International Journal of AI & Data Science
Review Volume 2, Issue 1, 2026

A Note on Hypothesis Testing with Random Sample Sizes and its Relationship to Bayes Factors

Oskar Skibskios[1]
Institute of Informatics, University of Warsaw, Poland

Article Info

Received: 08-09-2025 Revised:11-10-2025 Accepted:22-12-2025 Published:07-01-2026

Abstract: Practitioners frequently see frequentist and Bayesian hypothesis testing as "two separate worlds." Although there are theoretical connections, our objective here is to show a real-world scenario in which frequentist hypothesis testing requires caution and to illustrate a practical equivalency between frequentist and Bayesian testing. Specifically, selecting an α level in advance, like 0.05, and applying it to all possible sample sizes is inadmissible if the sample size is random (which is hardly unusual in practical problems where the sample size may be "all available experimental units"). To put it another way, a different overall procedure with the same overall type I error but more power can be found. This alternative process is based on Bayesian testing procedures, which is not by accident.

Key words: α-level, Bayes Factor, hypothesis testing, Jeffreys's paradox.

Introduction

It is commonly known that frequentist and Bayesian hypothesis testing differ significantly. This is known as Jeffrey's (or Lindley's) paradox, and it happens when a dataset leads to the null hypothesis being rejected using a standard frequentist hypothesis test, but the posterior probability of the null hypothesis is close to 1, leading to different conclusions depending on your statistical paradigm.

Theoretically, the causes of this discrepancy are widely recognised; see, for instance, Cox (1958), Lehmann (1958), Jeffreys (1961), Lindley (1971), and Seidenfeld et al. (1990). Here, we aim to give a simple, real-world example of how the paradox arises and some practical consequences of ignoring it, rather than offering a novel theoretical solution. The practical aspects of hypothesis testing are still being investigated in applied fields, as evidenced by the fact that foundational arguments are still being published in substantive journals (see, for instance, Miller's 2004 article). In our experience, most researchers are shocked to discover that it is not acceptable to plan on using a prespecified α level test for all sample sizes, such as α = 0.05, in an experiment where the sample size cannot be determined in advance (for example, the straightforward case of patient dropout, or an example where one samples all available units and is unsure in advance how many will be available). By adjusting α according to the sample size, a better procedure can be achieved overall. Keep in mind that this is not the same as sequential procedures, where the data determines the sample size. In this case, the sample size is independent of the data and may vary.

One of two sample sizes can be selected in our example, which is a straightforward experiment. Planning to perform a standard hypothesis test with α = 0.05, independent of sample size, is one possible analysis. Selecting distinct α levels for each of the two sample sizes is an alternate approach. We show that it is possible to select the different α levels in such a way that 1) the overall type I error rates for both procedures are equal, and 2) using the different α levels leads to better power. It is not by accident that the optimal selection of different α levels happens when the α levels are selected using Bayesian techniques, which reduce the α level as n increases.

In Section 2, the example is presented, some notation is defined, and an intuitive, easily visualised rule for avoiding the problems in the example is defined. The implications of this rule for testing hypotheses with normally distributed data are demonstrated in Section 3. We discuss the findings in Section 4.

An unpleasant example

For example, let X1,..., Xn N (μ, 1) and assume that we are testing H0: μ = 0 against HA: μ = 1.

Additionally, assume that n = 5 has a probability of 0.5 and n = 10 has a probability of 0.5.

Selecting α = 0.05 for any n that appears is one method. The power of the α = 0.05 test with n = 5 is 0.7228, whereas the power of the α = 0.05 test with n = 10 is 0.9354, according to some algebra. The overall procedure has a type I error probability of 0.05 and power (0.7228 + 0.9354)/2 = 0.8291 because it involves a 50-50 shot at each end.

An alternate method is to select α = 0.06 for n = 5 (power 0.7522) and α = 0.04 for n = 10 (power 0.9210). The total process has type I error probability (0.04 + 0.06)/2 = 0.05 and power (0.7522 + 0.9210)/2 = 0.8366, accounting for the random sample size. Keep in mind that this second method should be chosen because it has a higher power but the same overall type I error probability.

Naturally, one must wonder if there are other sets of α levels that perform even better, producing an even higher power but an overall type I error probability of 0.05. In the following, we present a straightforward criterion that makes it possible to see where the ideal α levels should be found. Using α = 0.0676 when n = 5 (power 0.7710) and α = 0.0324 when n = 10 (power 0.9059) yields the highest overall power (0.8384) while fixing the overall α level at 0.05, as we demonstrate in Section 3.1.

In order to create this example, α was first set at a constant level, then it was decreased for one sample size and increased by the same amount for the other sample size. This guarantees that the overall type I error rate of the new procedure is the same. The procedure will gain power if and only if the increase in power for one sample size is greater than the decrease in power for the other sample size, since increasing α always results in increasing power and decreasing α always results in decreasing power.

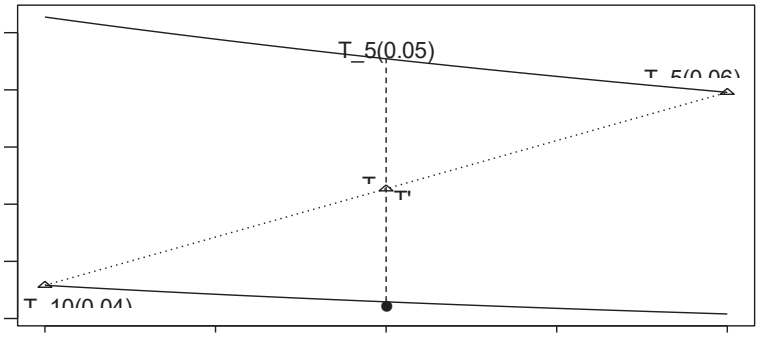
The derivative of the power function multiplied by the change in α accurately describes the change in power if the change in α is small. Therefore, the derivatives of the power function, which we explain below, can be used to find the best options for α. Although a general version of the Neyman-Pearson Lemma can be used to determine these results in an equivalent manner, our objective here is once again more pragmatic.

Some theory

Assume that X1,..., Xn f (x θ) and that we are comparing H0: θ = θ0 to H1: θ = θ1. Assume further that we observe a sample size of n = n1 with probability p and a sample size of n = n2 with probability 1 p (we address more than two possible values of n at the end of this section). Furthermore, assume that for every n and α, there is a uniformly most powerful (UMP) level α test Tn(α) with type I error α and type II error βn(α) (which will depend on f (x θ), θ0, and θ1). Observe that instead of using the power function, we have selected βn(α) as the type II error function. By doing this, the calculation avoids many (1 βn(α)) quantities. Assume further that Tn1 (αn1) is used if n = n1, and Tn2 (αn2) is used if n = n2. Thus, we may select different significance levels given n.

In our example, we used T5(0.05) and T10(0.05) (a constant α) in the first procedure and T5(0.06) and T10(0.04) in the second. The latter method was better, in our opinion.

The probabilities of type I and type II errors for the entire process can be calculated using this data. Type I errors have a probability of pαn1 + (1 – p)αn2, while type II errors have a probability of pβn1 (αn1) + (1 – p)βn2 (αn2). A pair (αn, αn) is considered inadmissible if there is another pair (α', α') such that at least one of the errors is strictly reduced and the probabilities of both errors are equal or reduced. A condition on (αn1, αn2) necessary for admissibility is shown in Theorem 2.1.



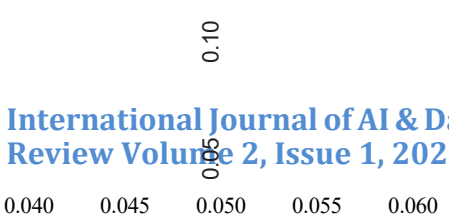


Figure 1: Example of an Inadmissible Test with Random Sample Sizes. The solid curves indicate the power of the test as α varies for $n = 5$ (top solid curve) and $n = 10$ (bottom solid curve). Test T uses $\alpha = 0.05$ when $n = 5$ and when $n = 10$ (marked with circles, with the average also denoted with a circle), while test T' uses $\alpha = 0.06$ when $n = 5$ and $\alpha = 0.04$ when $n = 10$ (marked with triangles, with the average also denoted with a triangle). Test T' has the same overall type I error probability, but has greater power.

Theorem 2.1 Let $X_1, \ldots, X_n \sim f(x|\theta)$ and suppose we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Furthermore assume the sample size n is random, where $n = n_1$ with probability p and $n = n_2$ with probability $1 - p$. Given an observed n, we test H_0 versus H_1 using the UMP level α_n test (assumed to exist). Let $\beta_n(\alpha_n)$ be the type II error probability for the UMP level α_n test, and let

probability of the entire procedure is the same and 2) the overall type II error probability of the entire procedure is reduced. Specifically

$$p\alpha_{n_1} + (1-p)\alpha_{n_2} = p\alpha' + (1-p)\alpha'$$

$$p\beta_{n_1}(\alpha'_{n_1}) + (1-p)\beta_{n_2}(\alpha'_{n_2}) < p\beta_{n_1}(\alpha_{n_1}) + (1-p)\beta_{n_2}(\alpha_{n_2})$$

Thus, to avoid this inadmissibility issue (given a constant type I error probability, one always want the smallest type II error probability), d_1 must equal d_2.

The proof is in the appendix.

Heuristically, Theorem 2.1 states that if the derivatives of the power functions are not identical, one can increase α for one value of n and decrease α for the other value of n such that 1) the size remains identical and 2) the power increases. You increase α in the direction of the larger derivative.

This may be seen visually in Figure 1. The solid curves in Figure 1 show $\beta_5(\alpha)$ (the top solid curve) and $\beta_{10}(\alpha)$ (the bottom solid curve) for α between 0.04 and 0.06. The two tests in the first procedure (constant α) are marked as labelled small circles on the plot. When the randomness in n is included, the overall error rates are (0.05, 0.1709) (the midpoint, or average, of the two tests), which is marked by a circle labelled T on the plot.

Note that the derivative of $\beta_5(\alpha)$ at $\alpha = 0.05$ is steeper than the derivative of $\beta_{10}(\alpha)$ at $\alpha = 0.05$. Theorem 2.1 therefore implies we can construct a dominating test by increasing α for $n = 5$ and decreasing α for $n = 10$. Essentially we move "downhill" for $n = 5$ more than we move "uphill" for $n = 10$ and thus achieve better power.

Suppose we choose $T_5(0.06)$ (with errors (0.06, 0.2478)) when $n = 5$ and $T_{10}(0.04)$ (with errors (0.04, 0.0790)) when $n = 10$. These tests are marked with triangles in the plot. Again taking the randomness of n into consideration, this procedure has overall error rates of (0.05, 0.1634), marked as a triangle labelled T' on the plot. Since the type I error rate is identical to T but the type II error rate is lower, T' dominates T and therefore T is inadmissible. We may do even better than these α values, as noted in the next section.

Note our initial discussion involves two values of n, but Theorem 2.1 can be directly applied to arbitrarily many possible sample sizes. Suppose, for any $n \geq 1$, p_n provides the probability the sample size will be n. The overall error probabilities are

$$\Pr(\text{type I}) = \sum_n p_n\alpha_n \quad \Pr(\text{type II}) = \sum_n p_n\beta_n(\alpha_n)$$

ne n (n_1 or n_2) and decrease α for the other n such that the type I error remains constant (by making the increase and decrease in each α equal) but the type II error decreases. Essentially one applies Theorem 2.1 to the terms involving n_1 and n_2 in the error probabilities, keeping the remaining α_n constant.

Since this result implies the derivatives of the type II error probabilities must be equal for *any* pair of possible sample sizes, the derivatives of the type II error probabilities must be equal for *all* possible sample sizes.

Hypotheses about Normal Means

Simple versus simple hypotheses

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, with σ^2 known. In this section we consider testing simple null and alternative hypotheses. Without loss of generality, we can consider testing $H_0 : \mu = 0$ against $H_1 : \mu = 1$, since with suitable linear transformations of the data these hypotheses are equivalent to $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$. Theorem 2.1 states that to acquire admissibility, we must choose the α_n se-quence such that the derivative of the type II error function $\beta_n(\alpha_n)$ is equal across all n. Theorem 2.1 does not state what this common value must be, but once α_n is specified for any single sample size, all the other α values are determined. In this section we derive how a common derivative translates into rejection regions for the test based on $\bar{X}_n$ and thus how selecting a rejection region for one n determines the rejection regions for all other n. On a related topic, we illustrate how this procedure relates to Bayesian hypothesis testing.

We begin by finding the derivative of the type II error function. Using stan-dard results (with φ being the standard normal pdf and Φ being the standard normal cdf), the UMP level α_n test rejects when $\bar{X}/(\sigma/\sqrt{n}) > z^*$ and has $1-\alpha_n$ type II error function

$$\beta_n(\alpha_n) = P\left(\frac{\bar{X}}{\sigma/\sqrt{n}} < z_{1-\alpha_n} \mid \mu = 1\right) = 1 - \Phi\left(\Phi^{-1}(1-\alpha_n) - \frac{1}{\sigma}\right).$$

Tak-ing the derivative of the type II error function to a constant, we find

$$C = d/d\alpha_n\, \beta_n(\alpha_n) = -\varphi\left(\Phi^{-1}(1-\alpha_n) - \frac{1}{\sigma}\right) \frac{n^{1/2}}{\varphi(\Phi^{-1}(1-\alpha_n))}$$

$$-C = \exp\left(-\frac{1}{2}\left(\Phi^{-1}(1-\alpha_n) - \frac{2n^{1/2}}{\sigma}\right)^2 + \frac{n}{2\sigma^2}\right).$$

Solving for α_n,

$$\alpha_n = 1 - \Phi\left(\frac{2\sigma^2 \log(-C) + n}{2\sigma n^{1/2}}\right). \qquad (3.1)$$

This choice of α_n is intuitive. After some simplification, one rejects the null hypothesis if

$$\bar{X}_n > \frac{\sigma^2 \log(-C)}{n} + \frac{1}{2} \qquad (3.2)$$

As n tends toward infinity, one rejects the null hypothesis if $\bar{X}_n$ is greater than 1/2. One chooses the hypothesis closest to the sample mean. This sequence of α_n values is not constant, but decreases exponen-tially fast. By Mill's ratio (Read 1985) equation (3.1) implies $2n = n + 2$ of course

This is the same rejection region as shown in equation (3.2), with $BF_0 = C$.
Of course, this sequence of α_n values is not unique. However, the central idea is that specifying α for one value of n specifies the value of α for all other values of n, otherwise the sequence would allow an inadmissible procedure. To finish our initial example from Section 2, note the distribution of n is known ($n = 5$ with probability 0.5, $n = 10$ with probability 0.5). If in addition to admissibility we add the requirement that the overall α level must be 0.05, we can use Equation (3.1) to determine α for all sample sizes. The added constraint that the overall α level is 0.05 requires that the chosen α_n follow the equation

$$0.5\ \alpha_5 + 0.5\ \alpha_{10} = 0.05$$

Since the only unknown in this equation is the constant C, we may solve for C numerically and find $\alpha_5 = 0.0676$ and $\alpha_{10} = 0.0324$. In general, if p_n is the probability the sample size is n, then setting the overall α level to α^* introduces the equation

$$\sum_n p_n \alpha_n = \alpha^*$$

Since all the α_n are determined by C, this is a single equation with a single unknown which yields a unique solution.

Possible extension to composite alternatives

The extension to composite alternative hypotheses requires use of a weight function, and is therefore more controversial. Suppose we are testing $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ using the most powerful level α_n test, $T_n(\alpha_n)$. In this case there are uncountably many values in the alternative hypothesis and the probability of a type II error is different for each value of the alternative. We account for this by using a weighted type II error function

$$\beta_n(\alpha) = \int_{H_1} \beta_n(\alpha_n, z) w(z) dz.$$

If the weight function w is an indicator function, $I_{[z=a]}$, then this reduces to the simple-simple case covered in Section 3.1. To compute the sequence α_n we again perform derive conditions under which the derivatives of the type II errors are equal.

Theorem 3.1 If w is integrable, continuous at 0, and $w(0) > 0$, then in order for $d/d\alpha_n\ \beta_n(\alpha_n)$ to remain constant as n increases, $\alpha_n = O\ (n \ln n)^{-1/2}$.

Proof

$$\frac{d}{d\alpha} \beta_n(\alpha_n) = \frac{d}{d\alpha} \int_0^\infty \Phi^{-1}\ (1 - \alpha_n) - \frac{z n^{1/2}}{\sigma} w(z) dz$$

$n \qquad n \quad 0$

By exchanging the derivative and the integral,

$$\frac{d}{d\alpha_n} \beta_n(\alpha_n) = \int_0^\infty \psi\ \Phi^{-1}(1 - \alpha_n) - \frac{\sigma}{\psi(\Phi^{-1}(1 - \alpha_n))} w(z) dz.$$

$$= \int_0^\infty \exp\left(-\frac{n}{2\sigma^2}\left(z - \frac{\sigma\Phi^{-1}(1 - \alpha_n)}{n^{1/2}}\right)^2\right) \exp\left(\frac{n}{2\sigma^2}\left(\frac{\sigma\Phi^{-1}(1 - \alpha_n)}{n^{1/2}}\right)^2\right) w(z) dz \quad (3.3)$$

If we let f_n be the density of a normal distribution with mean $\Phi^{-1}(1 - \alpha_n)\sigma/n^{1/2}$ and variance σ^2/n, then Equation (3.3) can be rewritten as

$$\frac{d}{} \beta_n(\alpha_n) = -\sigma(2\pi/n)^{1/2} \exp\left(\frac{n}{} \frac{\sigma\Phi^{-1}(1 - \alpha_n)}{}\right)^2 \int_\infty$$

$f_n(z)w(z)dz.$

$\frac{4nC^2}{w(0)^2\sigma^2 2\pi}$ ^{2}
so
$\alpha_n = O\ (n \log n)^{-1/2}$.
nlike simple alternative hypotheses, in the composite case the resulting α_n sequence does not exactly agree with testing based on Bayes Factors. However, the resulting asymptotic rates for α_n do agree (and the resulting α_n sequence is certainly not constant, as is typically done in practice). For testing the hypotheses $H_0 : \mu = 0$ against $H_A : \mu = 0$, the Schwarz approximation to the log of the Bayes factor is

$$\log BF_i = \frac{-1}{2\sigma^2} \sum_n (X - \bar{X})^2_i + \frac{1}{2\sigma^2} \sum_i (X - 0) - \frac{\log n}{2} + O(1).$$

This simplifies to

$$\log BF = \frac{n(\bar{X}_n)^2}{2\sigma^2} - \frac{\log n}{2} + O(1)$$

Choosing a reference value of the Bayes factor for performing the test, BF_0, we reject H_0 if

$$\frac{n(\bar{X}_n)^2}{2\sigma^2} - \frac{\log n}{2} > \log BF_0.$$

This simplifies to rejecting H_0 if

$$\bar{X}_n > \left(\frac{\sigma^2 \log n + 2\sigma^2 \log BF_0}{n}\right)^{1/2}$$

This rejection region corresponds to

$$\alpha_n = 1 - \Phi\left(\frac{n^{1/2}}{\sigma^2}\left(\frac{\sigma^2 \log n + 2\sigma^2 \log BF_0}{n}\right)^{1/2}\right)$$

$$= 1 - \Phi\left(\sqrt{\log(BF_0 n^{1/2})}\right) \qquad (3.5)$$

The α_n sequence corresponding to this rejection region is not the same as the α_n sequence shown in Equation (3.4). The asymptotic rate is the same, however. Applying Mill's ratio to Equation (3.5),

$$\alpha_n = \frac{1}{BF_0 n^{1/2}} \log(BF_0)^2 n^{-1/2},$$

so

$\alpha_n = O((n \ln n)^{-1/2})$.
which is the same rate shown for the admissible sequence.

Discussion

It should be noted that using the Bayesian paradigm of testing is not the same as suggesting that α decrease with n. Developing an admissible overall procedure with more experimental power than selecting a constant α level is the driving force behind the change.

What modifications might be made to the "standard" process of choosing an α level beforehand and applying that α level for all sample sizes is an obvious practical question. Unfortunately, we rarely know the sample size distribution in practice, so the only way to guarantee a given type I error probability is to choose a constant α. But this puts us in the uncomfortable position where we are aware that there is a better process. Choosing to test using Equation (3.1) or Equation (3.4), where the constant C may be selected to acquire a specific α for a specific n, is one way to completely avoid this.

There are additional intuitive advantages when the α level decreases. The argument that type I error is more crucial to prevent is used by many textbooks to support the process of selecting α beforehand rather than defining the power. Selecting $\alpha = 0.05$ as the "more important error" for moderate or large sample sizes leads to a type I error probability of 5% but a type II error probability that is negligible, obviously not controlling the "more important error." Take, for instance, the straightforward hypothesis testing problem of comparing $H0: \mu = 0$ to $H1: \mu = 1$ (assuming, for simplicity, $\sigma 2 = 1$). You decide to use $\alpha = 0.05$ for $n = 30$, which sets $\ln C = (4.2646)$ in Equation (3.1) and rejects for $X\approx 30 > 0.3578$. For $n = 100$, the necessary companion α is $2.3978\ 10{-}6$ (rejecting when $X\approx 100 > 0.4573$). Despite having a very small α, it still reaches power 1 ($2.87\ 10{-}8$), or nearly 100%.

Why choose $\alpha = 0.05$ at $n = 100$ and have such a high probability of type I error when both type I and type II error probabilities can be made negligible? One can obtain "the best of both worlds," where the probabilities of both errors are nearly zero for moderate or large sample sizes, by suitably reducing α.

References

D. Cox (1958). A few issues related to statistical inference. Mathematical Statistics Annals 29, 357-372.

H. Jeffreys (1961). Oxford University Press, Theory of Probability, Third Ed.

Raftery, A., and R. Kass (1995). Bayes Factors. American Statistical Society Journal 90, 773-795.

E. Lehmann (1958). Power and Significance Level. Mathematical Statistics Annals 29, 1167-1176.

D. Lindley (1971). A review of Bayesian statistics. Philadelphia: Society for Industrial and Applied Mathematics.

J. Miller (2004). Software technology experiments can benefit greatly from statistical significance testing. Systems and Software Journal, 73, 183-192.

C.B. Read (1985). The ratio of Mill. In Encyclopaedia of Statistics 5, 504-506 (Edited by S. Kotz, N.L. Johnson, and C.B. Read).

Kadane, J.B., Schervish, M.J., and Seidenfeld, T. (1990). choices without placing an order. 143–170 in Acting and Reflecting (Edited by W. Sieg).