Journal of Data Science **6**(2008), 75-87

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

Scott Berry[1] and Kert Viele[2]
[1]*Berry Consultants and* [2]*University of Kentucky*

Abstract: Frequentist and Bayesian hypothesis testing are often viewed as "two separate worlds" by practitioners. While theoretical relationships of course exist, our goal here is to demonstrate a practical example where one must be careful conducting frequentist hypothesis testing, and in that context illustrate a practical equivalence between Bayesian and frequentist testing. In particular, if the sample size is random (hardly unusual in practical problems where the sample size may be "all available experimental units"), then choosing an α level in advance such as 0.05 and using it for every possible sample size is inadmissible. In other words, one can find a different overall procedure which has the same overall type I error but greater power. Not coincidentally, this alternative procedure is based on Bayesian testing procedures.

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

1. Introduction

It is well known there are drastic differences between Bayesian and frequentist hypothesis testing. This is called Jeffrey's (or Lindley's) paradox, which occurs when a dataset results in rejecting the null hypothesis using a standard frequentist hypothesis test, but the posterior probability of the null hypothesis is near 1, creating opposite conclusions depending on your statistical paradigm.

Theoretically, the reasons for this difference are well known, see for example Cox (1958), Lehmann (1958), Jeffreys (1961), Lindley (1971), and Seidenfeld et. al (1990). Our goal here is not to present a new theoretical resolution for the paradox, but to provide a straightforward practical example of how the paradox occurs and some practical implications of ignoring it. Indeed, foundational arguments continue in substantive journals, see for example the article by Miller (2004), indicating that the practical aspects of hypothesis testing are still being explored in applied fields. While most theorists may consider it "well-known", in our experience most researchers are surprised to learn that in an experiment where the sample size cannot be determined in advance (for example the simple case of patient dropout, or an example where one samples all available units and is unsure in advance how many will be available) it is inadmissible to plan on using a prespecified α level test for all sample sizes such as $\alpha = 0.05$. One can achieve a better overall procedure by varying α with the sample size. Note this is a separate issue from sequential procedures, where the sample size is determined by the data. Here the sample size may vary, but does not depend on the data.

Our example involves a simple experiment where one of two sample sizes may be chosen. One possible analysis is to plan on conducting a standard hypothesis test using $\alpha = 0.05$ regardless of the sample size. An alternative method is to choose different α levels for each of the two sample sizes. We demonstrate that one can choose the differing α levels such that 1) the overall type I error rates for both procedures are the same, and 2) one achieves better power when using the differing α levels. Not coincidentally, the ideal choice of differing α levels occurs when the α levels are chosen based on Bayesian methods, which decrease the α level as n increases.

In Section 2 we show the example, define some notation, and define an intuitive, easy to visualize rule for avoiding the problems shown in the example. In Section 3 we illustrate the consequences of this rule for testing hypotheses with normally distributed data. In Section 4 we provide a discussion of the results.

2. Preliminaries

2.1 An unpleasant example

For illustration, let $X_1, \ldots, X_n \sim N(\mu, 1)$ and suppose we are testing $H_0 : \mu = 0$ against $H_A : \mu = 1$. Suppose also that $n = 5$ with probability 0.5 and $n = 10$ with probability 0.5.

One procedure is to choose $\alpha = 0.05$ for whichever n appears. Some algebra reveals that the power of the $\alpha = 0.05$ test for $n = 5$ is 0.7228 while the power for the $\alpha = 0.05$ test with $n = 10$ is 0.9354. The overall procedure involves a 50-50 shot at each end, so the overall procedure has type I error probability 0.05 and power $(0.7228 + 0.9354)/2 = 0.8291$.

An alternative procedure is to choose $\alpha = 0.06$ when $n = 5$ (resulting in power 0.7522) and $\alpha = 0.04$ when $n = 10$ (with power 0.9210). The entire procedure, taking into account the random sample size, has type I error probability $(0.04 + 0.06)/2 = 0.05$ and power $(0.7522 + 0.9210)/2 = 0.8366$. Note this second procedure has the same overall type I error probability but higher power, and thus should be preferred.

Of course, one must wonder whether there are other sets of α levels which perform even better in that they produce the 0.05 overall type I error probability but an even higher power. In what follows we show a simple criteria which allows one to visually see where the optimal α levels should occur. In Section 3.1 we use this criteria to illustrate that using $\alpha = 0.0676$ when $n = 5$ (power 0.7710) and $\alpha = 0.0324$ when $n = 10$ (power 0.9059) produces the highest overall power (0.8384) while fixing the overall α level at 0.05.

This example was created by beginning with a constant α level, and then decreasing α for one sample size and increasing α by the same amount for the other sample size. This assures the new procedure has the same overall type I error rate. Since increasing α always results in increasing power, and decreasing α always results in decreasing power, the overall procedure will gain power if and only if the increase in power for the one sample size is greater than the decrease in power for the other sample size.

If the change in α is small, the change in power is accurately described by the change in α multiplied by the derivative of the power function. Thus, the optimal choices for α can be found using the derivatives of the power function, which we describe below. Note that one can equivalently determine these results using a general form of the Neyman-Pearson Lemma, but again our goal here is more practical.

2.2 Some theory

Suppose $X_1, \ldots, X_n \sim f(x|\theta)$ and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Suppose further that with probability p we observe a sample size $n = n_1$ and with probability $1-p$ we observe a sample size of $n = n_2$ (we address more than two possible values of n at the end of this section). In addition, suppose for each n and α there exists a uniformly most powerful (UMP) level α test $T_n(\alpha)$, which has type I error α and type II error $\beta_n(\alpha)$ (which will depend on $f(x|\theta)$, θ_0, and θ_1). Note that we have chosen $\beta_n(\alpha)$ to be the type II error function rather than the power function. This avoids numerous $(1 - \beta_n(\alpha))$ quantities through the calculation. Suppose further that if $n = n_1$, we use $T_{n_1}(\alpha_{n_1})$, and if $n = n_2$, we use $T_{n_2}(\alpha_{n_2})$. Thus, given n, we potentially choose different significance levels.

In our example, our first procedure was to use $T_5(0.05)$ and $T_{10}(0.05)$ (a constant α) while our second procedure was to use $T_5(0.06)$ and $T_{10}(0.04)$. We found the latter procedure to be superior.

Using this information, we can compute the probabilities of type I and type II error for the entire procedure. The probability of type I error is $p\alpha_{n_1} + (1-p)\alpha_{n_2}$ and the probability of type II error is $p\beta_{n_1}(\alpha_{n_1}) + (1-p)\beta_{n_2}(\alpha_{n_2})$. We define a pair $(\alpha_{n_1}, \alpha_{n_2})$ to be inadmissible if there exists another pair $(\alpha'_{n_1}, \alpha'_{n_2})$ such that the probabilities of both errors are equal or reduced, and that at least one of the errors is strictly reduced. Theorem 2.1 illustrates a condition on $(\alpha_{n_1}, \alpha_{n_2})$ required for admissibility.

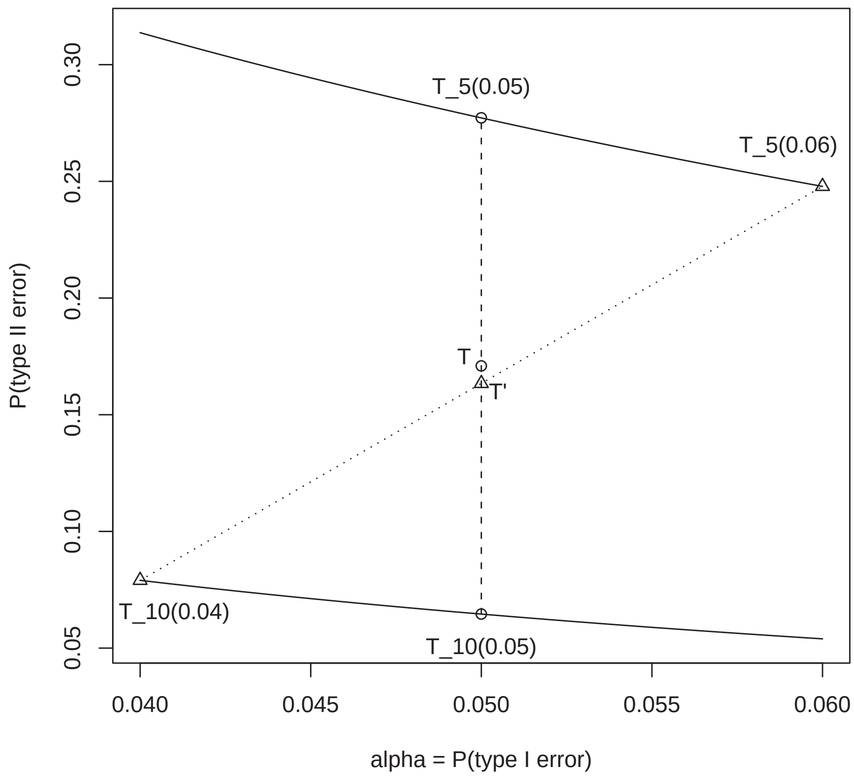

Figure 1: Example of an Inadmissible Test with Random Sample Sizes. The solid curves indicate the power of the test as α varies for $n = 5$ (top solid curve) and $n = 10$ (bottom solid curve). Test T uses $\alpha = 0.05$ when $n = 5$ and when $n = 10$ (marked with circles, with the average also denoted 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

$$d_1 = \frac{\mathrm{d}}{\mathrm{d}\alpha_{n_1}}\beta_{n_1}(\alpha_{n_1}) \qquad d_2 = \frac{\mathrm{d}}{\mathrm{d}\alpha_{n_2}}\beta_{n_2}(\alpha_{n_2})$$

If $d_1 \neq d_2$, then there exist α'_{n_1} and α'_{n_2} such that 1) the overall type I error probability of the entire procedure is the same and 2) the overall type II error probability of the entire procedure is reduced. Specifically

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

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

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

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

Suppose one chooses α levels α_n for each n but there exist n_1 and n_2 such that the α levels do not result in equal derivatives d_1 and d_2. Then one can increase α for one 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.

3. Hypotheses about Normal Means

3.1 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 sequence 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 standard results (with ϕ being the standard normal pdf and Φ being the standard normal cdf), the UMP level α_n test rejects when $\bar{X}/(\sigma/\sqrt{n}) > z^*_{1-\alpha_n}$ and has type II error function

$$\beta_n(\alpha_n) = P\left(\frac{\bar{X}}{\sigma/\sqrt{n}} < z^*_{1-\alpha_n} | \mu = 1\right) = 1 - \Phi\left(\Phi^{-1}(1-\alpha_n) - \frac{n^{1/2}}{\sigma}\right)$$

Setting the derivative of the type II error function to a constant, we find

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

which simplifies to

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

Solving for α_n,

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

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

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

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 of course not constant, but decreases exponentially fast. By Mill's ratio (Read 1985) equation (3.1) implies

$$\alpha_n \asymp \left(\frac{2\sigma n^{1/2}}{2\sigma^2\log(C) + n}\right)\exp\left\{-\frac{1}{2}\left[\frac{2\sigma^2\log(-C) + n}{2\sigma n^{1/2}}\right]^2\right\},$$

so

$$\alpha_n = O\left(n^{-1/2}\exp\left\{-\frac{n}{8\sigma^2}\right\}\right).$$

In fact, this sequence of α_n values corresponds exactly to tests based on Bayes Factors (Kass and Raftery 1995), with C determining the value of the Bayes Factor required to choose the alternative hypothesis. The Bayes factor against H_0 is

$$\begin{aligned} BF &= \frac{\Pr(X|\mu=1)}{\Pr(X|\mu=0)} \\ &= \exp\left\{\frac{-1}{2\sigma^2}\sum\left[(X_i-1)^2 - (X_i-0)^2\right]\right\} \\ &= \exp\left\{\frac{-n}{2\sigma^2} + \frac{2}{2\sigma^2}\sum X_i\right\}. \end{aligned}$$

If one chooses to reject the null hypothesis if BF is greater than a reference value BF_0, then one rejects if

$$\exp\left\{\frac{-n}{2\sigma^2} + \frac{2}{2\sigma^2}\sum X_i\right\} > BF_0,$$

which simplifies to rejecting if

$$\bar{X}_n > \frac{2\sigma^2 \ln BF_0 + n}{2n} = \frac{\sigma^2 \ln BF_0}{n} + \frac{1}{2}$$

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.

3.2 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, $\alpha_n = O\left((n \ln n)^{-1/2}\right)$.

Proof

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

By exchanging the derivative and the integral,

$$\frac{d}{d\alpha_n}\beta_n(\alpha_n) = \int_0^\infty \psi\left[\Phi^{-1}(1-\alpha_n) - \frac{z\sqrt{n}}{\sigma}\right]\left[\frac{-1}{\psi(\Phi^{-1}(1-\alpha_n))}\right] w(z)dz.$$

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

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

$$\frac{d}{d\alpha_n}\beta_n(\alpha_n) = -\sigma(2\pi/n)^{1/2}\exp\left\{\frac{n}{2\sigma^2}\left[\frac{\sigma\Phi^{-1}(1-\alpha_n)}{n^{1/2}}\right]^2\right\}\int_0^\infty f_n(z)w(z)dz.$$

Since w is continuous and positive at 0, then as n tends to infinity the integral converges to $\frac{1}{2}w(0)$. In order for the derivative to remain constant

$$C = -\sigma(2\pi/n)^{1/2}\exp\left\{\frac{n}{2\sigma^2}\left[\frac{\sigma\Phi^{-1}(1-\alpha_n)}{n^{1/2}}\right]^2\right\}\frac{1}{2}w(0),$$

which implies that

$$\alpha_n = 1 - \Phi\left[\left[2\log\left(\frac{-2Cn^{1/2}}{w(0)\sigma(2\pi)^{1/2}}\right)\right]^{\frac{1}{2}}\right]. \tag{3.4}$$

Applying Mills' ratio yields

$$\alpha_n \asymp \frac{w(0)\sigma}{-2Cn^{1/2}}\left[\log\left(\frac{4nC^2}{w(0)^2\sigma^2 2\pi}\right)\right]^{-\frac{1}{2}},$$

so

$$\alpha_n = O\left((n\log n)^{-1/2}\right).$$

Unlike 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 $H_A : \mu \neq 0$, the Schwarz approximation to the log of the Bayes factor is

$$\log BF = \frac{-1}{2\sigma^2} \sum (X_i - \bar{X}_n)^2 + \frac{1}{2\sigma^2} \sum (X_i - 0)^2 - \frac{\log n}{2} + O(1).$$

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

$$\begin{aligned} \alpha_n &= 1 - \Phi \left[\left(\frac{n}{\sigma^2} \frac{\sigma^2 \log n + 2\sigma^2 \log BF_0}{n} \right)^{\frac{1}{2}} \right] \\ &= 1 - \Phi \left[\left(2 \log (BF_0 n^{1/2}) \right)^{\frac{1}{2}} \right]. \end{aligned} \tag{3.5}$$

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),

$$\alpha_n \asymp \frac{1}{BF_0 n^{1/2}} \left[\log (BF_0)^2 n \right]^{-1/2},$$

so

$$\alpha_n = O((n \ln n)^{-1/2}).$$

which is the same rate shown for the admissible sequence.

4. Discussion

Note that the suggestion of decreasing α with n is not equivalent to adopting the Bayesian paradigm of testing. The change is motivated by creating an admissible overall procedure that has greater experimental power than choosing a constant α level.

An obvious practical question is what changes might be implemented from the "standard" procedure of selecting an α level in advance and using that α level for all sample sizes. Unfortunately, in practice we rarely know the distribution of the sample size, and thus except for choosing a constant α, we rarely have any other way of assuring a specified type I error probability. However, this leads us to the unpleasant situation where we know a better procedure exists. One way of avoiding this altogether is to choose to test based on Equation (3.1) or Equation (3.4), where the constant C may be chosen to acquire a specific α for a specific n.

A decreasing α level results in other intuitive benefits. Many textbooks justify the procedure of choosing α in advance, instead of specifying the power, under the reasoning that type I error is more important to avoid. For moderate or large sample sizes choosing $\alpha = 0.05$ as the "more important error" results in a probability of type I error of 5% but a type II error probability that is negligible, clearly not controlling the "more important error". For example, consider the simple hypothesis testing problem of testing $H_0 : \mu = 0$ against $H_1 : \mu = 1$ (assume for simplicity $\sigma^2 = 1$). You choose to use $\alpha = 0.05$ for $n = 30$, which rejects for $\bar{X}_{30} > 0.3578$ and thus sets $\ln C = (-4.2646)$ in Equation (3.1). The required companion α for $n = 100$ is $2.3978 * 10^{-6}$ (rejecting when $\bar{X}_{100} > 0.4573$). This is a very small α, but it still achieves power $1 - (2.87 * 10^{-8})$, almost 100%. One can make both type I and type II error probabilities negligible at $n = 100$, so why choose $\alpha = 0.05$ at $n = 100$ and have such a relatively high probability of type I error? By decreasing α appropriately, one can acquire "the best of both worlds", where the probabilities of both errors are close to 0 for moderate or large sample sizes.

Appendix

Proof of Theorem 2.1

Suppose $d_1 > d_2$ (the proof for the reverse inequality is analogous) and define $\bar{d} = (d_1 + d_2)/2$. Then, using the definition of derivative, for some constant $c > 0$ there must exist α'_{n_1} and α'_{n_2} such that

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

$$\beta_{n_1}(\alpha'_{n_1}) < \beta_{n_1}(\alpha_{n_1}) - \bar{d}(\alpha'_{n_1} - \alpha_{n_1}) \qquad \beta_{n_2}(\alpha'_{n_2}) < \beta_{n_2}(\alpha_{n_2}) + \bar{d}(\alpha_{n_2} - \alpha'_{n_2})$$

If such points didn't exist, it would contradict the limit involved in the definition of derivative. The size of the test using the pair $(\alpha'_{n_1}, \alpha'_{n_2})$ is

$$p\alpha'_{n_1} + (1-p)\alpha'_{n_2} = p\left[\frac{c(1-p)}{p} + \alpha_{n_1}\right] + (1-p)(\alpha_{n_2} - c) = p\alpha_{n_1} + (1-p)\alpha_{n_2}$$

Thus, the size of the test using the pair $(\alpha_{n_1}, \alpha_{n_2})$ is the same as the size of the test using the pair $(\alpha'_{n_1}, \alpha'_{n_2})$. In addition, the above equality establishes the relation

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

which in turn establishes

$$\begin{aligned} p\beta_{n_1}(\alpha'_{n_1}) + (1-p)\beta_{n_2}(\alpha'_{n_2}) &< p\left[\beta_{n_1}(\alpha_{n_1}) - \bar{d}(\alpha'_{n_1} - \alpha_{n_1})\right] \\ &\quad + (1-p)\left[\beta_{n_2}(\alpha_{n_2}) + \bar{d}(\alpha_{n_2} - \alpha'_{n_2})\right] \\ &= p\beta_{n_1}(\alpha_{n_1}) + (1-p)\beta_{n_2}(\alpha_{n_2}) \end{aligned}$$

Thus the pair $(\alpha'_{n_1}, \alpha'_{n_2})$ results in a test with smaller probability of type II error. Since the sizes were identical, this makes the test using the pair $(\alpha_{n_1}, \alpha_{n_2})$ inadmissible, and thus proves the theorem.

References

Cox, D. (1958). Some Problems Connected with Statistical Inference. *Annals of Mathematical Statistics* **29**, 357-372.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press.

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

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

Lindley, D. (1971). *Bayesian Statistics : A Review*. Society for Industrial and Applied Mathematics, Philadelphia.

Miller, J. (2004). Statistical Significance Testing: A Panacea for Software Technology Experiments. *Journal of Systems and Software* **73** 183-192.

Read, C.B. (1985). Mill's Ratio. In *Encyclopedia of Statistics* (Edited by S. Kotz, N.L. Johnson, and C.B. Read) **5**, 504-506.

Seidenfeld, T., Schervish, M.J., Kadane, J.B. (1990). Decisions Without Ordering. In *Acting and Reflecting* (Edited by W. Sieg), 143-170.

Received August 14, 2006; accepted October 31, 2006.

Scott Berry
Berry Consultants
3145 Chaco Canyon Drive
College Station, TX 77845, USA
scott@berryconsultants.com

Kert Viele
Department of Statistics
University of Kentucky
Lexington, KY 40506-0027, USA
viele@ms.uky.edu