Calibrating the evidence in experiments with applications to meta-analysis

Elena Kulinskaya,*Stephan Morgenthaler†and Robert G. Staudte‡

June 4, 2006

1 Introduction

How much more evidence is there in a 'highly significant' p-value of 0.01 relative to one 'just significant' at 0.05? Why does the replication of an experiment lead, on average, to a higher p-value than the one just obtained? To answer such questions one must go beyond the traditional p-value which is conditional on the data and thus interpretable only in the context of the experiment just performed. One can achieve this by considering the *random* p-value which has a uniform distribution under the null hypothesis but a highly skewed distribution under alternative hypotheses.

When considered as a random variable, the p-value becomes another test statistic, and thus one can ask, which test statistic, if not the p-value, best captures the notion of 'evidence' ? Morgethaler and Staudte (2005) suggest that the answer is a transformation which takes the test statistic into *evidence* T which has a normal distribution with mean τ and variance 1 for all values of the distributional parameters of the test statistic. The mean evidence τ should grow from 0 as the alternative moves away from the null; and, further, for a fixed alternative should grow at the same rate as the alternative can

*Imperial College, London,e.kulinskaya@imperial.ac.uk
†École Polytechnique Fédérale de Lausanne, stephan.morgenthaler@epfl.ch
‡La Trobe University, Melbourne, r.staudte@latrobe.edu.au

be estimated, typically the square root of the sample size. The advantages of variance stabilization have long been appreciated by statisticians, (see Anscombe (1948) and Efron (1982), for example. They include small sample normal approximations, and confidence intervals for τ which can easily be transformed into confidence intervals for a model parameter.

2 Calibrating the evidence in p-values

For definiteness consider the simple model in which the test statistic is the sample mean $\bar{X}_n$ having the normal distribution with unknown mean μ and standard deviation $1/\sqrt{n}$. For testing $\mu = 0$ against $\mu > 0$ the random p-value is $PV = \Phi(-\sqrt{n}\bar{X}_n)$. The probit transformation $p \to \Phi^{-1}(1-p)$ clearly transforms PV to $T = \sqrt{n}\bar{X}_n$ which has the normal distribution with mean $\tau = \sqrt{n}\,\mu$ and variance 1, thus satisfying for any $\mu > 0$ the desirable properties of evidence (see $E_1 - E_4$ below).

Instead of reporting a p-value of 0.05, we advocate reporting evidence $T = 1.645$, plus or minus standard error 1. Further, on this scale a p-value of 0.01 is reported as 2.33, plus or minus standard error 1. So 0.01 reflects only about 41% more evidence than 0.01 in this example, subject to equal standard errors of 1. To obtain twice the expected evidence, a p-value of 0.0005 is required. This is more in keeping with what Bayesian statisticians have been arguing for years, although a recent (Selke, Bayarri and Berger, 2001) Bayesian calibration scale for the p-value, when examined from the point of view espoused here, shows that posterior probabilities of the null underestimate the expected evidence in the p-value by at least one standard deviation over the range of interest.

In the context of Neyman Pearson hypothesis testing, the expected evidence is simply the sum of the probits of the false positive and false negative rates, so once the expected evidence is found, a formula for the power function of a test can be deduced. In addition, bits of evidence on the probit scale are easily combined, facilitating the computation and interpretation of evidence for joint alternatives in multiple related experiments. Standard meta-analytic theory applies, but with *known* weights, which circumvents a major problem in meta-analysis.

Evidence for the two-sided alternative $\mu \neq 0$ is *not* simply $\Phi^{-1}(1-p^{\pm})$, where $p^{\pm} = 2\Phi(-\sqrt{n}|\bar{X}_n|)$ is the two-sided p-value, for this transformation is

not variance stabilized, having a singularity at $\bar{X}_n = 0$. However, the transformation $p^{\pm} \to T^{\pm} = \max\{0, \Phi^{-1}(1-p^{\pm})\}$ is 'nearly' variance stabilized. This example raises the question of how general is the calibration scale.

3 Calibration of evidence

Let θ be an unknown effect for which it is desired to test $\theta = 0$ against $\theta > 0$, and let S be a test statistic which rejects H_0 for large values of S. We want a measure of one-sided evidence T to satisfy:

E_1. The one-sided evidence T is a monotone increasing function of S;

E_2. the distribution of T is normal for all values of the unknown parameters;

E_3. the variance $\mathrm{Var}[T] = 1$ for all values of the unknown parameters; and

E_4. the expected evidence $\tau = \tau(\theta) = \mathrm{E}_\theta[T]$ is monotone increasing in θ from $\tau(0) = 0$.

In the simple example of a normal model with known variance all of the above properties hold exactly for evidence defined by the Z-test statistic; that is, the standardized effect. In general, properties $E_2 - E_4$ will hold only approximately, but to a surprising degree, even for small sample sizes.

We somewhat arbitrarily describe values of T near 1.645 as *weak* evidence against the null. Values of T which are twice as large we call *moderate* evidence, and values which are 3 times as large as *strong* evidence. Thus our definition of weak evidence follows Fisher's low standard when the null is true, but we are otherwise measuring evidence against the null on a different calibration scale, one which allows for interpretation whether or not the null hypothesis holds.

4 Example

Let X have the Binomial(n, p) distribution, with $0 < p < 1$. For testing $p = p_0$ against $p > p_0$, The classical transformation $a_n(p) = 2\sqrt{n}\ \arcsin(\sqrt{p})$ with $\tilde{p} = (X + 3/8)/(n + 3/4)$ does have an approximate normal distribution,

with unit variance (see p. 123, Johnson, Kotz and Kemp, 1995). Therefore we define the evidence against the null hypothesis for this one-sided alternative by $T = a_n(\tilde{p}) - a_n(p_0)$. Then, at least approximately, T is unit normal with expected value

$$\tau(p) = \mathrm{E}_{n,p}[\,T] \approx \{a_n(p) - a_n(p_0)\} - \frac{p - 0.5}{2\sqrt{np(1-p)}} \quad . \tag{1}$$

This T roughly satisfies properties $E_1 - E_4$. As an example, when $n = 9$, this two-term approximation to $\tau = \mathrm{E}_{9,p}[\,T]$ shown above is accurate to 0.05 for all $p \geq 0.5$ and the standard deviation $\mathrm{SD}_{9,p}[\,T]$ is within 0.05 of the target 1 for all $0.5 \leq p \leq 0.8$.

The maximum amount of evidence in a Binomial(n, p) experiment against $p = 0.5$ in favor of $p > 0.5$ occurs when $X = n$ and is $T_{max}(n) \approx \sqrt{n}\ \pi/2$. Thus to obtain 'strong' evidence against the null, the minimum sample size one needs must satisfy $5 \approx \sqrt{n}\ \pi/2$, or $n = 10$, and then one must observe $X = 10$. In the orthodox view, this sounds fairly difficult, for the p-value of this event would be $2^{-10} \approx 0.001$. But the p-value is computed under the null, and the null may well be false.

Many other examples of variance stabilizing transformations for test statistics are available in the references given below, but the requirements for a measure of evidence $E_1 - E_4$ are somewhat stronger. They make it easier to interpret evidence, to compare evidence obtained from different experiments, and to obtain simple confidence intervals for τ which can be converted into intervals for θ.

5 Evidence for heterogeneity

Given K studies measuring potentially different effects θ_k, for $k = 1, \ldots, K$ one often tests the null hypothesis of equal effects, or *homogeneity*, with an asymptotic Chi-squared test based on $Q = \sum_k w_k(\hat{\theta}_k - \hat{\theta}_w)^2$; Cochran (1954).

Unfortunately, when the weights in Q need to be estimated, the distribution of Q converges extremely slowly to its limit. But suppose it is possible to find evidence in the kth study $T_k \sim N(\tau_k, 1)$, where $\tau_k = \sqrt{n_k}\, m_k$, $m_k = m(\theta_k)$ and m is a monotone function free of k. Also let $\bar{m} = \sum n_k m_k / N$

be the weighted transformed effect, where $N = \sum_k n_k$. Then m_k can be estimated by $\hat{m}_k = T_k/\sqrt{n_k}$ and Cochran's $Q = \sum_k n_k(\hat{m}_k - \hat{\bar{m}})^2$; it measures heterogeneity of the m_k's directly, and of the θ_k's indirectly. Moreover this $Q \sim \chi^2_{K-1}(\lambda_Q)$, with $\lambda_Q = \sum_k n_k(m_k - \bar{m})^2$.

A variance stabilizing transformation of this Q to evidence is given by $T_Q = \{Q - \nu/2\}^{1/2} - \{\nu/2\}^{1/2}$, which satisfies $T_Q \sim N(\mathrm{E}[T_Q], 1)$ with $\mathrm{E}[T_Q] = \{\lambda_Q + \nu/2\}^{1/2} - \{\nu/2\}^{1/2}$. This T_Q satisfies $E_1 - E_4$ approximately, and is therefore a measure of *evidence for heterogeneity*.

6 Combinations of evidence on the probit scale

How one combines evidence in $\mathbf{T} = (T_1, \ldots, T_K)$ obtained in K studies depends on how much evidence T_Q one finds for heterogeneity of the θ_k's and on what specific alternative to the joint null $\theta_1 = \theta_2 = \ldots = \theta_K$ one wants evidence for. If there is only weak evidence for heterogeneity, one can assume the standard fixed effects model (all $\theta_k = \theta$) and find the evidence for $\theta > 0$ using $T_w = \sum_k \sqrt{w_k}\, T_k$, where $\sum_k w_k = 1$. Then, because $\tau_k = \sqrt{n_k}\, m(\theta)$, $T_w \sim N(\tau_w, 1)$ with $\tau_w = \sum_k \sqrt{w_k}\, \tau_k = m(\theta) \sum_k \sqrt{w_k n_k}$. A possible choice for $w_k = n_k / N$. Obvious confidence intervals for τ_w are easily transformed into intervals for θ, if desired.

If one chooses a fixed, but *unequal* effects model then there are several possible alternative hypotheses. For example, one can define an overall effect as the θ which transforms into a weighted average of the transformed effects $m_1, \ldots, m_K$ and find evidence for $\theta > 0$. This methodology is illustrated for one- and two-sample t-tests in Kulinskaya and Staudte (2006). Finally, one can assume a random transformed effects model which introduces a variance component on the range of the map m. Then inference on $\mu = m(\theta)$ can be carried out and transformed back into inference for $\theta = m^{-1}(\mu)$.

7 Summary

By means of variance stabilization, many routine test statistics can be transformed onto a calibration scale that allows for easy interpretation of results, and comparison and combination of evidence obtained in similar independent

experiments. While the proposed framework only leads to measures of evidence which are approximately normal, this has not been a hindrance to the greater goals of interpretation, combination and repeatability of evidence. It is basically a meta-analytic framework with *known* weights.

References

[1] F.J. Anscombe. The transformation of Poisson, binomial and negative binomial data. *Biometrika*, 35:266–254, 1948.

[2] P.F. Azorin. Sobre la distribución t no central I,II. *Trabajos de Estadistica*, 4:173–198 and 307–337, 1953.

[3] B. Efron. Transformation theory: How normal is a family of distributions? *The Annals of Statistics*, 10(2):323–339, 1982.

[4] N.L. Johnson, S. Kotz, and N. Balakrishnah. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, 1994.

[5] N.L. Johnson, S. Kotz, and N. Balakrishnah. *Continuous Univariate Distributions*, volume 2. John Wiley & Sons, New York, 1995.

[6] N.L. Johnson, S. Kotz, and A.W. Kemp. *Univariate Discrete Distributions*. Wiley, New York, second edition, 1993.

[7] E. Kulinskaya and R. G. Staudte. Confidence intervals for the standardized effect arising in comparisons of two normal populations. 2006. La Trobe University Technical Report No. 2006-4.

[8] S. Morgenthaler and R.G. Staudte. Calibrating significant p-values. 2005. Submitted for publication.

[9] C.D. Mulrow, E. Chiquette, L. Angel, J. Cornell, C. Summerbell, B. Anagnosetelis, M. Brand, and R.Jr. Grimm. Dieting to reduce body weight for controlling hypertension in adults (Cochran Review). In *The Cochran Library,* Issue 3. John Wiley & Sons, Chichester, UK, 2004.

[10] T. Selke, M.J. Bayarri, and J.O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55:62–71, 2001.