Bayes Factors for Genome-Wide Association Studies: Comparison with p-values

Jon Wakefield

Running title: Bayes Factors for Genome-Wide Studies

Jon Wakefield

Address:

Departments of Statistics and Biostatistics,

University of Washington,

Box 357232

Seattle, WA 98195–7232,

USA

Telephone: 206–616–6292

Email: jonno@u.washington.edu

Abstract

The Bayes factor is a summary measure that provides an alternative to the p-value for the ranking of associations, or the flagging of associations as "significant". We describe an approximate Bayes factor that is straightforward to use and is appropriate when sample sizes are large. We consider various choices of the prior on the effect size, including those that allow effect size to vary with the minor allele frequency of the marker. An important contribution is the description of a specific prior that gives identical rankings between Bayes factors and p-values, providing a link between the two approaches, and allowing the implications of the use of p-values to be more easily understood. As a summary measure of noteworthiness p-values are difficult to calibrate since their interpretation depends on minor allele frequency and, crucially, on sample size. A consequence is that a consistent decision-making procedure using p-values requires a threshold for significance that reduces with sample size, contrary to common practice. We outline how Bayes factors may be simply calculated under a variety of sampling schemes, including those in which imputed SNPs are available, or the response is a survival endpoint.

Keywords: Bonferroni correction; p-values; Prior distributions; Strength of evidence.

Introduction

Genome-wide association studies (GWASs) are an exciting prospect for discovering genetic variants that are detrimental or protective for human disease (Hirschhorn and Daly 2005; Wang et al. 2005), and a number of important results have already been reported (DeWan et al. 2006; Sladek et al. 2007; Easton et al. 2007; Consortium 2007).

The most common summary measure for inference in a GWAS is the p-value; p-values have a number of well-documented drawbacks, however (Sterne and Smith 2001; Goodman 1999; Wacholder et al. 2004). Many recommendations for significance thresholds have appeared and the majority depend on the number of tests that are performed, and on neither the sample size of the study, nor the minor allele frequency (MAF) of the SNP. The use of a single threshold regardless of sample size implicitly implies that the ratio of costs of type I to type II errors varies with sample size. Specifically, consider two situations with low power in the first and high power in the second; if the p-value threshold is the same in boh situations then one is accepting that the cost of a type II error is higher in the second situation. Later we give a precise formulation of this argument.

The Bayes factor, defined as the ratio of the probability of the data under the null and alternative hypotheses, provides an alternative to the p-value for assessing the consistency of a set of data with a null hypothesis, as compared to the alternative. Bayes factors have been previously discussed in a genome-wide context (Wakefield 2007; Marchini et al. 2007; Consortium 2007), and also used in other genetic settings (Servin and Stephens 2007). The more widespread use of the Bayes factor has been hampered by the need for prior distributions to be specified for all of the unknown parameters in the model, and the need to evaluate multi-dimensional integrals, a complex computational task. In this paper we provide a more rigorous derivation of a recently-proposed asymptotic Bayes factor (Wakefield

2007), that avoids each of these requirements. An important additional contribution is the description of a specific prior on the effect size that leads to identical rankings between SNPs based on the p-value and on the asymptotic Bayes factor, providing an important conceptual link between the two and allowing a Bayesian interpretation of the p-value. In particular, the p-value implicitly assumes a prior in which larger effect sizes at lower MAFs are expected.

Methods

Bayes Factors

In a GWAS, a common approach (Balding 2006) is to fit the logistic model:

$$\frac{p_j}{1-p_j} = \exp(\alpha + \theta x_j) \tag{1}$$

where p_j is the probability of disease for an individual with $j = 0, 1, 2$ copies of the mutant allele at a particular SNP, and x_j is a variable that depends on the assumed genetic model. For $j = 0, 1, 2$ copies we have $x_j = 0, 1, 1$, for a dominant genetic model, $x_j = 0, 0, 1$ for a recessive genetic model, and $x_j = 0, 1, 2$ for a multiplicative genetic model (Sasieni 1997). The Bayes factor for the general two degree of freedom model can also be considered, but for simplicity of explanation we concentrate on genetic models in which there is a single parameter of interest θ. Model (1) may be easily extended to include matching and other variables for which adjustment is required, as detailed in the appendix.

Under a rare disease assumption, the relative risk corresponding to departure from the null model is given by RR$= \exp(\theta)$, and we wish to compare the null hypothesis $H_0 : \text{RR} = 1$ with the alternative $H_1 : \text{RR} \neq 1$. As described elsewhere (Wakefield 2008) there are two endeavors that may be carried out in the context of a genome-wide association study. The

first is to *rank* markers in terms of association, to provide a list of those that should be carried through to a next phase. The second is *calibration* of inference to make a final decision as to whether to call the marker "significant", i.e. associated with disease, or not. Each of these tasks may be carried out using the Bayes factor (BF) given by

$$\text{BF} = \frac{\Pr(\boldsymbol{y}|H_0)}{\Pr(\boldsymbol{y}|H_1)}$$

where $\boldsymbol{y}$ is the observed data, and corresponds to a vector of binary indicators when disease status is the phenotype. If the Bayes factor equals 1 then the data are equally likely under the null and the alternative, and the smaller/larger the Bayes factor becomes the more/less the alternative is favored. For a measure of "significance" the posterior odds on H_0 are required:

$$\text{Posterior odds on } H_0 \ = \text{BF} \times \ \text{Prior Odds on } H_0$$

where the prior odds on H_0 are given by $\pi_0/(1-\pi_0)$, with $\pi_0 = \Pr(H_0)$ the prior probability of the null. Ranking may be carried out directly on the basis of the Bayes factor if the prior odds of no association is constant across all SNPs, since the relative value rather than the absolute value is all that is needed.

The Bayes factor requires the specification of a prior distribution for all unknown parameters, and for logistic regression models it is computationally expensive to evaluate, which has lead to the search for simple approximations. In the Wellcome Case Control Consortium study (Consortium 2007), Bayes factors were calculated using the Laplace approximation (Kass and Raftery 1995). This approximation can be difficult to implement, however, since a search for the maximum of the multidimensional posterior is required for each association. Below we describe an alternative asymptotic Bayes factor that is based on the output from a simple logistic regression analysis; the only data input required for Bayes factor calculation is a confidence interval for the parameter of interest θ (or equivalently an estimate and standard

error). Maximization of a binomial likelihood is required when a logistic regression model is fitted, but this operation is routinely carried out by all statistical packages. The Bayes factor we describe has a simple closed form, which offers a number of benefits including ease of power calculations, and straightforward combination of evidence across studies.

Let $\widehat{\theta}$ and $\sqrt{V}$ represent the maximum likelihood estimate (MLE) and standard error from a logistic regression analysis; V depends on the genetic model, the case and control sample sizes, n_1 and n_0, and on the MAF (equation (8) gives the specific form). Asymptotically, that is as n_0 and n_1 increase, the MLE $\widehat{\theta}$ has the normal distribution $N(\theta, V)$. Combining this "likelihood" with a normal prior, $N(0, W)$, on the log relative risk, θ, gives the asymptotic Bayes factor

$$\text{ABF} = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{z^2}{2}\frac{W}{(V+W)}\right) \tag{2}$$

where $z = \widehat{\theta}/\sqrt{V}$ is the usual Wald statistic. High/low values of the asymptotic Bayes factor occur when z^2 is small/large and correspond to evidence for/against the null hypothesis. The appendix provides a rigorous derivation of the ABF.

A great advantage of the Bayes factor, (2), is that is depends on readily available summaries only, though if the sample sizes are not large there may be some loss in efficiency if $\{\widehat{\theta}, V\}$ do not summarize all the information contained in the full data concerning θ.

The crucial difference between inference based on the ABF and on the p-value calculated from the Wald statistic, which equals $2\{1 - \Phi(|z|)\}$ (where $\Phi(\cdot)$ is the distribution function of a standard normal random variable), is that ABF depends, in addition to z, on the power through the asymptotic variance V. The relationship between ABF and V is not monotonic. For fixed z (i.e. fixed p-value) and fixed prior variance W we examine the behavior of ABF as a function of the V. Figure 1 illustrates the behavior of the evidence for the alternative (1/ABF) against V for $z = 4$ and $W = 0.21^2$ (corresponding to a 95% belief that the

relative risk is less than 1.5). Recall that the significance level of the p-value is constant and that under the p-value approach no alternative is considered. In contrast, the Bayes factor compares the evidence between H_0 and H_1, assuming that one of them is true. Under the null $\widehat{\theta} \sim N(0, V)$ while under the alternative $\widehat{\theta} \sim N(0, V + W)$ and the Bayes factor is the ratio of these two quantities. For low values of V (high power) the evidence for H_1 is not strong since although the data (the z-score) are unlikely under H_0, they are unlikely under H_1 also — this behavior contrasts with the p-value under which very small departures from H_0 provide small p-values when the power is high. The evidence for H_1 increases rapidly as the power decreases, to a maximum at $V = W/(z^2 - 1)$. Beyond this point there is a decrease in the evidence for H_1 since the power is not sufficient to give strong evidence.

We briefly examine the asymptotic properties of ABF. Let $\widehat{\theta}_n$ be the MLE based on samples of size n where, for simplicity, we have assumed that $n_0 = n_1 = n$. In this case the asymptotic variance $V_n = F/n$ where F depends on the MAF and the genetic model but not on n. We have $\sqrt{n}(\widehat{\theta}_n - \theta) \to N(0, F)$ as $n \to \infty$, by the properties of MLEs and

$$\begin{aligned} \log \mathrm{ABF}_n &= \frac{1}{2} \log \left(1 + \frac{Wn}{F}\right) - \frac{n\widehat{\theta}_n^2}{2F} \times \frac{W}{\frac{F}{n} + W} \\ &= \log \left(1 + \frac{Wn}{F}\right) - \frac{\{\sqrt{n}(\widehat{\theta}_n - \theta) + \sqrt{n}\theta\}^2}{2F} \times \frac{W}{\frac{F}{n} + W}. \end{aligned}$$

When $\theta = 0$, $\log \mathrm{ABF}_n \to \infty$ as $n \to \infty$ and when $\theta \neq 0$, $\log \mathrm{ABF}_n \to -\infty$ so that ABF is consistent under both the null and the alternative and the correct model is chosen with probability 1 as the sample sizes increase.

When data from two studies (or phases) are available, with estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ and standard errors $\sqrt{V_1}$ and $\sqrt{V_2}$, the Bayes factor for the combined evidence is given by

$$\mathrm{ABF}(\widehat{\theta}_1, \widehat{\theta}_2) = \sqrt{\frac{W}{RV_1V_2}} \exp\left\{-\frac{1}{2}\left(z_1^2 RV_2 + 2z_1 z_2 R\sqrt{V_1V_2} + z_2^2 RV_1\right)\right\} \tag{3}$$

where $R = W/(V_1W + V_2W + V_1V_2)$ and z_1 and z_2 are the z-statistics arising from the two

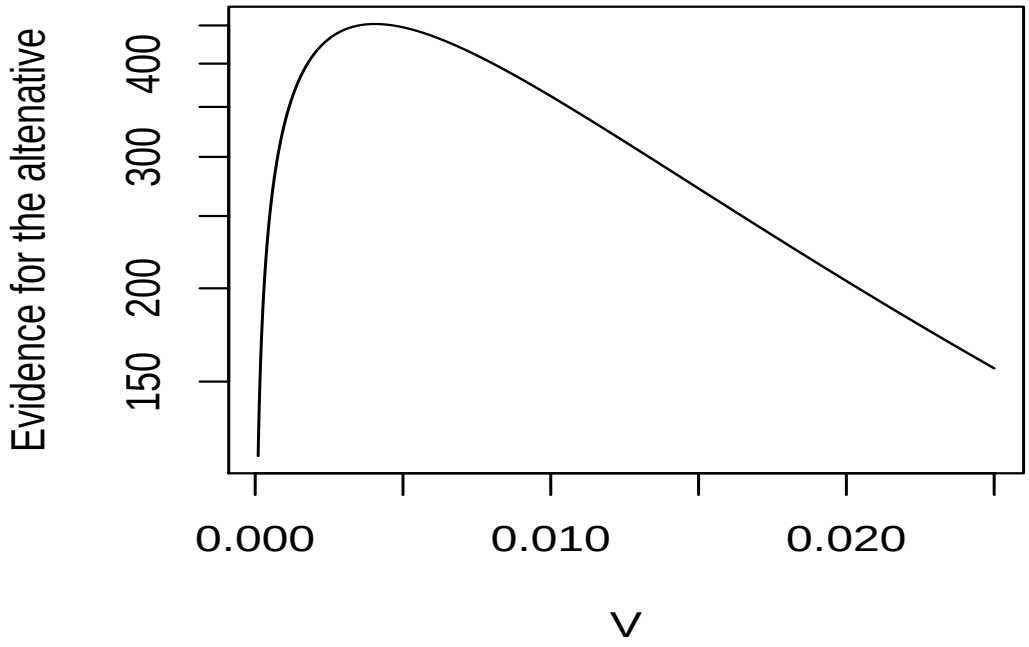


Figure 1: Evidence in favor of the alternative as a function of the asymptotic variance of the estimator, V, for fixed p-value. Small and large values of V corresponds to high and low power, respectively.

studies. Strong evidence in favor of the alternative requires large z statistics *of the same sign*. This formula can be simply extended to three or more studies.

For both single and multiple studies, and with $\text{PO} = \pi_0/(1-\pi_0)$ the prior odds on the null, the posterior probability of H_0 is given by the *Bayesian False Discovery Probability* Wakefield (2007):

$$\text{BFDP} = \frac{\text{ABF} \times \text{PO}}{1 + \text{ABF} \times \text{PO}}. \tag{4}$$

Prior Specification

The approximate Bayes factor sidesteps the need for specification of a prior on the nuisance parameters, but still requires a prior for the log relative risk of interest, θ, which is taken as normally distributed with mean 0 and variance W. The latter variance is the single specification that is needed, and we describe three particular choices.

Effect-MAF independence: The simplest choice is to take the variance, W, as independent of the MAF. The prior distribution of the relative risk, $\exp(\theta)$, is lognormal and we may specify an upper value RR_u, above which we believe that relative risks will occur with low probability. If the prior probability of a relative risk above RR_u is q we obtain $W = \{\log \text{RR}_u/\Phi^{-1}(1-q)\}^2$. For example, for a 5% chance that relative risks are above 2, $\text{RR}_u = 2, q = 0.05$ and $W = 0.42^2$.

Effect-MAF dependence: It has been argued that larger genetic effects will be associated with smaller MAFs (for discussion, see Wang et al. 2005), in which case the variance W should depend on the MAF. Selection would imply that large detrimental relative risks should not occur for common variants. A simple form that can mimic this behavior is:

$$W(M) = \delta_0 \exp(-\delta_1 \times M) \tag{5}$$

where M is the MAF. The parameters $\delta_0 > 0, \delta_1 > 0$ are chosen *a priori*. For example one may set the upper bound on the prior for the relative risk at two values of the MAF and then solve for δ_0, δ_1. Specifically, let M_{lo} and M_{hi} be the rare and non-rare MAF's at which we specify relative risks of $\mathrm{RR}_u^{lo} > \mathrm{RR}_u^{hi}$, above each of which we believe relative risks will lie with probability q. The variances at the rare and non-rare variants are:

$$W_{lo} = \{\log(\mathrm{RR}_u^{lo}/\Phi^{-1}(1-q)\}^2, \quad W_{hi} = \{\log(\mathrm{RR}_u^{hi}/\Phi^{-1}(1-q)\}^2$$

which may be solved to give:

$$\begin{aligned} \delta_1 &= \frac{\log(W_{lo}) - \log(W_{hi})}{M_{hi} - M_{lo}} \\ \delta_0 &= W_{lo} \exp(\delta_1 \times M_{lo}). \end{aligned}$$

An implicit p-value prior: In general, both ranking and significance of SNPs will differ when assessed using Bayes factors and p-values, and it is of great interest to see when the approaches can be reconciled. This unification occurs when the Bayes factor depends on the data only through z^2 in a monotonic fashion, since this is the only function of the data that determines the p-value. This is achieved if we eliminate V from ABF and occurs if we take the variance to be proportional to the asymptotic variance of the MLE:

$$W(M) = K \times V, \tag{6}$$

where K does not depend on the data (and in particular does not depend on n), to give

$$\mathrm{ABF} = \sqrt{1+K} \exp\left(-\frac{z^2}{2}\frac{K}{(1+K)}\right) \tag{7}$$

We want K to be independent of the data because we want a Bays factor that depends on the z score (and therefore the p-value) only. Such a prior was discussed with respect to the use of p-values in the context of a normal model by Cox and Hinkley (1974, p. 395–399).

Under the p-value prior, (6):

$$\log \text{ABF}_n = \frac{1}{2}\log(1+K) - \frac{\{\sqrt{n}(\widehat{\theta}_n - \theta) + \sqrt{n}\theta\}^2}{2F} \times \frac{K}{1+K}$$

which tends to $\frac{1}{2}\log(1+K)$ when $\theta = 0$ and not ∞ as is desirable. This is a consequence of the fact that for a fixed p-value threshold p_T the null will be incorrectly rejected a proportion p_T of the time under repeated sampling (which is intuitively why we need the p-value threshold to decrease with increasing sample size).

The implicit p-value prior (6) gives relatively strong belief that the effect size is small when the sample size is large and/or the MAF is not rare (since in both cases V is small). The dependence on the sample sizes n_0 and n_1 is alarming and does not make sense in the genome-wide context (in contrast to a designed experiment in which one would pick larger sample sizes when the expected effect was small, behavior that would be reflected in the prior also). For ranking, n_0 and n_1 will be constant across SNPs (give or take missing values) and so this aspect of the prior is not troubling.

The association between effect size and MAF is in the direction expected (larger effects at rarer MAFs) and is determined in a very specific manner by the depenence of V on the MAF. A convenient form for the asymptotic variance of $\widehat{\theta}$ is available from a score test (Slager and Schaid 2001), and is asymptotically equivalent to the logistic regression variance estimate:

$$V = \frac{n_0 + n_1}{n_0 n_1[(1-M)^2 x_0^2 + 2M(1-M)x_1 + M^2 x_2^2 - \{(1-M)^2 x_0 + 2M(1-M)x_1 + M^2 x_2\}^2]} \tag{8}$$

where M is the MAF, and x_0, x_1 and x_2 depend on the genetic model (examples of which were given above). The variance of the p-value prior (6) is not a simple function of the MAF, and so we graphically illustrate the shape, plotting against the comparison prior, (5). The variance of the latter can decrease rapidly with increasing MAF (with large values of δ_1), while the p-value prior exhibits a more gradual change. In Figure 2 we plot the priors for

MAFs of 0.10, 0.30 and 0.50; the p-value priors are drawn as the solid red lines while the comparison priors are the dashed lines density respectively. For the comparison prior δ_0 and δ_1 were chosen to give 95% points of 2.5 and 1.2 at MAFs of 0.05 and 0.50; K was chosen to give qualatively similar behavior though as we see, the relationship between the variance of the p-value prior and the MAF is not as strong as with the comparison prior.

The Specification of p-value Thresholds

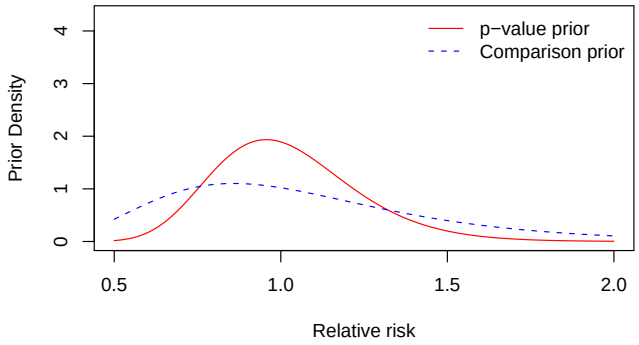
The above shows that rankings with p-values and Bays factors based on (6) will be identical, but for calibration the two approaches are more difficult to unify. As summarized elsewhere (Wakefield 2007) the Bayesian decision theory approach to calibration is to specify the costs of false non-discovery C_{FND} and false discovery C_{FD}, and then flag the SNP as "significant" if the posterior odds on H_0 drop below the ratio $\text{R} = C_{\text{FND}}/C_{\text{FD}}$. Hence an association will be called noteworthy if

$$\text{ABF} \times \text{PO} < \text{R} \tag{9}$$

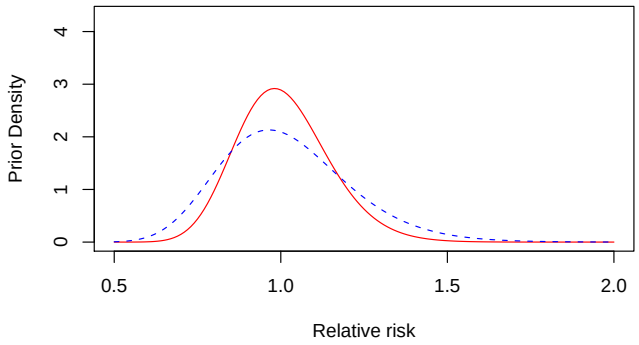
so that there are three elements to the decision problem, the ratio of the probabilities of the data under null and alternative, ABF, the prior odds on H_0, PO, and the ratio of costs, R.

For simplicity assume that the case and control sample sizes are equal, $n = n_0 = n_1$. Recall that for a MAF of M the implicit p-value prior is $W(M) = K \times F(M)/n$ where $F(M)$ is given in (8) and does not depend on sample size. To rectify the undesirable dependence of the prior on sample size, while retaining the effect-MAF relationship implied by the p-value, one can take $W(M) = K^\star \times F(M)$ to give

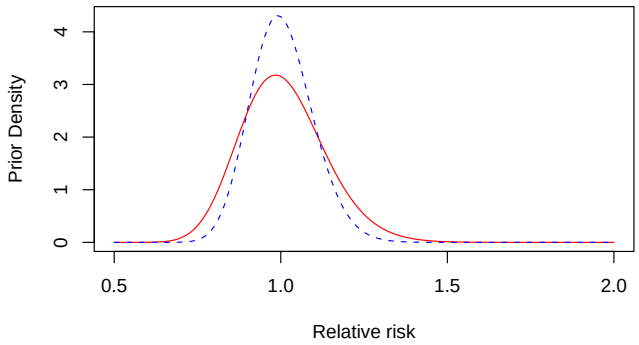
$$\begin{aligned} \text{ABF} &= \sqrt{1+nK^\star}\exp\left[-\frac{z^2}{2}\frac{nK^\star}{(1+nK^\star)}\right] &(10)\\ &= \sqrt{1+nK^\star}\exp\left[-\frac{1}{2}\Phi^{-1}\left(1-\frac{p}{2}\right)^2\frac{nK^\star}{1+nK^\star}\right] &(11)\end{aligned}$$



(a) MAF = 0.10



(b) MAF = 0.30



(c) MAF = 0.50

Figure 2: Implicit p-value prior (solid red lines) with prior variance on the log relative risk $W(M) = K \times V(M)$ (where M is the MAF), as compared to the prior $W(M) = \delta_0 \exp(-\delta \times M)$ (dashed blue lines); δ_0 and δ_1 are chosen so that at a MAF of 0.05 the 95% point of the prior is 2.5, and at 0.50 it is 1.2.

In contrast to the use of the prior $W = K \times V$ this ABF depends upon n, as well as upon the p-value. Hence two ABFs calculated from (11) with the same p-value, but with different sample sizes will provide different evidence for/against the null. This consequence is not pertinent to ranking, since rankings are based on comparisons with fixed sample sizes.

Now suppose one wishes to use a formal decision theory approach to picking a p-value threshold for calling an association noteworthy. Substituting (10) into (9) and rearranging, one finds that the z^2 threshold is

$$z_T^2 = 2\frac{(1+K^\star n)}{K^\star n}\log\left(\frac{\mathrm{PO}}{\mathrm{R}}\sqrt{1+K^\star n}\right) \tag{12}$$

We stress that we are assuming that both the prior odds, PO, and the ratio of costs, R, do not depend on the sample size or the MAF — further discussion of these assumptions is postponed until the discussion. The p-value threshold that should be used is therefore

$$p_T = 2\left[1-\Phi^{-1}\left(z_T\right)\right] \tag{13}$$

From (12) we see that the z^2 threshold increases (so that the p-value threshold decreases) as the prior odds of no association increases or as the ratio of costs of false discovery to false non-discovery increase, both as expected. It is difficult to make general statements about the dependence of the threshold on sample size since the formula is a complex function of n, as we saw in Figure 1.

To evaluate particular values of the threshold one must pick a value of $K^\star$ to calibrate the prior. If we take $K^\star = 1$ then this prior is equivalent to the unit-information prior (Kass and Wasserman 1995) which has been suggested as a "reference prior" for Bayes factors; this prior is not appealing in the GWAS setting in which substantive prior opinion on the size of possible effects exists. In what follows, for illustration, we choose a dominant model, evaluate the asymptotic variance at the null, and choose $K^\star$ so that the 95% points of the

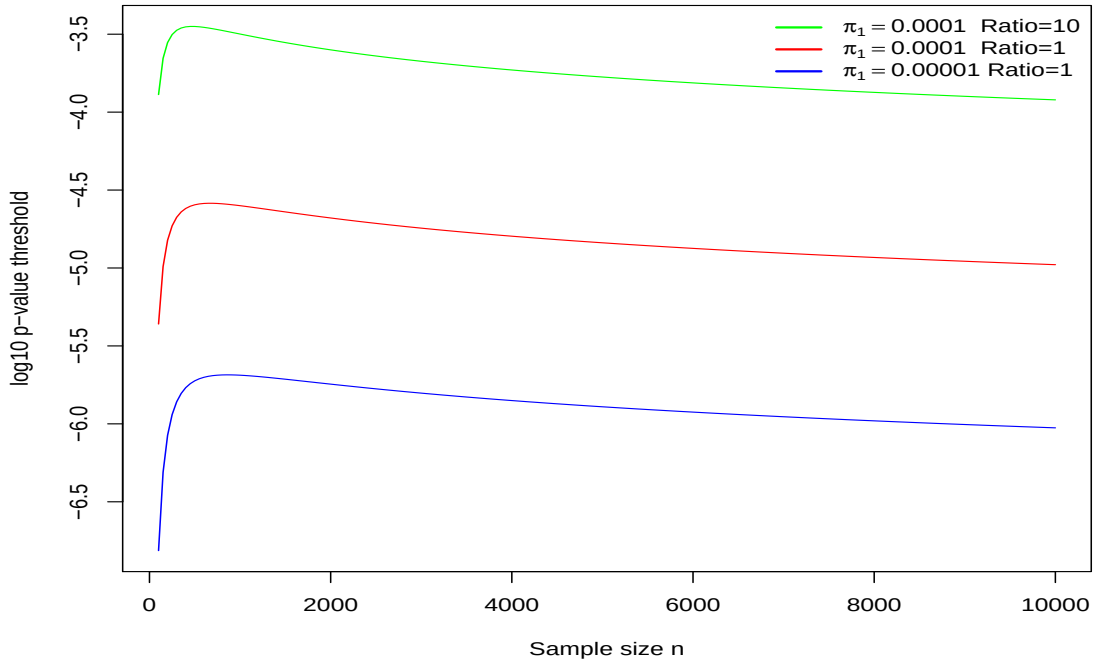


Figure 3: p-value threshold at which to call an association noteworthy versus cases and control sample sizes n, and for different values of the prior on the alternative π_1, and ratios of costs of false non-discovery to false discovery.

prior on the relative risk are 4.4 and 2.3 at MAFs of 0.05 and 0.50, respectively. Figure 3 shows the p-value threshold (on a log 10 scale) as a function of n, and for particular values of π_1, the prior on the alternative, and the ratio of costs R (neither of which, recall are assumed to depend on sample size or MAF). The curves are not horizontal which shows that it is important to make thresholds depend on sample size if a p-value approach is taken. The non-monotonic behavior is due to the complex relationship between the ABF and power (Figure 1). Requiring a smaller p-value for larger sample size (larger power), has been frequently advocated (Wacholder et al. 2004; Consortium 2007). From a decision theory perspective taking a common threshold across all sample sizes is only consistent with one or more of the prior variance on the effect size, the prior odds on an association, or the ratio of costs changing with n, and in a very stylized way. We stress that we are not advocating the use of the rule (13) in practice, as we would rather use a Bayes factor with the prior on the effect size reflecting our actual beliefs.

Multiple Testing

An important observation is that the threshold rule, (11), does not depend on the number of SNP associations to be considered, m, though interpretation, and in particular the expected number of false discoveries, will depend on this number. The small p-value thresholds shown in Figure 3 occur in large part due to the low *a priori* probability of a non-null hypothesis. Hence the threshold rule we have derived is in direct contrast to the Bonferroni threshold which is based on the number of tests considered, with the sample size being irrelevant. For example, Reisch and Merikang (1996) argue that when testing 10^6 associations a Bonferroni correction would suggest a level of 5×10^{-8}, while Dahlman et. al. (2002) suggest p-values in the range 10^{-7} to 10^{-8} if 250K–500K tests are carried out. In spite of the limitations of

the Bonferroni correction, including the relevance of controlling the probability of a single type I error, and questions over the relevant number of tests over which to control (Colhoun, McKeigue, and Davey-Smith 2003; Balding 2006) it remains the most common method of adjustment in GWASs (DeWan et al. 2006; Frayling et al. 2007).

If one believes that all of the nulls might be true then the Bonferrroni correction is more meaningful, and a similar adjustment can be obtained via a Bayesian approach (Westfall et al. 1995). Specifically suppose that when one specifies the prior over the set of m null associations one wishes to control the prior probability of all the nulls being true. The simplest way of achieving this is to take the prior on each null as $\Pi_0^{1/m}$ to give a prior probability of all nulls being true as Π_0. In contrast, if one specifies independent priors on each null to be π_0 the induced prior on all nulls being true is π_0^m. Under the latter, if $\pi_0 = 1 - 1/100000$ (so that we believe that in 1 in 100K SNPs are non-null), the prior on all 500K SNPs being null is 0.0067, i.e. very unlikely. If this prior truly reflects ones beliefs then controlling the family wise error rate (as is achieved by the Bonferonni correction) is an unappealing criterion.

We illustrate we use the prior $\Pi_0^{1/m}$ along with the asymptotic Bayes factor (10). A Bonferonni threshold is obtained by taking the p-value threshold corresponding to $\pi_0 = \Pi_0$, and then dividing this threshold by m. This may be compared with the threshold arising from (10) with the prior $\pi_0 = \Pi^{1/m}$, which we refer to as the power prior. Table 1 gives thresholds based on $\Pi_0 = 0.9999$, calculated for $n = 1000$. For a single test the thresholds under this prior is 6.2×10^{-8}. We see that there is a close correspondence between the Bonferonni and power prior thresholds. The message here is that the Bonferroni correction has a Bayesian justification, but only for a very extreme prior that will often be inapprorpiate in a GWAS.

Table 1: p-value thesholds (to the $\log_{10}$) for $\Pi_0 = 0.9999$. For the Bonferroni threshold we take the p-value threshold corresponding to $m = 1$, and divide by m. For the power prior we take $\pi_0 = \Pi_0^{1/m}$.

	Number of Tests m				
	1	1,000	100,000	317,000	500,000
Bonferroni Threshold	-6.17	-9.17	-11.17	-11.67	-11.87
Power Prior Threshold	-6.17	-9.27	-11.32	-11.83	-12.03

Discussion

We have considered the use of Bayes factors in genome-wide association studies and have illustrated that the use of the p-value corresponds to a particular prior on the relative risk parameter. This prior depends on the MAF in a qualatively reasonable way, with stronger effects anticipated at lower MAFs. The relationship between effect size and MAF is not strong, however (as illustrated in Figure 2), and lists of top-ranked SNPs from p-value and a Bayes factor approach with prior independence between effect size and MAF will often be similar, with differences only for SNPs with very low MAFs.

For final inference the use of the p-value is problematic, however, since its interpretation depends on sample size. We have shown, using a decision theory approach that if one assumes that neither the ratio of costs of type II to type I errors nor the prior odds do not depend on either the sample size or the MAF, and that the effect size does not depend on sample size, then it is not optimal to take a single p-value threshold for all sample sizes. Rather the threshold should decrease as a function of sample size, an approach that is already informally taken, based on experience and examination of interval estimates. A fixed threshold leads to a procedure that is inconsistent, in that the correct model is not chosen with probability

1 as the sample size increases.

With a formal Bayesian approach to testing, one may allow the ratio of costs and the prior odds to depend on sample size and MAF also. As data collection proceeds the cost of false discovery will increase relative to the cost of false non-discovery (early on we would like a long list), and such behavior can be formalized with a Bayesian decision theory approach. We may also wish the costs to depend on the MAF, so that there is a higher cost associated with a more common variant.

Andrews (1994), in a wide-ranging article, showed the relationship between Bayes factors and Wald, likelihood ratio and score statistics, under more general priors; the normal prior discussed above is a special case which is practically convenient. See also Efron and Gous (2001), and Johnson (2005, 2007); the latter derives properties of the Bayes factors based on test statistics.

For calculation, the Bayes factor described here requires just a point estimate/standard error, or a confidence interval. `R` code for one- and two-stage designs is available at: `http://faculty.washington.edu/jonno/cv.html`

The asymptotic Bayes factor described here can be used in any situation in which an estimate and standard error are available. A number of authors have considered regression using imputed unmeasured SNPs (Marchini et al. 2007; Servin and Stephens 2007). When data on such SNPs are analyzed the uncertainty in genotype must be acknowledged; and there are now a number of packages that allow valid inference to be made (Sinnwell and Schaid 2005). The implemented methods use weighted logistic regression model, with the weights given by the posterior probabilities of the genotypes, and a generalized estimating equation (with the clusters being the individuals), to account for the repeated observations on each individual (French et al. 2006). The use of estimates and standard errors from such approaches results

in a Bayes factor that is adjusted for measurement error in the genotype.

Recently two-phase sampling has been suggested as a method by which efficiency may be gained in a genetic epidemiology study (Chatterjee and Chen 2007). Specifically, at phase 1 inexpensive covariates may be measured on all individuals, with genetic and exposure information only gathered on an informative set of individuals at phase 2. The selection mechanism is carefully chosen to maximize information, but depends on the outcome and so must be accounted for in the estimation scheme. In other situations, survival data may be the endpoint of interest so that, for example, the Cox model may be the appropriate analysis tool. In both of these case a valid estimate and standard error is produced by the relevant software, and these can be used to evaluate the Bayes factor described here, so long as the sample size is large.

Appendix: Derivation of the Asymptotic Bayes Factor

In a case-control study we have binary disease indicators Y_i on $i = 1, ..., n_1$ cases, and $i = n_1 + 1, ..., n_1 + n_0$ controls. These data follow a binomial distribution with index 1 and probability p_i, the risk for individual i. We assume the logistic regression model:

$$\frac{p_i}{1 - p_i} = \exp(\boldsymbol{\alpha}\boldsymbol{z}_i + \theta x_i) \tag{14}$$

where $\boldsymbol{z}_i$ is a $1 \times p$ vector of confounders (which includes the intercept) with associated parameters $\boldsymbol{\alpha}$, and x_i depends on the genetic model and is a function of the number of mutant alleles possessed by individual i. The Bayes factor is given by $\Pr(\boldsymbol{y}|H_0)/\Pr(\boldsymbol{y}|H_1)$

where the numerator and denominator are integrals given, respectively, by

$$\begin{aligned}\Pr(\boldsymbol{y}|H_0) &= \int \Pr(\boldsymbol{y}|\boldsymbol{\alpha},\theta=0)\pi(\boldsymbol{\alpha})d\boldsymbol{\alpha} \\ \Pr(\boldsymbol{y}|H_1) &= \int \Pr(\boldsymbol{y}|\boldsymbol{\alpha},\theta)\pi(\boldsymbol{\alpha},\theta)d\boldsymbol{\alpha}d\theta\end{aligned}$$

where $\pi(\boldsymbol{\alpha},\theta)$ is the prior over all parameters, and $\pi(\boldsymbol{\alpha})$ is the prior over the nuisance parameters only. These integrals are analytically intractable for the binomial likelihood, and the specification of multivariate priors is cumbersome. We follow a different approach and assume we are in an asymptotic situation in which n_0 and n_1 are "large", a situation that is almost always satisfied in genome-wide association studies. In this case inference for the logistic model may be carried out on the basis of the asymptotic distribution:

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{\theta} \end{bmatrix} \sim N_{p+1}\left(\begin{bmatrix} \boldsymbol{\alpha} \\ \theta \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{00} & \boldsymbol{I}_{01} \\ \boldsymbol{I}_{01}^{\mathrm{T}} & I_{11} \end{bmatrix}^{-1} \right) \tag{15}$$

where $\boldsymbol{I}_{00}$ is the $p \times p$ matrix expected information concerning $\boldsymbol{\alpha}$, I_{11} is the expected information concerning θ, and $\boldsymbol{I}_{01}$ is the $p \times 1$ vector of cross terms. In (Wakefield 2007) it was assumed that $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\theta}$ were independent. Here we relax this assumption, reparameterize the model and consider $(\boldsymbol{\alpha},\theta) \rightarrow (\boldsymbol{\beta},\theta)$ where

$$\boldsymbol{\beta} = \boldsymbol{\alpha} + \frac{\boldsymbol{I}_{01}}{I_{00}}\theta$$

which yields

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\theta} \end{bmatrix} \sim N_{p+1}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \theta \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{00}^{\star} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & I_{11} \end{bmatrix}^{-1} \right) \tag{16}$$

where $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\alpha}} + (\boldsymbol{I}_{01}/I_{00}) \times \widehat{\theta}$ and $\mathbf{0}$ is a $p \times 1$ vector of zeros. Hence, asymptotically, $p(\widehat{\boldsymbol{\beta}},\widehat{\theta}|\boldsymbol{\beta},\theta) = p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta}) \times p(\widehat{\theta}|\theta)$. We assume independent priors on $\boldsymbol{\beta}$ and θ, $\pi(\boldsymbol{\beta},\theta) = \pi(\boldsymbol{\beta})\pi(\theta)$

and calculate the Bayes factor working with "data" $\{\widehat{\boldsymbol{\beta}}, \widehat{\theta}\}$. Under H_0:

$$\begin{aligned} p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|H_0) &= \int p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|\boldsymbol{\beta}, \theta = 0)\pi(\boldsymbol{\beta})d\boldsymbol{\beta} \\ &= \int p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} \times p(\widehat{\theta}|\theta = 0) \end{aligned}$$

and under H_1:

$$\begin{aligned} p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|H_1) &= \int\int p(\widehat{\boldsymbol{\beta}}, \widehat{\theta}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}, \theta)d\boldsymbol{\beta}d\theta = \int\int p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta})p(\widehat{\theta}|\theta)\pi(\boldsymbol{\beta})\pi(\theta)d\boldsymbol{\beta}d\theta \\ &= \int p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} \int p(\widehat{\theta}|\theta)\pi(\theta)d\theta \end{aligned}$$

Hence the Bayes factor based on $(\widehat{\boldsymbol{\beta}}, \widehat{\theta})$ is given by:

$$\text{ABF} = \frac{p(\widehat{\theta}|\theta = 0)}{\int p(\widehat{\theta}|\theta)\pi(\theta)d\theta} \tag{17}$$

The reparamaterization trick works because of the assumption of independent priors on $\boldsymbol{\beta}$ and θ, which does not imply independent priors on $\boldsymbol{\alpha}$ and θ. We emphasize that we need never specify the prior on $\boldsymbol{\beta}$, because terms involving $\boldsymbol{\beta}$ cancel in the Bayes factor calculation. Under the prior $\theta \sim N(0, W)$ the Bayes factor (17) becomes

$$\text{ABF} = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{z^2}{2}\frac{W}{(V+W)}\right).$$

where $V = I_{11}^{-1}$. The reparameterization described here is that which is used when the linear model:

$$Y_i = \alpha + x_i\theta + \epsilon_i$$

is written as

$$Y_i = \beta + (x_i - \overline{x})\theta + \epsilon_i$$

which, of course, yields uncorrelated least squares estimators $\widehat{\beta}, \widehat{\theta}$. The approach employed here is similar to the "null orthogonality" reparameterization of Kass and Vaidyanathan (1992).

Acknowledgments

This work was partially supported by grant 1 U01–HG004446–01 from the National Institutes of Health. The author would like to thank two anonymous referees for constructive comments that aided in clarification of the material.

References

Andrews (1994). The large-scale correspondence between. *Econometrica 62*, 1207–1232.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics 7*, 781–791.

Chatterjee, N. and Chen, Y.H. (2007). Maximum likelihood inference on a mixed conditionally and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. *Journal of the Royal Statistical Society, Series B 69*, 123–142.

Colhoun, H.M., McKeigue, P.M., and Davey-Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. *The Lancet 361*, 865–872.

Consortium, TheWellcomeTrustCaseControl (2007). Genome-wide association study between 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*, 661–678.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Dahlman, I., Eaves, I.A., Kosoy, R., Morrison, V.A., Heward, J., Gough, S.C.L., Allahabadia, A., Franklyn, J.A., Tuomilehto, J., Tuomilehto-Wolf, E., Cucca, F., Guja, C., Ionescur-Tirgoviste, C., Stevens, H., Carr, P., Nutland, S., McKinney, P., Shield,

J.P., Wang, W., Cordell, H.J., Walker, N., Todd, J.A., and Concannon, P. (2002). Parameters for reliable results in genetic association studies in common disease. *Nature Genetics 30*, 149–150.

DeWan, A., Liu, M., Hartman, S., Zhang, S.ShaoMon, Liu, D.T., Zhao, C., Tam, P.O.S, Chan, W.M., Lam, D.S.C, Snyder, M., Barnstable, C., Pang, C.P., and Hoh, J. (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science 314*, 989–992.

Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D.P., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., and Luben, R. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature 447*, 1–9.

Efron, B. and Gous, A. (2001). Scales of evidence for model selection: Fisher versus Bayes (with discussion). In P. Lahiri (Ed.), *Model Selection*, pp. 208–256. Institute of Mathematical Statistics Lecture Notes, Monograph Series.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., and et al., C.M.Lindgren (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science 316*, 889–894.

French, B., Lumley, T., Monks, S.A., Rice, K.M., Hindorff, L.A., Reiner, A.P., and Psaty, B.M. (2006). Simple estimates if haplotype relative risks in case-control data. *Genetic Epidemiology 30*, 485–494.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: the P value fallacy. *Annals of Internal Medicine 130*, 995–1004.

Hirschhorn, J.N. and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics 6*, 95–108.

Johnson, V.E. (2005). Bayes factors based on test statistics. *Journal of the Royal Statistical*

Society, Series B 67, 689–701.

Johnson, V.E. (2007). Properties of Bayes factors based on test statistics. *Scandanavian Journal of Statistics*. Published on-line, October 31st, 2007.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kass, R.E. and Vaidyanathan, S.K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B 54*, 129–144.

Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics 39*, 906–913.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science 273*, 1516–1517.

Sasieni, P.D. (1997). From genotypes to genes: doubling the sample size. *Biometrics 53*, 1253–1261.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantative traits. *Public Library of Science Genetics 3*, 1296–1308.

Sinnwell, J.P. and Schaid, D.J. (2005). haplo.stats: statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. *American Journal of Human Genetics 70*, 425–434.

Sladek, R., Rocheleau, G., Ring, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., and et al, B.Balkau (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature 445*, 881–885.

Slager, S.L. and Schaid, D.J. (2001). Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Human Heredity 52*, 149–153.

Sterne, J.A.C. and Smith, G.Davey (2001). Sifting the evidence – what's wrong with significance tests? *British Medical Journal 322*, 226–231.

Wacholder, S., Chanock, S., Garcia-Closas, M., El-ghormli, L., and Rothman, N. (2004). Assessing the probability that a postitive report is false: an approach for molecular epidmiology studies. *Journal of the National Cancer Institute 96*, 434–442.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics 81*, 208–227.

Wakefield, J.C. (2008). Reporting and Interpretation in Genome-Wide Association Studies. *International Journal of Epidemiology*. Published on-line February 11th, 2008.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics 6*, 109–118.

Westfall, P.H., Johnson, W.O., and Utts, J.M. (1995). A Bayesian perspective on the Bonferroni adjustment. *Biometrika 84*, 419–427.

Figure Legends

Figure 1: Evidence in favor of the alternative as a function of the asymptotic variance of the estimator, V, for fixed p-value. Small and large values of V corresponds to high and low power, respectively.

Figure 2: Implicit p-value prior (solid red lines) with prior variance on the log relative risk $W(M) = K \times V(M)$ (where M is the MAF), as compared to the prior $W(M) = \delta_0 \exp(-\delta \times M)$ (dashed blue lines); δ_0 and δ_1 are chosen so that at a MAF of 0.05 the 95% point of the prior is 2.5, and at 0.50 it is 1.2.

Figure 3: p-value threshold at which to call an association noteworthy versus cases and control sample sizes n, and for different values of the prior on the alternative π_1, and ratios of costs of false non-discovery to false discovery.