8 (2): 92–116. <https://doi.org/10.1016/j.accinf.2007.04.001>
- Drake, P. D., and M. J. Nigrini. 2000. Computer assisted analytical procedures using Benford's law. *Journal of Accounting Education* 18 (2): 127–146. [https://doi.org/10.1016/S0748-5751\(00\)00008-7](https://doi.org/10.1016/S0748-5751(00)00008-7)
- Durtschi, C., W. Hillison, and C. Pacini. 2004. The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting* 5 (1): 17–34. <https://api.semanticscholar.org/CorpusID:17973012>
- Edmonds, M., T. Miller, and A. Savage. 2019. Accounts receivable: An audit simulation. *Journal of Accounting Education* 47: 75–92. <https://doi.org/10.1016/j.jaccedu.2019.04.001>
- Edwards, W., H. Lindman, and L. J. Savage. 1963. Bayesian statistical inference for psychological research. *Psychological Review* 70 (3): 193–242. <https://doi.org/10.1037/h0044139>
- Elliott, R. K., and J. R. Rogers. 1972. Relating statistical sampling to audit objectives. *Journal of Accountancy* 134 (1): 46–55.
- European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Brussels, Belgium: European Commission. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>
- Fisher, R. A. 1934. *Statistical Methods for Research Workers*, 5th edition. London, U.K.: Oliver & Boyd.
- Gillett, P. R., and R. P. Srivastava. 2000. Attribute sampling: A belief-function approach to statistical audit evidence. *Auditing: A Journal of Practice & Theory* 19 (1): 145–155. <https://doi.org/10.2308/aud.2000.19.1.145>

- Good, I. J. 1967. A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 29 (3): 399–418. <https://doi.org/10.1111/j.2517-6161.1967.tb00705.x>
- Goodman, S. 2008. A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology* 45 (3): 135–140. <https://doi.org/10.1053/j.seminhematol.2008.04.003>
- Gunel, E., and J. Dickey. 1974. Bayes factors for independence in contingency tables. *Biometrika* 61 (3): 545–557. <https://doi.org/10.1093/biomet/61.3.545>
- Hardt, M., E. Price, and N. Srebro. 2016. *Equality of opportunity in supervised learning*. Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, December 5–10.
- Heck, D. W., U. Boehm, F. Böing-Messing, P.-C. Bürkner, K. Derks, Z. Dienes, Q. Fu, X. Gu, D. Karimova, H. A. L. Kiers, et al. 2023. A review of applications of the Bayes factor in psychological research. *Psychological Methods* 28 (3): 558–579. <https://doi.org/10.1037/met0000454>
- Hubbard, R., and R. M. Lindsay. 2008. Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology* 18 (1): 69–88. <https://doi.org/10.1177/0959354307086923>
- International Auditing and Assurance Standards Board (IAASB). 2021. *Handbook of International Quality Control, Auditing Review, Other Assurance, and Related Services Pronouncements, Part I*. New York, NY: International Federation of Accountants.
- Jamil, T., A. Ly, R. D. Morey, J. Love, M. Marsman, and E.-J. Wagenmakers. 2017. Default “Gunel and Dickey” Bayes factors for contingency tables. *Behavior Research Methods* 49 (2): 638–652. <https://doi.org/10.3758/s13428-016-0739-8>
- Jeffreys, H. 1939. *Theory of Probability*, 1st edition. Oxford, U.K.: Oxford University Press.
- Jeffreys, H. 1961. *Theory of Probability*, 3rd edition. Oxford, U.K.: Oxford University Press.
- Johnstone, D. 2018. Accounting theory as a Bayesian discipline. *Foundations and Trends in Accounting* 13 (1–2): 1–266. <https://doi.org/10.1561/14000000056>
- Johnstone, D. J. 1986. Tests of significance in theory and practice. *Journal of the Royal Statistical Society: Series D (The Statistician)* 35 (5): 491–498.
- Johnstone, D. J. 1990. Sample size and the strength of evidence: A Bayesian interpretation of binomial tests of the information content of qualified audit reports. *Abacus* 26 (1): 17–35. <https://doi.org/10.1111/j.1467-6281.1990.tb00230.x>
- Johnstone, D. J. 1994. A statistical paradox in auditing. *Abacus* 30 (1): 44–49. <https://doi.org/10.1111/j.1467-6281.1994.tb00341.x>
- Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90 (430): 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Kearns, M., S. Neel, A. Roth, and Z. S. Wu. 2018. *Preventing fairness gerrymandering: Auditing and learning for subgroup fairness*. Proceedings of the 35<sup>th</sup> International Conference on Machine Learning, Stockholm, Sweden, July 10–15.
- Keyzers, C., V. Gazzola, and E.-J. Wagenmakers. 2020. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience* 23 (7): 788–799. <https://doi.org/10.1038/s41593-020-0660-4>
- Kim, J. H., K. Ahmed, and P. I. Ji. 2018. Significance testing in accounting research: A critical evaluation based on evidence. *Abacus* 54 (4): 524–546. <https://doi.org/10.1111/abac.12141>
- Kinney, W. R. 1975. Decision theory aspects of internal control system design/compliance and substantive tests. *Journal of Accounting Research* 13: 14–29. <https://doi.org/10.2307/2490473>
- Lehmann, E. L., and J. P. Romano. 2006. *Testing Statistical Hypotheses*. Springer Science & Business Media.
- Leslie, D. A. 1984. *Analysis of the audit framework focusing on inherent risk and the role of statistical sampling in compliance testing*. Proceedings of the 1984 Touche Ross University of Kansas Symposium of Auditing Problems, Lawrence, KA, May 17–18.
- Li, C., K. K. Raman, L. Sun, and R. Yang. 2020. The SOX 404 control audit and the effectiveness of additional audit effort in lowering the risk of financial misstatements. *Review of Quantitative Finance and Accounting* 54 (3): 981–1009. <https://doi.org/10.1007/s11156-019-00814-7>
- Lindley, D. V. 1993. The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics* 15 (1): 22–25. <https://doi.org/10.1111/j.1467-9639.1993.tb00252.x>
- Love, J., R. Selker, M. Marsman, T. Jamil, D. Dropmann, J. Verhagen, A. Ly, Q. F. Gronau, M. Smira, S. Epskamp, et al. 2019. JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software* 88 (2): 1–17. <https://doi.org/10.18637/jss.v088.i02>
- Ly, A., J. Verhagen, and E.-J. Wagenmakers. 2016. Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology* 72: 19–32. <https://doi.org/10.1016/j.jmp.2015.06.004>
- Lynch, S. M., and B. Bartlett. 2019. Bayesian statistics in sociology: Past, present, and future. *Annual Review of Sociology* 45 (1): 47–68. <https://doi.org/10.1146/annurev-soc-073018-022457>
- Ma’arif, M., N. Mohd Satar, A. A. Abdul Jalal, M. Samah. 2020. Detecting ERP data fraud using the first digits formula of Benford’s law. *Science International* 32 (4): 439–444. <http://www.sci-int.com/pdf/637323930858416050.edited.pdf>
- Martel-Escobar, M., F. J. Vázquez-Polo, and A. Hernández-Bastida. 2018. Bayesian inference in auditing with partial prior information using maximum entropy priors. *Entropy* 20 (12): 919. <https://doi.org/10.3390/e20120919>

- Nigrini, M. J. 2012. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, volume 586. John Wiley & Sons.
- Nigrini, M. J., and L. J. Mittermaier. 1997. The use of Benford's law as an aid in analytical procedures. *Auditing* 16 (2): 52–67.
- PCAOB. 2020. Auditing Standards. Washington, DC: PCAOB.
- Richard, M., and J. Vecer. 2021. Efficiency testing of prediction markets: Martingale approach, likelihood ratio and Bayes factor analysis. *Risks* 9 (2): 31. <https://doi.org/10.3390/risks9020031>
- Rouder, J. N. 2014. Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review* 21 (2): 301–308. <https://doi.org/10.3758/s13423-014-0595-4>
- Rouder, J. N., R. D. Morey, P. L. Speckman, and J. M. Province. 2012. Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology* 56 (5): 356–374. <https://doi.org/10.1016/j.jmp.2012.08.001>
- Rouder, J. N., P. L. Speckman, D. Sun, R. D. Morey, and G. J. Iverson. 2009. Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* 16 (2): 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Sarafoglou, A., J. M. Haaf, A. Ly, Q. F. Gronau, E.-J. Wagenmakers, and M. Marsman. 2023. Evaluating multinomial order restrictions with bridge sampling. *Psychological Methods* 28 (2): 322–338. <https://doi.org/10.1037/met0000411>
- Scott, W. R. 1973. A Bayesian approach to asset valuation and audit size. *Journal of Accounting Research* 11 (2): 304–330. <https://doi.org/10.2307/2490195>
- Sorensen, J. E. 1969. Bayesian analysis in auditing. *The Accounting Review* 44 (3): 555–561.
- Steele, A. 1992. *Audit Risk and Audit Evidence: The Bayesian Approach to Statistical Auditing*. London, U.K.: Academic Press.
- Stewart, T. R. 2012. *Technical Notes on the AICPA Audit Guide Audit Sampling*. New York, NY: American Institute of Certified Public Accountants.
- Stewart, T. R. 2013. *A Bayesian Audit Assurance Model with Application to the Component Materiality Problem in Group Audits*. Amsterdam, The Netherlands: Vrije Universiteit.
- van den Acker, C. 2000. Belief-function representation of statistical audit evidence. *International Journal of Intelligent Systems* 15 (4): 277–290. [https://doi.org/10.1002/\(SICI\)1098-111X\(200004\)15:4<277::AID-INT1>3.0.CO;2-P](https://doi.org/10.1002/(SICI)1098-111X(200004)15:4<277::AID-INT1>3.0.CO;2-P)
- van Doorn, J., D. van den Bergh, U. Böhm, F. Dablander, K. Derks, T. Draws, N. J. Evans, Q. F. Gronau, M. Hinne, Š. Kucharský, et al. 2021. The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review* 28 (3): 813–826. <https://doi.org/10.3758/s13423-020-01798-5>
- Varma, D. T., and D. A. Khan. 2012. Fraud detection in supply chain using Benford distribution. *International Journal of Research in Management* 5 (2): 90–96. <https://ssrn.com/abstract=3150290>
- Wagenmakers, E.-J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review* 14 (5): 779–804. <https://doi.org/10.3758/BF03194105>
- Wagenmakers, E.-J., Q. F. Gronau, and J. Vandekerckhove. 2019. Five Bayesian intuitions for the stopping rule principle. <https://doi.org/10.31234/osf.io/5ntkd>
- Wagenmakers, E.-J., M. D. Lee, T. Lodewyckx, and G. J. Iverson. 2008. Bayesian versus frequentist inference. In *Bayesian Evaluation of Informative Hypotheses*, edited by H. Hoijtink, I. Klugkist, and P. A. Boelen. New York, NY: Springer.
- Wasserstein, R. L., and N. A. Lazar. 2016. The ASA statement on p-values: Context, process, and purpose. *The American Statistician* 70 (2): 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Wei, A., and A. E. Vellwock. 2020. Is COVID-19 data reliable? A statistical analysis with Benford's law. <https://doi.org/10.13140/RG.2.2.31321.75365/1>
- Westfall, P. H., W. O. Johnson, and J. M. Utts. 1997. A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84 (2): 419–427. <https://doi.org/10.1093/biomet/84.2.419>
- Wetzels, R., and E.-J. Wagenmakers. 2012. A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review* 19 (6): 1057–1064. <https://doi.org/10.3758/s13423-012-0295-x>
- Wetzels, R., R. P. P. Grasman, and E.-J. Wagenmakers. 2012. A default Bayesian hypothesis test for ANOVA designs. *The American Statistician* 66 (2): 104–111. <https://doi.org/10.1080/00031305.2012.695956>
- Wetzels, R., J. G. Raaijmakers, E. Jakab, and E.-J. Wagenmakers. 2009. How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review* 16 (4): 752–760. <https://doi.org/10.3758/PBR.16.4.752>
- Wetzels, R., D. Matzke, M. D. Lee, J. N. Rouder, G. J. Iverson, and E.-J. Wagenmakers. 2011. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science* 6 (3): 291–298. <https://doi.org/10.1177/1745691611406923>
- Yin, X. 2020. Audit evidence concept, classification and collection techniques in China and the US. *Global Journal of Management and Business Research* 19 (3): 1–6. <https://journalofbusiness.org/index.php/GJMBR/article/view/2995>