

A Decision-Theoretic Formulation of Fisher's Approach to Testing

Kenneth Rice,

Department of Biostatistics, University of Washington

`kenrice@u.washington.edu`

November 12, 2010

Abstract

In Fisher's interpretation of statistical testing, a test is seen as a 'screening' procedure; one either reports some scientific findings, or alternatively gives no firm conclusions. These choices differ fundamentally from hypothesis testing, in the style of Neyman and Pearson, which do not consider a non-committal response; tests are developed as choices between two complimentary hypotheses, typically labeled 'null' and 'alternative'. The same choices are presented in typical Bayesian tests, where Bayes Factors are used to judge the relative support for a null or alternative model. In this paper, we use decision theory to show that Bayesian tests can also describe Fisher-style 'screening' procedures, and that such approaches lead directly to Bayesian analogs of the Wald test and two-sided p -value, and to Bayesian tests with frequentist properties that can be determined easily and accurately. In contrast to hypothesis testing, these 'screening' decisions do not exhibit the Lindley/Jeffreys paradox, that divides frequentists and Bayesians.

1 Introduction and Review

Testing, and the interpretation of tests, is a controversial statistical area (Schervish, 1996; Hubbard and Bayarri, 2003; Berger, 2003; Murdoch et al., 2008; Sackrowitz and Samuel-

Cahn, 1999; Pereira et al., 2008). Christensen (2005) reviews this controversy, discussing both the strong disagreement between Fisherian and Neyman-Pearson schools of testing as well as frequentist-Bayesian differences. In this article we aim to dissipate some of this controversy, by describing Fisher’s approach to testing within the Bayesian decision-theoretic paradigm. This approach gives valid Bayesian tests with straightforward frequentist properties. Moreover, the decision-theoretic development relies on the user making a clear statement of their testing goals, which may or may not align with hypothesis-testing. It is hoped this will help the user to choose which type of test, if any, is appropriate for their application.

In this paper, we characterize statistical tests as calibrated procedures, producing binary decisions based on data and assumptions about the true state of nature. Specific decisions and calibrations define particular types of tests.

A ‘hypothesis test’ produces a decision choosing between two complementary hypotheses; either $H_0 : \theta \in \Theta_0$ or $H_1 : \theta \in \Theta_1$. Hypothesis tests are calibrated by considering penalties for incorrect decisions, i.e. for errors of Type I or Type II, made when we incorrectly choose H_1 or H_0 , respectively. (We note that many modern interpretations differ slightly from Neyman and Pearson’s original development, and in particular do not interpret this choice as “acceptance” of a true state of nature. See, for example, Hubbard and Armstrong (2006) or Hurlbert and Lombardi (2009).) Frequentist hypothesis tests are calibrated through specifying the Type I error rate, or an upper bound on this quantity. Optimal frequentist hypothesis tests are obtained (minimizing the Type II error rate, under this calibrating constraint) when the test statistic is the likelihood ratio, or its asymptotically-justified transformation to the likelihood ratio test p -value. In contrast, Bayesians typically calibrate hypothesis tests by ascribing relative costs to Type I and II errors. Their optimal decision, the Bayes rule, minimizes the expected total costs, where the expectation is taken over posterior uncertainty. This calibration leads to use of Bayes Factors (see, for example, Moreno and Girón, 2006). Other, less conventional calibrations of Bayesian testing can be seen in e.g. Madruga et al. (2001), or Casella, Hwang and Robert (1993). While it is well-known that use of p -values and Bayes Factors can lead to hypothesis tests with strongly divergent results – the ‘Lindley/Jeffreys’ paradox (Lindley, 1957) – the underlying binary choice of either H_0 or H_1 is common to both paradigms.

We define a ‘Fisherian’ test as one where the underlying decision is whether a report on some scientific quantity is merited, or not. The Fisherian test is a screening procedure where, given prior knowledge and observed data, our assessment is ‘interesting’ or ‘not worth comment’. This latter choice is not a confirmation or refutation, but is simply non-committal, and cannot be a Type II error. We have used the term ‘Fisherian’ to attribute this formulation to Fisher, who interpreted negative results to mean that “nothing could be concluded.” (Schmidt, 1996). (‘Popperian’ would also be an appropriate term, after Karl Popper’s philosophy of science; see Mayo and Cox (2006), pg 80.) To be completely clear, in Fisherian testing the ‘null’ decision means that, although *someone* might conclude *something* from the data, the present analyst chooses to withhold judgment.

In many applications Fisherian tests may not be the approach of choice. However, motivation for calibrated statistical screening can be seen in e.g. contemporary genome-wide association studies (Pearson and Manolio, 2008), where variations at very many genetic loci are assessed for association with a disease outcome. In this popular form of study, results from ‘interesting’ loci are reported, and are typically followed up with independent data – it is standard practice to defer claims of true association until positive replication is obtained. Loci with non-significant results are not followed up, but in practice neither are these loci claimed to have truly null associations. In fact, beyond their non-significance these loci receive no comment, and Fisher’s conclusion of ‘nothing’ is apt. We see that, for both positive and negative outcomes, the Fisherian test’s conclusions directly reflect real contemporary scientific decisions.

Before continuing, we note that in this paper the term ‘Bayesian’ only implies use of Bayes’ calculus, expressing knowledge via probability distributions. Similarly, ‘frequentist’ implies consideration of repeatedly applying procedures to similarly-generated datasets. These criteria are not mutually exclusive, and we will not treat them in that way. Characterization of the procedures as ‘objective’ or ‘subjective’ is a choice we leave entirely to the reader.

2 Fisherian testing: a Bayesian approach

In this section we develop the Fisherian testing argument, in the language of decision theory. (This formalizes recent consideration of ‘screening’ behavior to calibrate applied Bayesian tests (Datta and Datta, 2005; Guo and Heitjan, 2010)). We shall make a binary decision $h \in \{0, 1\}$ that is pertinent to $\theta \in \mathbb{R}$, a univariate parameter of interest. The decision $h=0$ is interpreted as not making a report about θ , and ‘concluding nothing’. The alternative, $h=1$, is a decision to report findings. Both decisions have costs dependent on the true state of nature, and these will be specified in a loss function. For simplicity our approach is fully parametric; we proceed with analysis conditional on the assumed validity of a parametric model, whose distribution function depends on θ and a fixed finite number of nuisance parameters.

Our loss function for binary decision h may be expressed in two parts. First, if we decide to ‘report’ ($h=1$) then we have to state what is known about θ . This motivates making a decision d which is real-valued, i.e. which is measured in the same units as θ . In this paper we will use a point estimate as the decision d , and will penalize d according to the familiar squared error loss $(d - \theta)^2$. Intuitively, if our point estimate is far from θ , this may be interpreted as the need for extensive follow-up to establish its value with accuracy, requiring more data, and incurring loss. Secondly, if we ‘conclude nothing’ ($h=0$), then we must pay for missing important results. This can be interpreted as either embarrassment to the analyst or as a loss to society, failing to realize a potential benefit. If some ‘null’ value θ_0 exists (defined here as a true state of nature we would be content to ‘miss’ reporting) then it makes sense to penalize $h=0$ heavily if θ is far from θ_0 , and very little if $\theta \approx \theta_0$. This motivates loss proportional to $(\theta - \theta_0)^2$. Importantly, when $h=0$ our loss will *ignore* d , as the loss is unchanged for any $d \in \mathbb{R}$. In this way the decision is completely non-committal about the true value of θ , and reflects Fisher’s conclusion of ‘nothing’.

All that remains is to combine the two squared-error losses. As these are measured in the same units, we can reasonably just add them together. Simple addition of the two terms gives a utility which exactly equates squared loss around d (when $h=1$) with squared loss around θ_0 (when $h=0$). More generally, we wish to express skepticism about making the decision to ‘report’, which is achieved by making the loss around θ_0 cheaper by a constant factor $\gamma \leq 1$. γ

is a ratio of costs; it is positive and unitless, and states how many units of inaccuracy (around d) one would trade for a unit of embarrassment (around θ_0).

Motivated in this manner, we have the loss function

$$L_\gamma = \gamma^{1/2}(1 - h)(\theta - \theta_0)^2 + \gamma^{-1/2}h(\theta - d)^2.$$

(Our choice to weight the contributions by $\gamma^{1/2}$ and $\gamma^{-1/2}$ is to emphasize that the product of the two weight terms is fixed; this form also appears in Section 3.) To set the value of γ , it may help to consider specific circumstances. For example, with $\theta_0 = 0$, if ‘missing’ a true value of $\theta = 3$ is as bad as the embarrassment caused by reporting $d = \theta + 2$ or $\theta - 2$, then $\gamma = 2^2/3^2 = 0.44$ can be used. Other choices of γ are given at the end of this section.

The Bayes rule for L_γ can be determined straightforwardly; conditional on deciding $h=1$, we minimize the expected value of L_γ by choosing d to be the posterior mean. We denote this as $d = \mathbb{E}[\theta|Y]$, where Y denotes the observed data. As noted above, if $h=0$ any value of d provides equal utility. Hence, conditional on data Y , the Bayes rule decides $h=1$ if and only if

$$\begin{aligned} \text{Var}[\theta|Y] &\leq \gamma \mathbb{E}[(\theta - \theta_0)^2|Y] \\ &= \gamma (\text{Var}[\theta|Y] + \mathbb{E}[\theta - \theta_0|Y]^2). \end{aligned}$$

Re-arranging this, we see that the Bayes rule decides $h=1$ if and only if

$$\frac{\mathbb{E}[\theta - \theta_0|Y]^2}{\text{Var}[\theta|Y]} \geq \frac{1 - \gamma}{\gamma}.$$

Our restriction to $0 < \gamma \leq 1$ above means that the threshold for deciding $h=1$ can be any positive number.

It is immediately apparent that the Bayes rule is of a similar form to the Wald test, which rejects the null hypothesis whenever

$$\frac{(\hat{\theta} - \theta_0)^2}{\widehat{\text{Var}}(\hat{\theta})} \geq \chi_{1, 1-\alpha}^2.$$

Here $\hat{\theta}$ is an estimate of θ , such as the maximum likelihood estimate, and the denominator is an estimate of the variance of $\hat{\theta}$, such as the reciprocal of the estimated Fisher information. As with the Bayesian test, different choices of α allow the threshold for rejection to take any positive value.

In a frequentist calibration of Wald tests, the choice of α is motivated by considering binary decisions in large numbers of experiments for which θ is exactly θ_0 . In our Bayesian approach, the choice of γ is instead motivated by considering relative costs of quantities directly related to θ , the true state of nature, which need not be exactly θ_0 . Despite these differences, by Bernstein-von Mises theorem (i.e. asymptotic agreement of likelihood- and posterior-based methods) equating $(1 - \gamma)/\gamma$ and $\chi_{1, 1-\alpha}^2$ provides asymptotic equivalence of the MLE-based Wald test and our Bayesian test, under mild regularity conditions on the prior and model (van der Vaart 2000, Chapter 10). In particular, using $\gamma = 0.21$ (i.e. a ratio of costs of approximately 1:5) gives a threshold for deciding $h=1$ at $(1 - \gamma)/\gamma = \chi_{1, 0.95}^2 = 3.84 = 1.96^2$, and good agreement with the common use of $\alpha = 0.05$. This is illustrated in Section 4.

3 Wald test p -values: a Bayesian approach

In Section 2 we developed Fisherian tests in a Bayesian framework, viewing them as an optimal choice between two types of loss being traded-off at a given rate. An equivalent optimization problem (formally, the mathematical ‘dual’ of this problem) is to determine the optimal trade-off rate. That is, if the loss is a weighted sum of embarrassment and inaccuracy, but now we get to choose the weighting based on our posterior knowledge, what choice minimizes our expected loss?

A loss function quantifying this dual decision problem is

$$L' = \frac{1}{\sqrt{1+w}}(\theta - \theta_0)^2 + \sqrt{1+w}(d - \theta)^2,$$

where $d \in \mathbb{R}$ as before, and $w \in \mathbb{R}^+$ is a positive real-valued decision describing the optimal weighting. In line with the skepticism noted in Section 2, we have set up the decision so that the loss attached to one unit of inaccuracy ($\sqrt{1+w}$) is greater than that attached to one unit of embarrassment ($1/\sqrt{1+w}$).

The Bayes rule for d is $\mathbb{E}[\theta|Y]$ as before, hence the Bayes rule for w minimizes

$$\frac{1}{\sqrt{1+w}}\mathbb{E}[(\theta - \theta_0)^2 | Y] + \sqrt{1+w} \text{Var}[\theta|Y] = \frac{1}{\sqrt{1+w}} \left(\text{Var}[\theta|Y] + \mathbb{E}[\theta - \theta_0 | Y]^2 \right) + \sqrt{1+w} \text{Var}[\theta|Y].$$

We see that the Bayes rule sets

$$w = \frac{\mathbb{E}[\theta - \theta_0|Y]^2}{\text{Var}[\theta|Y]}.$$

Modulo the influence of the prior on the posterior moments, the Bayes rule for w is therefore the classical two-sided Wald test statistic, $(\hat{\theta} - \theta_0)^2 / \widehat{\text{Var}}(\hat{\theta})$. Transforming w via the cumulative distribution function of χ_1^2 , we see that the Wald test's p -value is a Bayes rule, up to the same asymptotic approximation as in Section 2.

Just as the p -value can be interpreted as a summary of all the Wald tests one might perform (at all α thresholds), the decision w provides a summary of what the original Bayesian testing decision would be, at different γ . Re-phrased, both of these are optimal decisions in their own right; p tells us the maximum α at which the null would be rejected, and w assesses, in one particular manner, the optimal balance between losses associated with possible scientific reporting (inaccuracy) versus losses from not reporting (embarrassment).

In Fisherian tests, we see that the two-sided p -value has a close Bayesian analog. This agreement between Bayes and frequentism is not seen in hypothesis tests, where the behavior of two-sided p -values contrasts sharply with Bayesian measures comparing the evidence that $\theta = \theta_0$ versus some other value. Attempts to interpret the p -value as such a measure have been strongly criticized (Schervish, 1996; Berger and Sellke, 1987); the results in this paper suggest that these disputes need not occur in every form of testing.

4 Example: Hardy-Weinberg equilibrium

In this section we develop Fisherian tests of Hardy-Weinberg equilibrium (HWE), a parametric constraint encountered in categorical genetic data. For the common case of a bi-allelic marker, the data from n unrelated subjects can be denoted

$$\begin{array}{c|ccc} \text{genotype} & AA & Aa & aa \\ \hline \text{count} & n_{AA} & n_{Aa} & n_{aa} \end{array},$$

and the data modeled as n draws from a multinomial distribution, with cell probabilities (p_{AA}, p_{Aa}, p_{aa}) in the unit simplex. Under HWE the two alleles are assigned independently,

and the cell probabilities lie on a curve parameterized as

$$(p_{AA}, p_{Aa}, p_{aa}) = (p_A^2, 2p_A(1 - p_A), (1 - p_A)^2)$$

where p_A denotes the probability that an allele is of type A . Departures from HWE can be explained by several phenomena (Weir, 1996); we shall focus on potential inbreeding, which could invalidate the assumed independence.

Under our Fisherian testing approach, we will either ‘conclude nothing’ or report some level of inbreeding. This is in line with many applications of HWE tests, which screen for notable divergence from HWE ($h=1$) but otherwise do not affect subsequent analysis ($h=0$). To quantify divergence from HWE, we define the inbreeding coefficient;

$$\theta = \frac{2(p_{aa} + p_{AA}) - 1 - (p_{aa} - p_{AA})^2}{1 - (p_{aa} - p_{AA})^2}.$$

Positive and negative values of θ correspond (respectively) to excess and deficient numbers of Aa observations, relative to HWE (Weir, 1996). We will take the null value to be $\theta_0 = 0$, which corresponds to exactly no inbreeding. For simplicity of exposition, we will use a *Dirichlet*(1, 1, 1) prior on (p_{AA}, p_{Aa}, p_{aa}) ; this gives uniform support to all values of (p_{AA}, p_{Aa}, p_{aa}) , and support over the range of $-1 < \theta < 1$. Of note, it gives zero support to $\theta = \theta_0 = 0$, i.e. actually ruling out states of nature perfectly free from inbreeding. This assumption would also be typical of an elicited, substantive prior.

We implemented this test for all values of n_{AA}, n_{Aa} and n_{aa} that have total sample size $n = 200$. Numerical integration using quadrature provides the posterior mean and variance of θ ; the Bayes rule decisions d and h then follow as described in Section 2. Using $\gamma = 0.21$, as suggested in Section 2, we obtain Figure 1. We see that decisions $h=0$ cluster around the line denoting perfect HWE.

The frequentist properties of this rule are also easily computed; under perfect HWE and for a given p_A we can compute the long-run proportion of datasets under which we make decision $h=1$. This proportion is plotted in Figure 2, together with the same frequency obtained using a standard Pearson’s χ^2 test. This asymptotically-justified test can also be motivated by direct consideration of θ , hence its inclusion here. (To reflect its use in practice, whenever the Pearson

χ^2 test statistic is infinite or undefined ($0/0$), $h=0$ was returned. Without this step, the χ^2 test has a grossly-inflated Type I error rate for extreme p_A .)

Despite using a prior which rules out models with perfect HWE (i.e. with $\theta = 0$), the Bayes procedure actually outperforms the Pearson test, which is calibrated at this value. For central p_A , although the Bayes rule is slightly anti-conservative, the Pearson test is worse at this sample size. For extreme p_A (where appeals to asymptotic properties are weakest) the Bayesian addition of prior information moderates extreme behavior of the test statistic, precluding values such as ∞ or $0/0$.

[Figure 1 about here.]

[Figure 2 about here.]

5 Contrasts with hypothesis-testing approaches

The justification of Fisherian testing with L_γ and related loss functions in Sections 2 and 3 is intended to be short and uncomplicated, and Section 4 is intended to illustrate that its use need not be difficult or controversial. However, in our experience the Neyman-Pearson school of hypothesis-testing is deeply ingrained in statistical thought, and many statisticians seem to have trouble conceptualizing testing other than in the Neyman-Pearson setup. Christensen (2005), Schervish (1996), and Lavine and Schervish (1999) all give excellent reviews of what various existing frequentist and Bayesian tests do provide; in this section we expand on how these familiar approaches can differ from our Fisherian tests.

Being non-committal may be an appropriate outcome. Experienced statisticians will know that different datasets support conclusions such as “the hypothesis is right”, “the hypothesis is wrong” and – perhaps all too often – “this data is so unhelpful that no-one can tell”. Fisherian tests which conclude ‘nothing’ ($h=0$) accord with the outcome that ‘no-one can tell’. Trying to make such a conclusion with hypothesis-testing Bayes Factors is unsatisfactory. To represent the null, one can attempt to include a very flexible ‘Model Zero’ among the list of putative models, but Bayes Factors can be very sensitive to the choice of prior on nuisance parameters, so one’s

decision may be sensitive to essentially arbitrary choices. In particular, for the limiting case of improper flat priors the Bayes Factor is not well-defined, and testing in this way cannot proceed at all. In contrast, our Fisherian tests explicitly permit non-committal Bayesian decisions, without any requirement that one construct non-committal priors or posteriors.

Model-selection is not necessary for testing. In parametric settings, two-sided Bayesian hypothesis tests force users to choose between two distinct data-generating mechanisms. ‘Lump and smear’ mixtures (also known as ‘spike and slab’ mixtures) are used, where ‘lumps’ in the model (or prior) support precise null hypotheses, and ‘smears’ of continuous support describe what is known about θ under the alternative. Tests based on these models typically decide in favor of either the lump sub-model or the smear, depending on which hypothesis is better supported. As seen in Section 4, Fisherian testing does not require these multi-part models. While not precluding lump and smear approaches, we have seen that Fisherian testing can be performed with priors and likelihoods which are smooth in θ , giving zero prior weight to θ being exactly θ_0 . In many applied settings it can also seem artificial to elicit precise prior support for the singular value $\theta = \theta_0$ (Garthwaite and Dickey, 1992), making Fisherian tests more natural for use with smooth priors.

Larger sample sizes may motivate reduced significance thresholds. While the long-run statistical properties of a the standard fixed- α approach are unambiguous, the scientific appropriateness of fixing the Type I error rate without regard to n may be challenged. Specifically, under the standard approach, as n increases (keeping everything else fixed) equally ‘significant’ results can provide more evidence against H_0 . To many, it is unsatisfactory that decision rules of the form $p < \alpha$ ignore this increase in evidence. One resolution of the phenomenon (Cox and Hinkley, 2000; Wakefield, 2008) notes that large n might be *a priori* motivated by beliefs that θ is near θ_0 , and thus hard to detect. The increased skepticism suggested by larger n cancels the greater evidence from the data, restoring decisions based on $p < \alpha$. Of course, our Bayesian derivation in Section 2 allows skepticism to be built into priors, and putting more prior mass near $\theta = \theta_0$ will lead to more skeptical results in finite samples, when γ is kept constant. But the decision-theoretic approach also allows n to enter in a different way, through the choice of γ . Quite plausibly, n may be small because we – or our funding body – are largely indifferent to

the value of θ . Such indifference is not reflected in the prior’s summary of scientific knowledge about θ , nor in the likelihood, but is naturally expressed in a loss function. If θ will only be of interest if it is far from θ_0 , this motivates using a small γ . If we did a larger study because we are keen to know θ and have little inclination to ‘conclude nothing’, then this motivates a bigger choice of γ , until at $\gamma = 1$ we reach the situation where $h=1$ for any data. This most extreme case is not pathological; if any point estimate from our study will be of interest, and it is appropriate that the ‘null’ decision $h=0$ never occurs.

6 Conclusion

We have provided a direct, constructive, decision-theoretic derivation of Fisher’s ‘screening’ interpretation of testing, where we trade-off comparable quantities at rates determined through *a priori* consideration of our interest in a scientific parameter of interest. In the cases discussed, it answers Schervish’s (1996) question of what p -values can measure, by deriving the p -value as a measure of relative support for two competing decisions. Perhaps most importantly, the decision-theoretic approach inherently makes up-front, unambiguous statements about how the testing procedure is justified, and does so in terms of θ , the natural scale for scientific inquiry. Furthermore, in this derivation we have tried to make the role of the different tools very clear; the prior and model express what we assume about θ and how data is generated, while the loss function expresses what we want to say about θ – and how badly we want to say it.

This clarity is appealing, because explicit statements of the question of interest force one to consider what *is* appropriate. In the Fisherian approach to testing, should we ‘conclude nothing’ from data? In high-throughput settings like genome-wide studies, current practices suggest this is acceptable. Elsewhere, if the inferential goal is a comprehensive summary of what is known, then reporting ‘nothing’ is inappropriate, and the form of L_γ indicates it should not be used. (We note that any single binary decision or one-number summary would also not suffice as such a summary). In perhaps still rarer applied settings, the scientific question may be a direct choice between two putative states of nature. If knowledge of θ is not of interest beyond knowing whether $\theta = \theta_0$ or some other value, then here too we see that L_γ and its Bayes rules

are inappropriate.

Across many settings, we therefore find that the ‘language’ of utility can clarify and quantify how one should make inference. Distinguishing between use and misuse of procedures is often subtle and difficult, even for experienced statisticians with ample common sense, so we feel that more widespread use of decision-theory may be beneficial.

References

- Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32.
- Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence (C/R: P123-133, 135-139, 1201-1201). *Journal of the American Statistical Association*, 82:112–122.
- Casella, G., Hwang, J. T. G., and Robert, C. (1993). A paradox in decision-theoretic interval estimation. *Statistica Sinica*, 3:141–155.
- Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59(2):121–126.
- Cox, D. R. and Hinkley, D. V. (2000). *Theoretical Statistics*. Chapman & Hall Ltd.
- Datta, S. and Datta, S. (2005). Empirical bayes screening of many p-values with applications to microarray studies. *Bioinformatics*, 21(9):1987–1994.
- Garthwaite, P. H. and Dickey, J. M. (1992). Elicitation of prior distributions for variable-selection problems in regression. *The Annals of Statistics*, 20:1697–1719.
- Guo, M. and Heitjan, D. (2010). Multiplicity-calibrated Bayesian hypothesis tests. *Biostatistics*, in press.
- Hubbard, R. and Armstrong, J. (2006). Why we don’t really know what statistical significance means: a major educational failure. *Journal of Marketing Education*, 28:114–120.

- Hubbard, R. and Bayarri, M. J. (2003). Confusion Over Measures of Evidence (p 's) Versus Errors (α 's) in Classical Statistical Testing. *The American Statistician*, 57(3):171–178.
- Hurlbert, S. and Lombardi, C. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, 46(5):311–349.
- Lavine, M. and Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, 53:119–122.
- Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192.
- Madruga, M. R., Esteves, L. G., Esteves, L. G., and Wechsler, S. (2001). On the Bayesianity of Pereira-Stern tests. *Test (Madrid)*, 10(2):291–299.
- Mayo, D. and Cox, D. (2006). Frequentist statistics as a theory of inductive inference. *Lecture Notes-Monograph Series*, 49:77–97.
- Moreno, E. and Girón, F. J. (2006). On the frequentist and bayesian approaches to hypothesis testing. *Statistics & Operations Research Transactions*, pages 3–28.
- Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). P-Values are Random Variables. *The American Statistician*, 62(3):242–245.
- Pearson, T. and Manolio, T. (2008). How to interpret a genome-wide association study. *Journal of the American Medical Association*, 299(11):1335–1344.
- Pereira, C. A. d. B., Stern, J. M., and Wechsler, S. (2008). Can a significance test be genuinely bayesian? *Bayesian Analysis*, 3(1):79–100.
- Sackrowitz, H. and Samuel-Cahn, E. (1999). P values as random variables — Expected P values. *The American Statistician*, 53:326–331.
- Schervish, M. J. (1996). P values: What they are and what they are not. *The American Statistician*, 50:203–206.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2):115–129.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Wakefield, J. (2008). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology*, page in press.

Weir, B. (1996). *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

List of Figures

- 1 Bayes rule for all data sets with $n = 200$, using L_γ with θ as the inbreeding coefficient, a Dirichlet prior on the unit simplex, and ratio of costs $\gamma = 0.21$. The dashed line indicates values that would be in perfect Hardy-Weinberg equilibrium 16
- 2 Frequentist evaluation of the Bayes rule in Figure 1, under $\theta = 0$. For comparison, Pearson χ^2 tests based on the inbreeding coefficient are also evaluated. When the χ^2 test statistic is infinite or undefined, $h=0$ is returned; without this modification, the Type I error rate is grossly inflated at low p_A 17

Bayes rule for significance tests of HWE, $n=200$, $\gamma = 0.21$

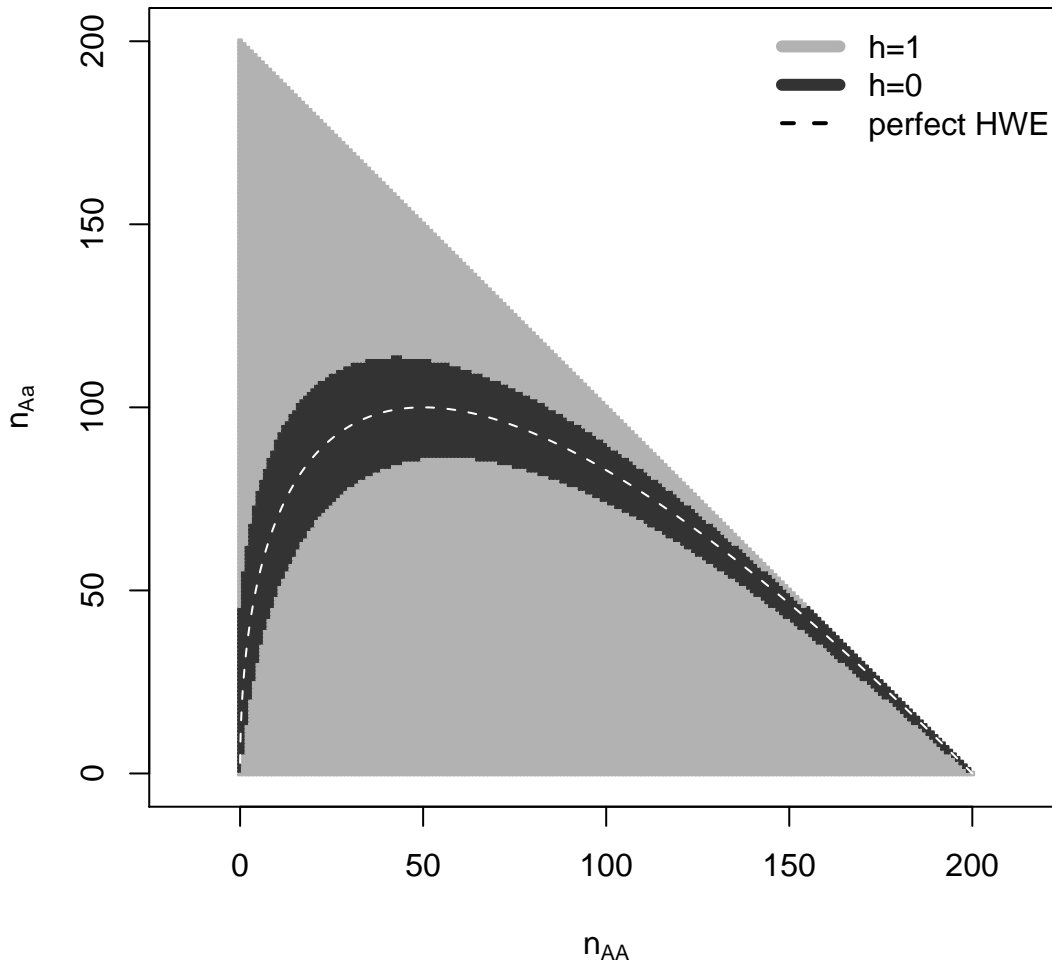


Figure 1: Bayes rule for all data sets with $n = 200$, using L_γ with θ as the inbreeding coefficient, a Dirichlet prior on the unit simplex, and ratio of costs $\gamma = 0.21$. The dashed line indicates values that would be in perfect Hardy-Weinberg equilibrium

Significance Tests of HWE/inbreeding: n=200

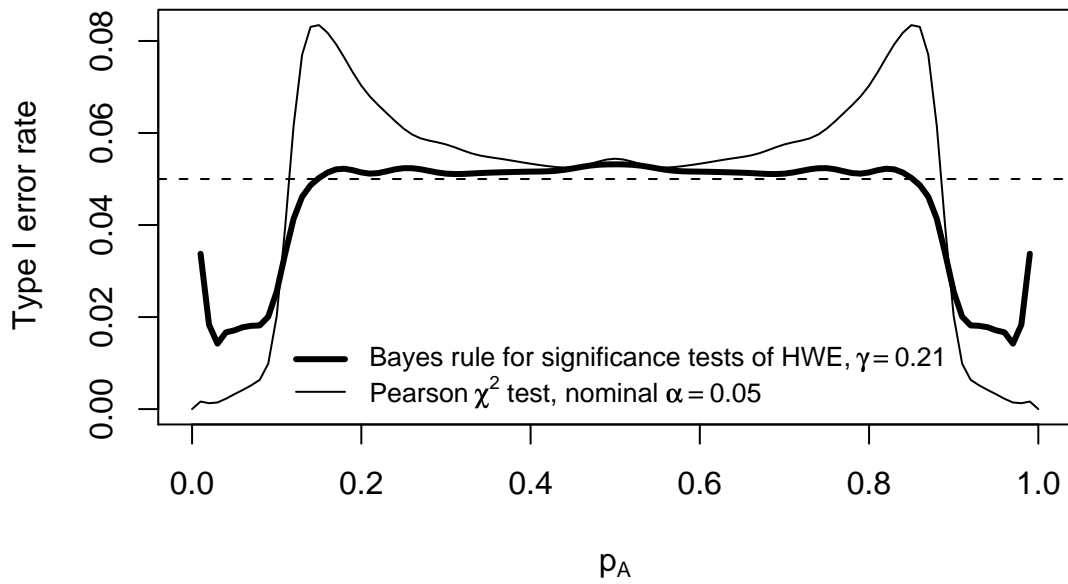


Figure 2: Frequentist evaluation of the Bayes rule in Figure 1, under $\theta = 0$. For comparison, Pearson χ^2 tests based on the inbreeding coefficient are also evaluated. When the χ^2 test statistic is infinite or undefined, $h=0$ is returned; without this modification, the Type I error rate is grossly inflated at low p_A .