The Decline and Fall of Type II Error Rates

Steve VERRILL and Mark DURST

For general linear models with normally distributed random errors, the probability of a Type II error decreases exponentially as a function of sample size. This potentially rapid decline reemphasizes the importance of performing power calculations.

KEY WORDS: Asymptotic relative efficiency; Experimental design; Hodges-Lehmann efficiency; Linear models; Mills' ratio; Minimum detectable difference; Noncentral F; Normal tail; Pitman efficiency; Power; Sample size.

1. INTRODUCTION

Introductory statistics students learn that hypothesis tests involve two types of error. If we reject a true null hypothesis, a Type I error occurs. If we fail to reject a false null hypothesis, a Type II error occurs. Students become adept at using tables to find critical values that control the probability of a Type I error. However, the calculations needed to control the probability of a Type II error can be complex and typically receive less emphasis in introductory courses. This is unfortunate as the scientists and engineers who take these courses and later design experiments sometimes fail to perform proper power calculations (power $= 1 -$ Prob(Type II error)). Instead their sample sizes are sometimes based on past practice or on available resources. As a result their experiments can be underdesigned (sample sizes are too small) or overdesigned (sample sizes are unneccessarily large). See Lenth (2001) for an overview of sample size issues.

In one sense a researcher does not design an experiment to achieve a certain Type I error probability. Instead, at the analysis stage, the researcher just enters a critical value table via the targeted probability and the appropriate degrees of freedom. Given that the scientist's model and distribution assumptions hold, the scientist is assured that the desired Type I error probability will result. Thus an analysis of Type I error is relatively easy to incorporate into an introductory statistics course or into a plan of experimentation. On the other hand, achieving a desired level of Type II error requires much more forethought. Prior to performing an experiment, a researcher must specify:

1. What Type I error rate (probability of falsely detecting an effect) are they willing to accept?
2. What differences do they want to be able to detect? Perhaps a 5% difference in means is not of practical importance, but the researcher wants to be fairly certain of detecting a 25% difference.
3. What is the variability in the property that is being tested?
4. What Type II error rate (probability of failing to detect a real effect) are they willing to accept?

Given this information they can calculate the necessary sample sizes for a wide variety of experiments.

If the necessary sample size is beyond the capability of the researcher, then they must be willing to consider obtaining more resources or abandoning the experiment rather than waste resources on an experiment that is unlikely to be definitive. If the necessary sample size is smaller than what standard practice dictates, then the researcher can save resources by adopting the size dictated by the power calculations.

The problem with failing to understand the implications of power calculations goes beyond a possible waste of resources. It goes to the validity of claims produced by improperly designed experiments. If the intent is to demonstrate that one process is better than another (e.g., that one mean is larger than another), and it turns out that the difference observed in the experiment is not statistically significant, researchers are generally sophisticated enough to note that a difference might still exist but that the sample sizes were simply too small to detect the difference statistically. They can always try again with a better designed experiment. They do not necessarily draw the possibly false conclusion that no difference exists. However, it is sometimes the case that researchers simply want to establish that a new (possibly cheaper) product or process is no worse than an existing product or process. If their experiment detects no statistical difference between the new and old products, then they might be tempted to conclude that the claim that they wanted to establish is indeed established. If, however, their experiment was not properly designed and their sample sizes were too small, the Type II error rate (the probability of failing to detect a difference that actually exists) associated with the experiment could be large and the lack of statistical significance would not represent good evidence that the new process is no worse than the old.

This article demonstrates that for a large class of experiments (those whose results can be represented by linear models with normally distributed random errors) the probability of a Type II error declines exponentially as a function of sample size. That is,

$$\text{Prob(Type II error)} = 1 - \text{power} < K \exp(-b \times n) \quad (1)$$

for constants $K, b > 0$, where n is a measure of the sample size. Results of this type are well known to those who work with asymptotic relative efficiencies (ARE) (see, e.g., Serfling 1980, sec. 10.5). However, the ARE approach might be somewhat opaque to many statistics students. Here we work through calculations that should be more accessible. We first illustrate the exponential rate of decrease in a special case, and then we establish it for general linear models.

The potentially rapid decline in the probability of a Type II error as sample size increases has important implications for researchers as they design their experiments. There can be a fairly sharp boundary between successful and unsuccessful experi-

Steve Verrill is Mathematical Statistician, U.S. Department of Agriculture Forest Products Laboratory, 1 Gifford Pinchot Drive, Madison, WI 53726 (E-mail: sverrill@fs.fed.us). Mark Durst is Staff Scientist, Lawrence Berkeley National Laboratory, currently at Amgen, Thousand Oaks, CA 91320.

 DOI: 10.1198/000313005X70353

ments. An underdesigned experiment can fairly rapidly become a successful and then an overdesigned experiment. At the close of Section 3, we provide the addresses of two Internet-based linear model power calculation programs that we have developed to aid in the design of experiments.

2. A SIMPLE COMPARISON OF TWO POPULATIONS, KNOWN VARIANCE

To characterize this case, we need a useful fact about the tail behavior of normal distributions. Versions of this fact have appeared previously in the statistical literature. See, for example, the discussions of "Mills' ratio" in Kendall and Stuart (1977) and Johnson and Kotz (1970). The particular form of the fact described in Lemma 1 is due to Gordon (1941). His proof is considerably more complex than the proof we give here.

Lemma 1. For $x < 0$,

$$x^2/(x^2+1) < \Phi(x)/\left(\phi(x)/(-x)\right) < 1, \tag{2}$$

and for $x > 0$,

$$x^2/(x^2+1) < (1-\Phi(x))/(\phi(x)/x) < 1, \tag{3}$$

where $\Phi(x)$ is the N(0,1) cumulative distribution function and $\phi(x)$ is the N(0,1) probability density function.

Proof: Let $x < 0$. We have

$$\begin{aligned}\phi(x)/(-x) &= \int_{-\infty}^{x} \frac{d}{dt}(\phi(t)/(-t))dt \\ &= \int_{-\infty}^{x} \phi(t)(1+(1/t^2))dt.\end{aligned}$$

Thus,

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt < \int_{-\infty}^{x} \phi(t)(1+(1/t^2))dt = \phi(x)/(-x),$$

and

$$\begin{aligned}\phi(x)/(-x) &= \int_{-\infty}^{x} \phi(t)(1+(1/t^2))dt \\ &< \int_{-\infty}^{x} \phi(t)(1+(1/x^2))dt = \Phi(x)(x^2+1)/x^2\end{aligned}$$

and result (2) follows.

Because for $x > 0$, $1-\Phi(x) = \Phi(-x)$, and $\phi(x) = \phi(-x)$, result (3) is an immediate consequence of result (2).

Now suppose that we have $n/2$ observations from a $N(\mu_1, \sigma^2)$ population and $n/2$ from a $N(\mu_2, \sigma^2)$ population, σ known. We would like to test the null hypothesis that $\mu_1 = \mu_2$ versus the alternative that $\mu_1 \neq \mu_2$. In this case the test statistic is the ratio $z = (\bar{X}_2 - \bar{X}_1)/(\sigma\sqrt{4/n})$, and we reject the null hypothesis if $z < -a$ or $z > a$ for an appropriate critical value, a.

The power associated with this test equals

$$\begin{aligned}&\text{Prob}\left(\frac{\bar{X}_2-\bar{X}_1}{\sigma\sqrt{4/n}} < -a\right) + \text{Prob}\left(\frac{\bar{X}_2-\bar{X}_1}{\sigma\sqrt{4/n}} > a\right) \\ &= \Phi\left(-a - \frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma}\right) + 1 - \Phi\left(a - \frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma}\right).\end{aligned}$$

Thus,

$$\begin{aligned}\text{Prob(Type II error)} &= 1 - \text{power} \\ &= \Phi\left(a - \frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma}\right) \\ &\quad -\Phi\left(-a - \frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma}\right) \quad (4)\end{aligned}$$

so

$$\text{Prob(Type II error)} < \Phi\left(a - \frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma}\right), \tag{5}$$

and

$$\text{Prob(Type II error)} < 1 - \Phi\left(-a - \frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma}\right). \tag{6}$$

By Lemma 1, for $\mu_2 > \mu_1$ and n large enough, the quantity on the right side of Equation (5) is of the order

$$\exp\left(-\left(a - \frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma}\right)^2/2\right) \Big/ \left(\frac{(\mu_2-\mu_1)\sqrt{n}}{2\sigma} - a\right)$$

which, for large n, is dominated by

$$\exp\left(-a^2/2 + \frac{a(\mu_2-\mu_1)\sqrt{n}}{2\sigma} - \frac{(\mu_2-\mu_1)^2 n}{8\sigma^2}\right)$$

which, in turn, is dominated for large n by $\exp(-b \times n)$ for any b less than $(\mu_2-\mu_1)^2/(8\sigma^2)$.

From result (6), the $\mu_2 < \mu_1$ case follows in a similar fashion.

Although this is a large sample result, a roughly exponential decline in the probability of a Type II error can actually hold for "small" samples. For example, for a .05 significance level, $\sigma/(\mu_2-\mu_1) = 1$, and $n = 10, 20, 30, 40, 50$, the Type II error probabilities for a z test of the equality of the means are .65, .39, .22, .11, and .06. (In this example, the coefficient of variation is assumed to be roughly equal to the percent difference in the means. Hence the $\sigma/(\mu_2-\mu_1) = 1$ relationship between standard deviation and the difference in the means.) The logs of these values are plotted against n in Figure 1. This plot appears to be approximately linear. That is, Prob(Type II error) $\approx K\exp(-b \times n)$.

3. THE GENERAL LINEAR MODEL

Now suppose that we have the linear model

$$\mathbf{y} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where $\mathbf{y}$ is the $m \times 1$ vector of responses, $\mathbf{X}$ is the $m \times p$ design matrix, and $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector. Following Scheffé (1959), suppose that we want to test the hypothesis $\mathbf{c}_1^T\boldsymbol{\beta} = \eta_1$, ..., $\mathbf{c}_q^T\boldsymbol{\beta} = \eta_q$ where the $\mathbf{c}_i^T\boldsymbol{\beta}$'s are estimable and the $\mathbf{c}_i$'s are linearly independent. [For example, in a balanced one-way analysis of variance with J "treatments" and I replicates of each treatment, $m = I \times J$, $p = J$, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j^T = (0, \ldots, 0\ 1, \ldots, 1\ 0, \ldots, 0)$ and $\mathbf{x}_j$ contains $I \times (j-1)$ initial 0's, I 1's, and $I \times (J-j)$ ending 0's. In this case we are interested in testing the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_p$ or $\beta_1 - \beta_2 = \cdots = \beta_1 - \beta_p = 0$. We have $\mathbf{c}_1^T =$

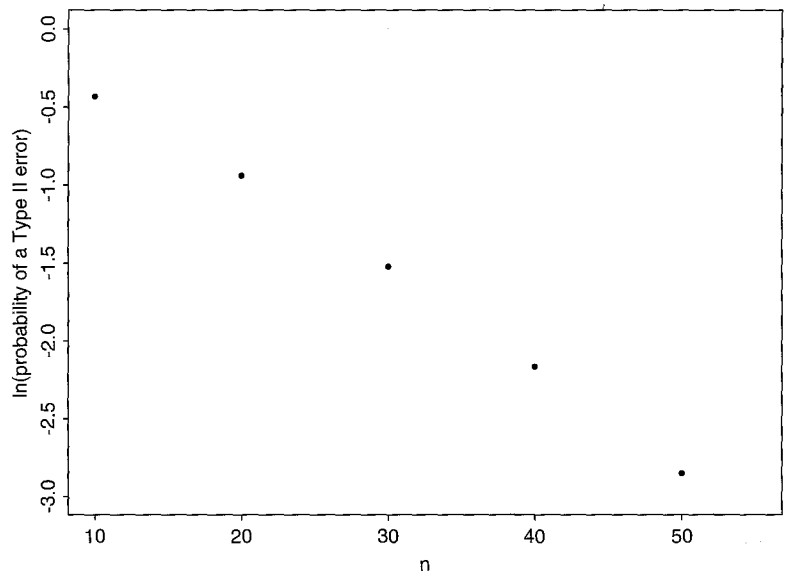

Figure 1. The approximately linear decline of ln(Prob(Type II error)) with increasing sample size in the two sample example. In this example, we have n/2 observations from a N(μ_1, σ^2) population, and n/2 observations from a N(μ_2, σ^2) population. σ is known, and we use a z statistic to test the hypothesis $\mu_1 = \mu_2$ versus the alternative $\mu_1 \neq \mu_2$. The probability of a Type II error is given by Equation (4). In the plot we present ln(probability of a Type II error) versus n for the case in which $\sigma/(\mu_2 - \mu_1) = 1$.

$(1\ -1\ 0, \ldots, 0), \ldots, \mathbf{c}_{p-1}^T = (1\ 0, \ldots, 0\ -1)$, and $\eta_1 = \cdots = \eta_{p-1} = 0$.] Because the $\mathbf{c}_i^T\boldsymbol{\beta}$'s are estimable and the $\mathbf{c}_i$'s are linearly independent, we can find unique linearly independent vectors, $\mathbf{a}_1, \ldots, \mathbf{a}_q$, that lie in the linear span of the columns of $\mathbf{X}$ and satisfy $\mathbf{a}_i^T\mathbf{X} = \mathbf{c}_i^T$.

Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_q)$, $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_q)$, and $\boldsymbol{\eta}^T = (\eta_1, \ldots, \eta_q)$. Under the null hypothesis, $\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{C}^T\boldsymbol{\beta} = \boldsymbol{\eta}$ and

$$\mathbf{A}^T\mathbf{y} \sim \mathrm{N}\left(\mathbf{A}^T\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{A}^T\mathbf{A}\right) = \mathrm{N}\left(\boldsymbol{\eta}, \sigma^2\mathbf{A}^T\mathbf{A}\right),$$

or

$$\mathbf{A}^T\mathbf{y} - \boldsymbol{\eta} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{A}^T\mathbf{A}\right),$$

and

$$(\mathbf{A}^T\mathbf{A})^{-1/2}(\mathbf{A}^T\mathbf{y} - \boldsymbol{\eta}) \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_{q\times q}\right).$$

Under the alternative hypthesis, $\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\eta}$ and

$$(\mathbf{A}^T\mathbf{A})^{-1/2}(\mathbf{A}^T\mathbf{y} - \boldsymbol{\eta}) \sim \mathrm{N}\left((\mathbf{A}^T\mathbf{A})^{-1/2}(\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}), \sigma^2\mathbf{I}_{q\times q}\right),$$

Thus, under the null hypothesis, the standardized F test numerator sum of squares

$$SS_N \equiv (\mathbf{A}^T\mathbf{y} - \boldsymbol{\eta})^T(\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{y} - \boldsymbol{\eta})/\sigma^2$$

is distributed as a central chi-squared random variable with q degrees of freedom, while under the alternative hypothesis, SS_N is distributed as a noncentral chi-squared with q degrees of freedom and noncentrality parameter

$$\lambda_m = \left(\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}\right)^T\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}\right)/\sigma^2. \quad (7)$$

In this case the Type II error probability associated with the standard ANOVA F test equals the probability that a noncentral $F_{q,m-r,\lambda_m}$ random variable lies below the appropriate critical value, x_m, derived from a central $F_{q,m-r}$ random variable (here r is the rank of the $\mathbf{X}$ matrix). As the noncentrality parameter increases this probability decreases. Larger differences among treatment means (resulting in larger $\left(\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}\right)^T\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}\right)$ values) and smaller variances will lead to larger values of λ_m and a reduction in the probability of a Type II error. More importantly for the purposes of this article, increased sample sizes will lead to larger noncentrality parameters and thus reductions in the probability of a Type II error.

How, specifically, does the noncentrality parameter change as the sample size changes? Suppose that we multiply the design by a factor of k so that $n = k \times m$. Then the new $n \times p$ design matrix contains k copies of the original design matrix and the new $\mathbf{a}_i$ must contain k copies of the original $\mathbf{a}_i$, each copy divided by k. Thus, the new noncentrality parameter is just k times the old noncentrality parameter, and

$$\lambda_n = \lambda_{k\times m} = k \times \lambda_m = n \times (\lambda_m/m) \propto n. \quad (8)$$

The new denominator degrees of freedom in the standard linear model F test statistic is $n - r$.

As noted earlier, in this case the Type II error probability equals the probability that a $F_{q,n-r,\lambda_n}$ random variable will lie below the appropriate critical value, x_n, derived from a central $F_{q,n-r}$ random variable. [Note how the four inputs discussed in Section 1 are incorporated into the calculation of appropriate sample size: 1) The Type I error probability appears in the choice of the critical value. 2) The differences show up as nonzero $\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}$ values in the noncentrality parameter calculation. 3) The variability appears as σ^2 in the noncentrality parameter calculation. 4) Given a base design and corresponding noncentrality parameter, λ_m, the targeted Type II error probability is obtained by finding the lowest k value such that the Type II error probability calculated using noncentrality parameter $\lambda_n = \lambda_{k\times m} = k \times \lambda_m$ falls below the targeted level. An

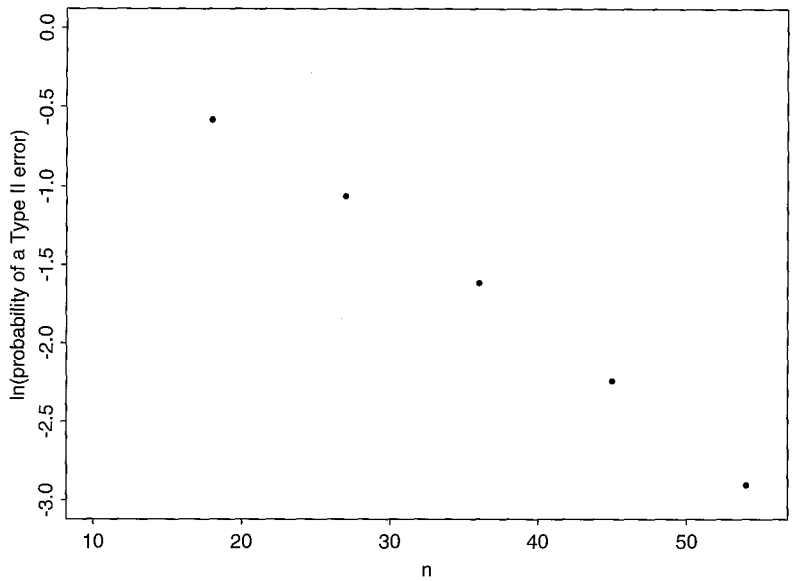

Figure 2. The approximately linear decline of ln(Prob(Type II error)) with increasing sample size in the 3 × 3 ANOVA example. In this example, we have a 3 × 3 ANOVA design with 2, 3, 4, 5, or 6 replicates per cell. We use a standard F statistic to test the hypothesis that the three means associated with the first factor are all equal versus the alternative hypothesis that there are differences among the three means. The probability of a Type II error in this case is given by result (9). The mean values for the three levels of the factor were taken to be .9, 1.0, and 1.1. The σ value was taken to be .15. In the plot we present ln(probability of a Type II error) versus n = 3 × 3 × number of replicates.

acceptable design is then one that is k replicates of the base design.]

Because x_n decreases as n increases, we have

$$\begin{aligned}\text{Prob(Type II error)} &= 1 - \text{power} \\ &= F_{q,n-r,\lambda_n}(x_n) < F_{q,n-r,\lambda_n}(x) \quad (9)\end{aligned}$$

for an appropriate fixed x.

To proceed with our demonstration of an exponential decline in the Type II error probability as a function of sample size we now need the following theorem. Its proof is provided in the appendix of Verrill and Durst (2005).

Theorem 1. Let $v_2 \to \infty$ as $n \to \infty$. Let v_1 and x be fixed. Then for any $d > 2$, there exists an $N_d > 0$ such that, for all $\lambda > 0$, $n > N_d$ yields

$$F_{v_1,v_2,\lambda}(x) < K_d \exp(-\lambda/d) \quad (10)$$

for some constant K_d.

Results (9) and (10) yield

$$\text{Prob(Type II error)} = 1 - \text{power} < K_d \exp(-\lambda_n/d) \quad (11)$$

for $n > N_d$, or, since, by (8), $\lambda_n = n(\lambda_m/m)$,

$$\text{Prob(Type II error)} = 1 - \text{power} < K_d \exp\left(-\frac{n(\lambda_m/m)}{d}\right) \quad (12)$$

for any constant $d > 2$ and $n > N_d$. This establishes that the probability of a Type II error declines exponentially as a function of n.

Note that in the case considered in Section 2,

$$\lambda_n = \frac{(\mu_2 - \mu_1)^2 n}{4\sigma^2}$$

so from (11) we would expect

$$\text{Prob(Type II error)} = 1 - \text{power} < K_d \exp\left(-\frac{(\mu_2 - \mu_1)^2 n}{4\sigma^2 d}\right)$$

for any $d > 2$. This is equivalent to the bounding rate identified at the end of Section 2.

As in the simple case discussed in Section 2, although the exponential decrease in the probability of a Type II error is an asymptotic result, a roughly exponential decline can hold for small samples. For example, given a 3 × 3 ANOVA design, a .05 significance level, μ values for one of the factors equal to .9, 1.0, and 1.1, a σ value of .15, and 2, 3, 4, 5, and 6 replicates per cell, the Type II error probabilities for the standard F test of the equality of the μ's are .56, .35, .20, .11, and .06. (See http://www1.fpl.fs.fed.us/power.glm.html for a program that permits general linear model power calculations to be performed using the World Wide Web. See http://www1.fpl.fs.fed.us/power.html for a simpler program that calculates power for balanced ANOVAs.) The logs of these values are plotted against n in Figure 2. Again this plot appears to be approximately linear. That is, Prob(Type II error) $\approx K \exp(-b \times n)$.

4. DESIGN IMPLICATIONS

From (12) we can obtain the Δn_{half} that (approximately) halves the Type II error probability. We have

$$\begin{aligned}1/2 &= (\text{Error probability})_2/(\text{Error probability})_1 \\ &= \exp\left(-\frac{(n_2 - n_1)(\lambda_m/m)}{d}\right),\end{aligned}$$

or

$$\Delta n_{\text{half}} = n_2 - n_1 = d \times m \times \ln(2)/\lambda_m.$$

Thus, large treatment differences or small variances (which yield large λ_m values) can yield very rapid declines in Type II error

probabilities. In this case, as noted in the introduction, an underdesigned experiment can fairly rapidly become a successful and then an overdesigned experiment.

5. AN ASIDE ON ASYMPTOTIC RELATIVE EFFICIENCIES (ARES)

In the introduction we remarked that those who study AREs are familiar with exponential declines in the probability of a Type II error. In particular in the Hodges-Lehmann approach to ARE, one fixes the probability of a Type I error and the difference one wants to detect, and compares tests based on the rate at which the probability of a Type II error declines. Thus, in the normal theory linear model case, the focus will in effect be on the constant b in (1). A larger b will correspond to an asymptotically more efficient test—given a fixed difference that one wants to detect, the probability of a Type II error will decline more rapidly.

On the other hand, in the Pitman approach to asymptotic relative efficiencies, one *fixes* the Type II (and Type I) error probability and observes the manner in which the minimal detectable difference declines as sample size increases. (The minimal detectable difference is the smallest difference for which power $\geq 1 -$ the fixed Type II error probability.) In this case (constant Type II error probability), the noncentrality parameter given by (8) must converge to a constant so we have

$$\lambda_m \propto 1/n$$

or, from (7), $\left(\mathbf{A}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}\right)$, the "difference" that we are trying to detect must be declining as $1/\sqrt{n}$. Thus, as one would expect given the equivalence between hypothesis tests and confidence intervals, for fixed Type I and Type II error probabilities, the minimal detectable difference declines at the same rate as confidence interval lengths. Pitman ARE differences among hypothesis tests will show up as differences in multipliers of the basic $1/\sqrt{n}$ rate of decline of the minimal detectable difference. Smaller multipliers will correspond to more asymptotically efficient tests—given a fixed Type II error probability, the minimal detectable difference will be smaller.

6. SUMMARY

We have established that for tests of hypotheses in general linear models, the probability of a Type II error declines exponentially. To do so, we have made use of a Mills' ratio lemma that permits one to approximate the tail behavior of the normal distribution and (see Verrill and Durst 2005, appendix) Tang's (1938) asymptotic expansion of the noncentral F distribution.

The potentially rapid decline in the probability of a Type II error reemphasizes the importance of performing power calculations. For reasons of experimental efficiency (and in underdesigned cases, statistical validity), it is important to neither underdesign nor overdesign a study. We have provided Web resources that facilitate the performance of linear model power calculations. Links to additional Web resources for performing power calculations can be found at http://www.stat.uiowa.edu/~rlenth/Power/ and http://members.aol.com/johnp71/javastat.html#Power.

[Received July 2004. Revised June 2005.]

REFERENCES

Gordon, R. D. (1941), "Values of Mills' Ratio of Area to Bounding Ordinate and of the Normal Probability Integral for Large Values of the Argument," *Annals of Mathematical Statistics*, 12, 364–366.

Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions—2*, Boston, MA: Houghton Mifflin.

Kendall, M., and Stuart, A. (1977), *The Advanced Theory of Statistics* (vol. 1), New York: MacMillan.

Lenth, R. V. (2001), "Some Practical Guidelines for Effective Sample Size Determination," *The American Statistician*, 55, 187–193.

Scheffé, H. (1959), *The Analysis of Variance*, New York: Wiley.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Tang, P. C. (1938), "The Power Function of the Analysis of Variance Tests with Tables and Illustrations of Their Use," *Statistical Research Memoirs*, 2, 126–149.

Verrill, S. P., and Durst, M. J. (2005), "The Decline and Fall of Type II Error Rates," Research Paper FPL-RP-628, Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory.