Published in final edited form as:
Am Stat. 2011 January 1; 65(3): . doi:10.1198/tast.2011.10117.

A Note on Comparing the Power of Test Statistics at Low Significance Levels

Nathan Morris and **Robert Elston**
Case Western Reserve University Cleveland, OH 44106-7281, USA

Abstract

It is an obvious fact that the power of a test statistic is dependent upon the significance (alpha) level at which the test is performed. It is perhaps a less obvious fact that the *relative* performance of two statistics in terms of power is also a function of the alpha level. Through numerous personal discussions, we have noted that even some competent statisticians have the mistaken intuition that relative power comparisons at traditional levels such as $\alpha = 0.05$ will be roughly similar to relative power comparisons at very low levels, such as the level $\alpha = 5 \times 10^{-8}$, which is commonly used in genome-wide association studies. In this brief note, we demonstrate that this notion is in fact quite wrong, especially with respect to comparing tests with differing degrees of freedom. In fact, at very low alpha levels the cost of additional degrees of freedom is often comparatively low. Thus we recommend that statisticians exercise caution when interpreting the results of power comparison studies which use alpha levels that will not be used in practice.

Keywords

Power; Small Significance Levels

1 Introduction

Hypothesis testing is an established part of any conventional statistics education, as are the familiar phrases “P-Value,” “type I error,” “alpha level” and “power.” Every first-year student of statistics learns that if the P-Value is less than $\alpha = 0.05$ one should “reject the null hypothesis.” While this cutoff is arbitrary, it is nevertheless tradition, and so we have grown used to it. We have developed our professional intuitions around this sacred, but quite arbitrary, standard. For example, traditional claims about the robustness of the central limit theorem in terms of Type I error are based on $\alpha = 0.05$. While such claims may be challenged even when $\alpha = 0.05$, one could certainly wonder if such rules of thumb are even remotely valid when alpha is orders of magnitude smaller than traditional levels (Bradley, 1978). In this note, however, we will pretend that all such considerations relating to the proper control of type I error at extremely small alpha levels are adequately addressed. Instead, we will consider power, and show that intuitions developed for $\alpha = 0.05$ may not apply at much lower alpha levels.

To motivate our discussion, let us consider genome-wide association studies (Wang et al., 2005). There are positions in the genome where the DNA of some individuals differs from that of others by a single nucleotide. These variations among individuals are referred to as

Author’s Footnote:
Nathan Morris is Instructor, Case Western Reserve University, OH 44106 (njm18@case.edu); and Robert Elston is Chair of the Department of Epidemiology and Biostatistics, Case Western Reserve University, OH 44106.

single nucleotide polymorphisms (SNPs). Usually, SNPs have only two possible variations (i.e., two possible alleles), and we will refer to the more common allele as "A" and the less common allele as "a." Because humans have two copies of most of their DNA, there are three possible combinations of alleles (i.e., three possible genotypes) that an individual can have: "AA," "Aa" and "aa." We make no distinction between "Aa" and "aA." A typical genome-wide association study will assay something on the order of 10^6 SNPs for each individual. A hypothesis test is then performed for each SNP to determine if it is associated in some way with the trait of interest. This leads to a Bonferoni corrected alpha level of $\alpha = 0.05/10^6 = 5 \times 10^{-8}$ to control the family wise error rate at 0.05.

If the trait of interest (y) is quantitative, then we may easily view the test in the linear model framework. For SNP j and individual i we may code a predictive variable x_{ij} as 0, 1 or 2 for genotypes "AA," "Aa" or "aa," respectively (i.e., x_{ij} represents the number of "a" alleles carried by individual i at SNP j). Two possible single SNP linear models are: (m_1) $y_i = \beta_0+\beta_1 x_{ij}+\varepsilon_i$ and (m_2) $y_i=\beta_0+\beta_1 x_{ij}+\beta_2 x_{ij}^2+\epsilon_i$. For model m_1 we would test the null hypothesis $H_0^{(1)}:\beta_1=0$ vs. $H_a^{(1)}:\beta \neq 0$, while for model m_2 we would test the hypothesis $H_0^{(2)}:\beta_1=\beta_2=0$ vs. $H_a^{(2)}:\beta_1 \neq 0$ or $\beta_2 \neq 0$. Comparing the power of a test of $H_0^{(1)}$ with a test of $H_0^{(2)}$ involves comparing two tests with differing degrees of freedom. It is true that in some sense these are distinct hypotheses, and technically, $H_0^{(1)}$ may be true at the same time $H_0^{(2)}$ is false. However, the rejection of either hypothesis will lead us to the conclusion that the SNP being investigated is associated with the disease. Some have suggested testing for association between y and two neighboring markers at a time (Kim et al., 2010). In this case we have numerous possible linear models such as: (m_3) $y_i = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{i(j+1)} + \varepsilon_i$ and (m_4) $y_i = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{i(j+1)} + \beta_3 x_{i(j+1)} x_{ij} + \varepsilon_i$. For m_3 we would test the hypothesis $H_0^{(3)}:\beta_1=\beta_2=0$ vs. $H_a^{(3)}:\beta_1 \neq 0$ or $\beta_2 \neq 0$, while for m_4 we would test the hypothesis $H_0^{(4)}:\beta_1=\beta_2=\beta_3=0$ vs. $H_a^{(4)}:\beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$. Again, comparing these two tests involves comparing two tests with differing degrees of freedom, though from a biological viewpoint they are equivalent in that they both may be used to demonstrate that a particular genomic region is associated with a trait. Clearly more complicated tests may be created with quadratic terms and higher-order interaction terms leading to tests with a large number of degrees of freedom.

Suppose we are comparing the power of two test statistics that are based upon two models m_a and m_b. For example, m_a and m_b could be m_1 and m_2 above. We define ν_a as the number of coefficients in model m_a, and ν_b is defined similarly. We will assume that $\nu_a < \nu_b$. For m_a we perform a test (using, say, the Wald test) of $H_0^{(a)}:\beta_1=\ldots=\beta_{\nu a}$, and for m_b we perform a test of $H_0^{(b)}=\beta_1=\ldots=\beta_{\nu_b}$. We refer to the test statistics as T_a and T_b, respectively. Let $\chi^2_{\nu,\delta}$ represent a non-central chi-squared distribution with degrees of freedom ν and a non-centrality parameter (NCP) Δ. Also, let n represent the sample size. In order to make our discussion simple we will simply assume that under the null hypothesis $T_a \sim \chi^2_{\nu_a}$ and $T_b \sim \chi^2_{\nu_b}$. Under a given alternative, we will assume that the distributions can be adequately characterized by the two non-centrality parameters (NCPs) $\Delta_a = n\lambda_a$ and $\Delta_b = n\lambda_b$, so that $T_a \sim \chi^2_{\nu_a,\delta_a}$ and $T_a \sim \chi^2_{\nu_b,\delta_b}$. Some discussion of how to calculate such NCPs for genome-wide association studies may be found for example in Kim et al. (2010) and Yang et al. (2010).

We now introduce a key piece of notation:

$$q = (\delta_b - \delta_a)/\delta_a, \quad (1)$$

or equivalently $\delta_b = \delta_a(1 + q)$. Thus q, which is independent of the sample size, represents the fractional increase in the NCP that can be achieved by using a hypothesis test based on the higher dimensional model m_b instead of the lower dimensional model m_a. When $q = 0$ then the NCPs of the two models are equal. Unfortunately, T_b has additional degrees of freedom compared to T_a, and therefore at a given alpha level the cutoff for rejecting H_0 using T_b is larger than the cutoff for rejecting using T_a. There is a cost/benefit analogy to be drawn; we can purchase a larger NCP (as quantified by q) at the cost of increased degrees of freedom. We will only be willing to pay this price if q is large compared to the increase in degrees of freedom.

In what follows, we pose two simple questions, and show that the answers to these questions depend on alpha. All of the computations below were performed in R 2.9.2. The numerical accuracy of the chi-squared CDF at low alpha levels in R has been evaluated by Bangalore et al. (2009).

2 Results

First, we ask: what must q be for T_a and T_b to have equal power? That is, to borrow from the cost benefit analogy, we ask what value of q would allow us to break even with costs equal to the benefits. To answer this question, we first consider a test statistic (T) such that $T \sim \chi^2_\nu$ distribution under H_0 and $T \sim \chi^2_{\nu,\delta}$ under some alternative hypothesis. In this case, the power for a given α and ν is $1 - F_{\nu,\delta}\left(F^{-1}_{\nu,0}(1-\alpha)\right)$ where $F_{\nu,\delta}\left(F^{-1}_{\nu,0}(1-\alpha)\right)$ is the cumulative distribution function of the non-central chi-squared distribution with degrees of freedom ν and $NCP = \Delta$ evaluated at the $1 - \alpha$ quantile of the central chi-squared distribution with ν degrees of freedom. Because the power of T at a fixed α is a strictly increasing (thus invertible) function of Δ, we may solve for the Δ which would yield a power (p) where $p \in (\alpha, 1)$. We define the function $\Delta(\alpha, p, \nu)$ as the solution to this equation (i.e. the Δ such that $1 - F_{\nu,\delta}\left(F^{-1}_{\nu,0}(1-\alpha)\right) = p$. We wish to find the value of q that would give equal power (p) for T_a and to T_b at a given α, and we define the function $Q(\alpha, p, \nu_a, \nu_b)$ as the answer to this question. From equation (1) we have

$$Q(\alpha, p, \nu_a, \nu_b) = [\Delta(\alpha, p, \nu_b) - \Delta(\alpha, p, \nu_a)]/\Delta(\alpha, p, \nu_a). \quad (2)$$

Figures 1a and 1b show plots of $Q(\alpha, p, 1, 2)$ and $Q(\alpha, p, 1, 3)$, which are obviously dependent on alpha. Figure 1a and 1b may be interpreted as saying that, in general, for a given p, α, ν_a and ν_b, we should use T_b if q is above what is suggested by the corresponding curve and we should use T_a otherwise. What is perhaps most remarkable about Figures 1a and 1b is the decreasing behavior of the function. This indicates that the cost of an additional degree of freedom, at least in this context, is less at a low alpha level than at a standard 0.05 alpha level. That is very small. The general behavior of the curves is similar over a broad range of choices for ν_a and ν_b.

Our second question is this: given some values of q and the power of T_a, how much power would be gained or lost by using T_b instead of T_a? Using the same notation described above, we want to know

$$\left[1 - F_{\nu b,(1+q)\Delta(\alpha,p,\nu_a)} - p\right]/p \quad (3)$$

In equation (3) the non-centrality parameter $(1 + q)\ \Delta\ (a, p, \nu_a)$ comes from equation (1) because $\Delta_b = (1 + q)\Delta_a$. In Figures 1c and 1d we show that, in the worst case scenario, when $q = 0$ (i.e., the extra degree(s)of freedom buy(s) no increase in the NCP) the percentage of power lost is fairly constant across alpha levels. However, if there is even a small increase in the NCP, the behavior at different alpha levels may be quite different. The curves shown in Figures 1c and 1d have the power (p) of T_a fixed at 65%. As may be expected, for larger p the dependence on alpha is less dramatic (i.e., the slopes are not as large) because a power cannot be larger than 100%.

3 Conclusion

We may wonder why we instinctively believe that power comparisons between different statistics will not be dependent on alpha. It may be our natural propensity to generalize, or it may be our experience with point hypotheses and uniformly most powerful (UMP) tests (Lehmann and Romano, 2005). UMP tests are typically not only most powerful for every value of the parameter in the parameter space, but also for every "attainable" value of $a \in [0, 1]$ (Hogg and Craig, 1995, p. 411). In the examples we present, no UMP test exists because we can easily find genetic models that would make the test of $H_0^{(a)}$ more powerful than $H_0^{(b)}$ and vice versa. Therefore, power comparison studies must compare power over a broad range of plausible genetic models.

With the development of numerous high throughput molecular techniques such as array-based genotyping, many modern applications in statistics involve high dimensional scans, and thus require multiple-test corrections leading to extremely small alpha levels. Many methodological papers set out to compare the power of various test statistics for such high dimensional applications. This note serves as a warning that power comparisons done with $a = 0.05$ may not generalize to these realistic applications, especially when comparing statistics with differing degrees of freedom.

Acknowledgments

Morris was supported by NCI Grant R25T CA094186. Elston was supported by the NCRR grant P41 RR03655 and NCI grant P30 CAD43703.

References

Bangalore S, Wang J, Allison D. How Accurate are the Extremely Small P-values Used in Genomic Research: An Evaluation of Numerical Libraries. Computational Statistics and Data Analysis. 2009; 53:2446–2452. [PubMed: 20161126]

Bradley J. Robustness. British Journal of Mathematical and Statistical Psychology. 1978; 31:144–152.

Hogg, R.; Craig. Introduction to Mathematical Statistics. Fifth Edition. Prentice-Hall; New Jersey: 1995.

Kim S, Morris N, Won S, Elston R. Single marker and two marker association tests for unphased case control genotype data, with a power comparison. Genetic epidemiology. 2010; 34:67–77. [PubMed: 19557751]

Lehmann, E.; Romano, J. Testing statistical hypotheses. Springer Verlag; New York: 2005.

Wang W, Barratt B, Clayton D, Todd J. Genome-wide Association Studies: Theoretical and Practical Concerns. Nature Reviews Genetics. 2005; 6:109–118.

Yang J, Wray N, Visscher P. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. Genetic epidemiology. 2010; 34(3):254–257. [PubMed: 19918758]

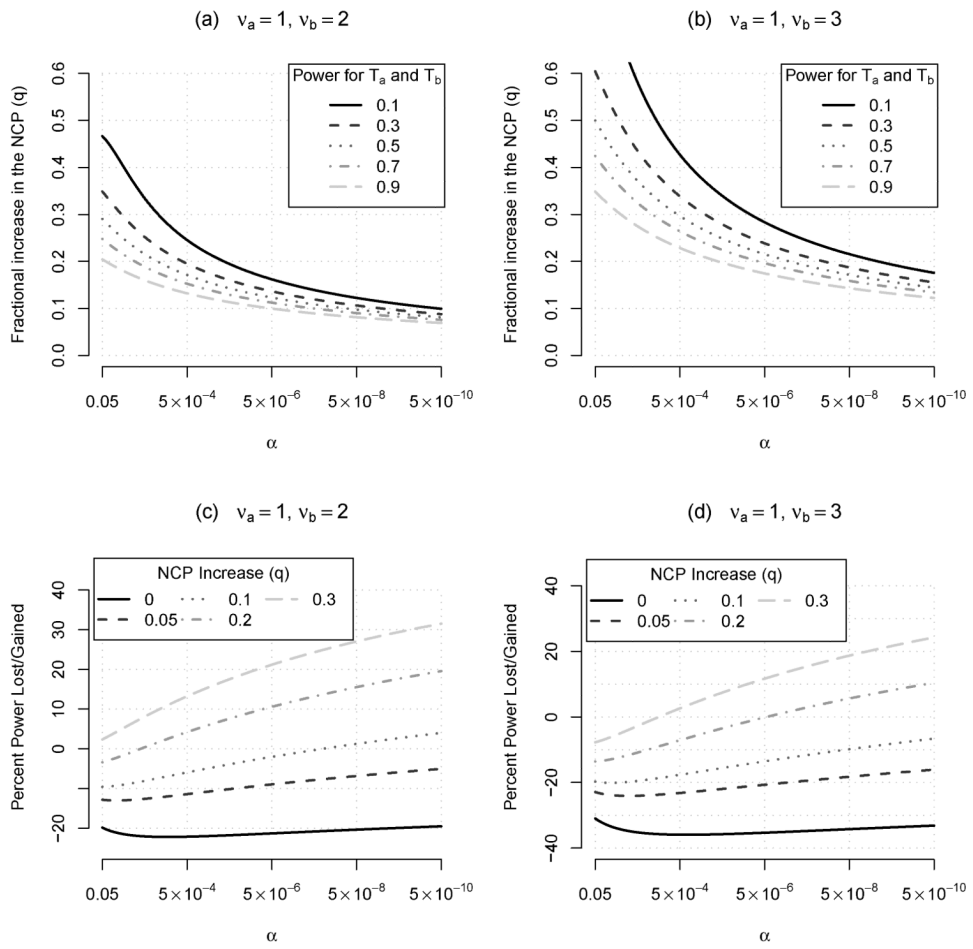

Figure 1.

Power comparisons are a function of alpha level. By definition $q = (\delta_b - \delta_a)/\delta_a$ is the relative increase in the NCP purchased by the additional degrees of freedom. Note the log scale on the x-axis of all four graphs. (a,b) For different values of α, power (*p*) of T_a, and degrees of freedom, we show how large q must be for T_b and Ta to have equal power. (c,d) For different values of *q* and α and with the power of T_a fixed at 65% we show the percentage of power gained by using T_b (i.e., $100\% \times [\left[1 - F_{\nu_b(1+q)\Delta(\alpha,0.65,\nu_a)}\left(F^{-1}_{\nu_b,0}(1-\alpha)\right) - 0.65\right]/0.65$).