INVITED COMMENTARY

International Journal of Sports Physiology and Performance, 2008, 3, 547-557

Inference About Magnitudes of Effects

Richard J. Barker and Matthew R. Schofield

In a recent commentary on statistical inference, Batterham and Hopkins[1] advocated an approach to statistical inference centered on expressions of uncertainty in parameters. After criticizing an approach to statistical inference driven by null hypothesis testing, they proposed a method of "magnitude-based" inference and then claimed that this approach is essentially Bayesian but with no prior assumption about the true value of the parameter. In this commentary, after we address the issues raised by Batterham and Hopkins, we show that their method is "approximately" Bayesian and rather than assuming no prior information their approach has a very specific, but hidden, joint prior on parameters. To correctly adopt the type of inference advocated by Batterham and Hopkins, sport scientists need to use fully Bayesian methods of analysis.

Keywords: statistics, Bayesian, inference, magnitude, effect

In a recent commentary on statistical inference, Batterham and Hopkins[1,2] claim:

1. null hypothesis testing is the "almost universal approach to inferential statistics" and further that ". . . hypothesis testing is illogical, because the null hypothesis of no relationship or no difference is always false – there are no truly zero effects in nature."
2. ". . . the strict definition of the confidence interval is hotly debated" and that "it is the likely range of the true, real, or population value of the statistic."

Batterham and Hopkins[1] then go on to suggest a form of inference in which the parameter space is partitioned into regions corresponding to various qualitative aspects of the parameter. A probability is assigned to each of these regions, with that probability determined by the sampling distribution of the estimator. Batterham and Hopkins[1] claim that this approach is essentially Bayesian[3] but with no prior assumption about the true value of the parameter, a claim that has gained some acceptance.[4]

Here we consider these claims in detail and go on to show that the inference proposed by Batterham and Hopkins is an approximately Bayesian procedure with a very specific prior on parameters.

The authors are with the Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand.

The dominant form of statistical inference over the past 100 years is called *frequentist*, in which statistical inference is based on procedures that are derived to have good long-run frequency properties. For example, a 95% confidence interval has the property that in repeated application it will include the parameter of interest in 95% of cases. The main alternative to frequentist inference is Bayesian inference, in which prior knowledge about an unknown quantity is combined with information from the data to construct what is called a *posterior distribution*. This posterior distribution summarizes our new state of knowledge once we have collected our data. Here, we argue that Bayesian inferences offer many advantages over frequentist methods of statistical analysis, but also that users should be both aware and up-front about prior choices that they have made.

Null-Hypothesis Testing

Despite its widespread use, few statisticians would agree that null-hypothesis testing is a "universal approach to inferential statistics." In addition, few would defend the widespread use of null-hypothesis testing,[5] but nor would many condemn its use outright.[6] Null-hypothesis testing and *P*-values do have some use. For example, in drug abuse testing, the null hypothesis formalizes the presumption of innocence. Null-hypothesis tests are also useful in assessing model fit, where they are used to check for gross inconsistencies between the data and the model.

A common objection to null-hypothesis testing is that null hypotheses are never true: "there are no truly zero effects in nature."[1] However, it is important to highlight the distinction between models and reality. The null hypothesis exists in the context of a model, which is an abstraction of the true process(es) that produced our data. As scientists we are not interested in "truly zero effects" but in "effectively" zero effects. A hypothesis that an effect is effectively zero may be perfectly reasonable. For example, it is unlikely that the color of one's underwear has much influence on the time to run 100 m, although with enough investigation we might find counter-examples. Nevertheless as a working hypothesis, it seems reasonable to suppose that underwear color has effectively no influence. Thus, null hypotheses need not be controversial.

Interval Estimation

Formally, a $100(1 - \alpha)\%$ confidence set is a random set of values (usually an interval) with the property that under repeated use it includes the unknown parameter with probability $1 - \alpha$. Among statisticians, there is no debate about the meaning of a confidence interval although we accept that the strict definition may not be understood by many users.

It must be understood that our "confidence" is in the procedure used to construct the interval. It is not a probabilistic statement about likely values of the parameter. The parameter is regarded as fixed, but unknown. When we report a 95% confidence interval, we are claiming (implicitly) that if we constructed a large

number of such intervals in this way, we would expect 95% of them to include the parameter and 5% to exclude it. Thus, for any such interval we can be confident that it contains the parameter.

A common misinterpretation is that a confidence interval is a probability statement about the parameter rather than the interval. The theoretical basis of most of the statistics taught to students is called *frequentist*, as it is based on the long-run frequency properties of estimators and tests. In frequentist theory, probability statements can only apply to random variables or events. Because a parameter is not a random variable but a fixed and unknown constant, it makes no sense to refer to the probability of fixed quantities in frequentist theory. The distinction between a probability statement about the random interval and a probability statement about the parameter is subtle. It is hardly surprising, then, that this misunderstanding of confidence intervals is almost universal among users.[7]

As noted by Batterham and Hopkins,[1] there is an alternative to frequentist theory, known as *Bayesian inference,* that does admit probability statements about parameters. In Bayesian theory, our uncertainty about a parameter is expressed using probability distributions. The uncertainty we have about a parameter before we have collected data is expressed by a prior distribution. Once we have collected data, we update our knowledge about the parameter by computing a new distribution for the parameter, one that is conditional on the data. This is known as the *posterior distribution*. We can use the posterior distribution to legitimately make probability statements about parameters.

To illustrate, suppose that following a marathon 3 out of 14 female athletes show signs of elevated hormone levels indicating stress. What can we say about the proportion of similar athletes in the population that this sample represents? Assuming that athletes respond independently, we can model the number of athletes showing signs of stress as a binomial random variable with parameters n = 14 and p. The standard (frequentist) solution would be to provide an estimate and confidence interval for p, say, $\hat{p} = 0.214$ with an approximate 95% CI of (0, 0.429). This solution is approximate because a small-sample confidence interval with exactly 95% coverage does not exist. In fact, this confidence interval, calculated using $\hat{p} \pm \sqrt{(\hat{p}(1-\hat{p}))/n}$, has coverage nowhere near 95% for most values of p in small samples. For further information, see Agresti and Coull.[8]

To carry out Bayesian inference, we need to specify a prior on p, the population proportion. This prior summarizes what we know about p before the experiment. Wanting to let the data dominate the prior, we use a uniform $U(0, 1)$ prior. The resulting posterior density is a beta $Be(4, 11)$ density, which is plotted in Figure 1. From this we can compute an exact credible interval for p of (0.062, 0.456). Importantly, Bayesian inference allows us to make a probability statement about the parameter p conditional on our one data set and the model, including the prior: the probability that p lies between 0.062 and 0.456 is exactly 95%. In contrast, the frequentist interval provides a statement of confidence in the procedure: we can be confident that the interval contains p because of the repeated sampling properties of our interval estimator. We cannot say that the probability that p lies between 0 and 0.42 is 95%.

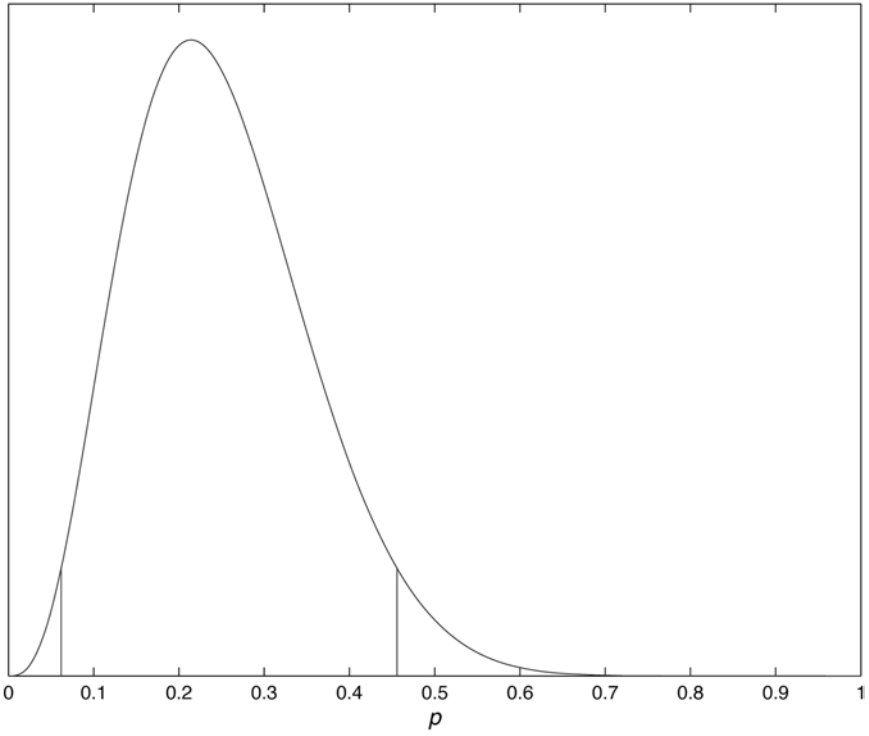

Figure 1 — Posterior density for the population proportion p, where $y = 3$ positive responses are observed in a sample of $n = 14$ individuals assuming a binomial $B(n, p)$ model for y and a uniform $U(0, 1)$ prior for p. The two vertical lines mark the limits of the 95% credible interval for p.

Magnitude-Based Inference

In their Figure 2, reproduced in Figure 2 here, Batterham and Hopkins[1] refer to inferences that can be drawn when the parameter space can be partitioned into regions corresponding to "harmful," "trivial," or "beneficial" effects. Note that the conclusions we have included in Figure 2 are different from those reported by Batterham and Hopkins[1] but are those that are justified by the confidence intervals. As a way of subject-matter interpretation of a confidence interval, this approach is both reasonable and valid.

However, Batterham and Hopkins[1] dismiss the interpretations in Figure 2 as "crude" and go on to extend this idea by providing probabilities that the parameter falls in the three regions defined by the partitioning. These probabilities are calculated from the assumed distribution of the estimator for the parameter. This step involves a fundamental shift from frequentist to Bayesian theory as it is only Bayesian inference that allows us to make probability statements about unknown

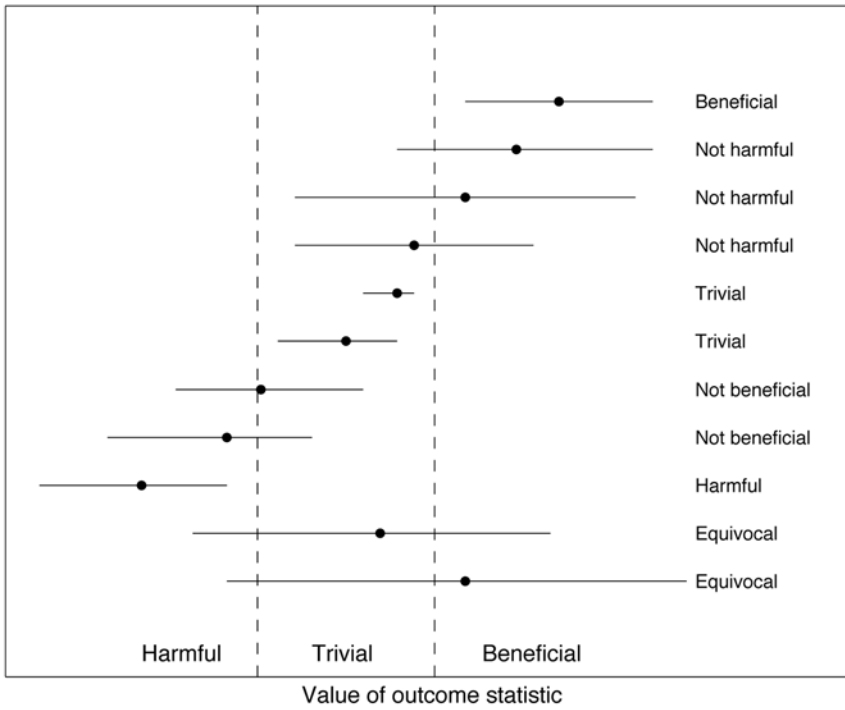

Figure 2 — Inferences that can be drawn from different confidence intervals when referenced against a partitioning of the parameter space into regions corresponding to "harmful," "trivial," and "beneficial" effects. Adapted from Batterham and Hopkins.[1]

values of parameters. Under frequentist theory, the "crude" interpretations of the confidence intervals are all that can be justified. To make the probabilistic interpretations suggested by Batterham and Hopkins,[1] we have no choice but to adopt a fully Bayesian approach complete with specification of prior probabilities.

Bayesian Inference

The key idea behind Bayesian inference is that all uncertainty should be represented by probability distributions. In Bayesian thinking, probability distributions are applied to anything about which we are uncertain, including parameters. Bayesians use probability distributions as a model for their uncertainty about parameters even though they regard parameters as fixed quantities at the time of the experiment.

The need for prior distributions in Bayesian inference has led to the common criticism that Bayesian inference is subjective because different choices of priors will lead to different posterior distributions. A common solution is to propose reference priors (meaning default priors) that represent vague knowledge about the parameters. These issues are explored further below.

Batterham and Hopkins[1] suggest that the approach that they advocate is essentially Bayesian and one that makes no prior assumption about the true value of parameters. This is untrue and in fact can never be true. Effectively, what they are proposing is a specific type of reference prior.

Consider a simple example where we have data $y_1, \ldots, y_n$ that we have decided can be modeled using a normal $N(\mu, \sigma^2)$ distribution with mean μ and variance σ^2, both unknown. For inference about μ, Batterham and Hopkins provide a spreadsheet that is available from http://newstats.org/generalize.html. They use this for carrying out the analysis in which they treat the theoretical probability distribution of the estimator as a posterior distribution. In this spreadsheet, the probabilities associated with a partitioning of the parameter space are obtained using a *t*-distribution for μ that is centered on the estimator $\bar{y}$ with degrees of freedom $\nu = n - 1$ and scale parameter $s / \sqrt{n}$. To obtain this distribution as a posterior for μ, it can be shown[2] that we require a prior for μ that is uniform on $(-\infty, \infty)$ and an independent one for σ^2 that is proportional to $1/\sigma^2$. The first problem is that these priors are improper, meaning that they are not density functions (eg, the area under the curve is infinite, not 1.0). The second and more important problem is that the prior for σ^2 is far from flat (Figure 3) and puts greatest prior mass on small values for σ^2.

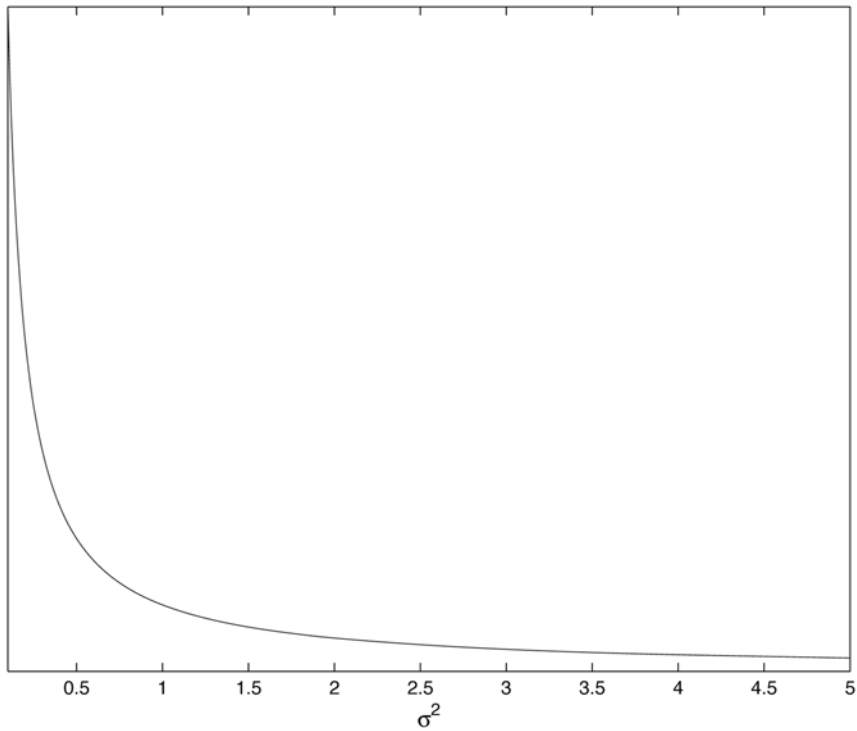

Figure 3 — Plot of the function $1/\sigma^2$ vs σ^2. This is the implicit prior on σ^2 used in the method of Batterham and Hopkins[1] when making inference about the parameter μ for data that are normally distributed with mean μ and variance σ^2 when both parameters are unknown.

Far from requiring no prior, the method proposed by Batterham and Hopkins[1] has an implicit prior on the parameters μ and σ^2. In fact, any procedure in which the posterior density is the sampling distribution of the estimator corresponds to an implicit prior on parameters. Such priors function as unspecified assumptions and may not be the ones that the user wants to make.

Problems With "Flat" Priors

One difficulty with the concept of flat prior distributions is that flat on one scale is not flat on another. Suppose we specify a "flat" prior for the standard deviation σ by using a uniform $U(0, \theta)$ distribution where θ specifies an upper limit. The corresponding induced prior for σ^2 is far from flat and puts greatest prior mass on small values for σ^2 (Figure 4).

If one plans to use flat priors, it is important to consider carefully the scale on which the prior is to be flat. For single parameter problems, an alternative to the uniform distribution prior is to use a prior known as the *Jeffreys prior*, which is designed to express the same prior belief no matter what scale is used. For multiparameter problems, specification of vague priors is more difficult.[9] As a general

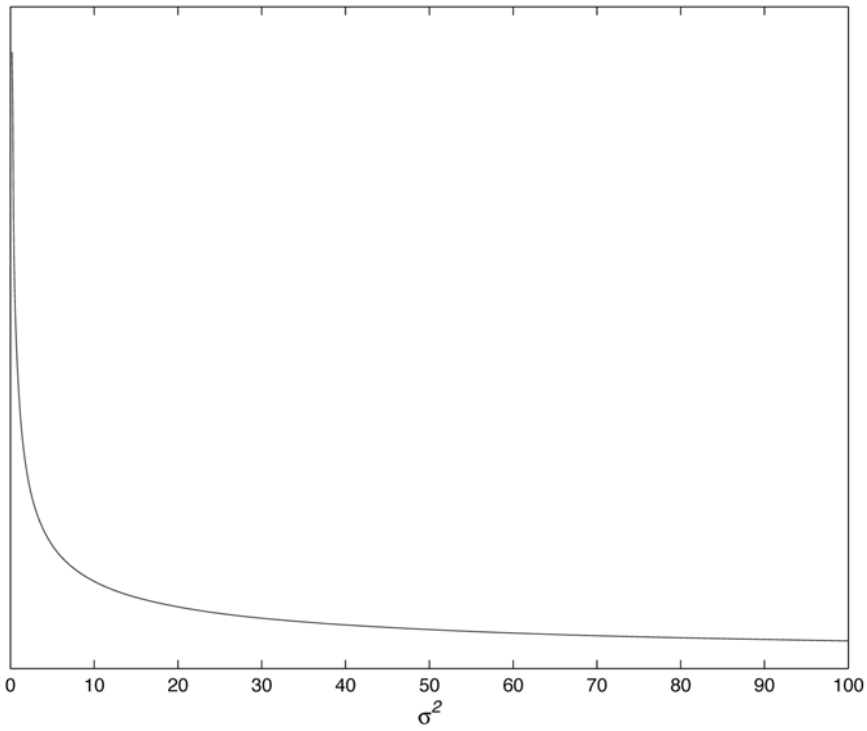

Figure 4 — The prior distribution obtained for σ^2 if a uniform $U(0, 10)$ prior distribution is used for σ, which is then transformed to σ^2.

rule, the choice of prior is less important as the amount of data increases. When the choice of prior is uncertain, we advocate exploration of the sensitivity of the posterior density to this choice as shown in the following example. If posterior inference is sensitive to different priors, we either have to accept an element of equivocation in inference or, preferably, collect more data. Simply put, when we have few observations, preconceptions matter. This is true regardless of whether we adopt a frequentist or a Bayesian standpoint in our inference.

In the example above on hormone levels in marathon runners, the uniform $U(0, 1)$ prior, which is also a beta $Be(1, 1)$ distribution, represents one commonly used reference prior. The Jeffreys prior for this problem is a beta $Be(1/2, 1/2)$ prior. The prior information that a $Be(a, b)$ prior contributes can be summarized as though we have already observed a successes in $a + b$ trials, before we observed y. Thus, we could also contemplate a beta prior with $a = b = 0$. However, we recommend against its use because when $y = 0$ or $y = n$, the resulting posterior distribution is improper. An alternative is a logit-normal distribution, in which a normal prior is used for $\text{logit}(p) = \ln(p/1 - p))$ with a mean of zero and large variance. This prior is a lot like the beta $Be(0, 0)$ prior, and we also recommend against its general use. For this example, however, the prior choice makes relatively little difference to posterior inference (Figure 5).

Discussion

Interpreting the likely magnitude of an effect in the context of its practical implication is important, and it is useful to be able to associate probabilities with various partitions of the parameter space. However, the method of Batterham and Hopkins lacks a proper theoretical foundation except in the context of Bayesian inference. In particular, the only coherent way of assigning probabilities of events conditional on the data is through adopting a fully Bayesian framework. This requires explicit use of priors.

As we have shown, the approach of Batterham and Hopkins is far from an essentially Bayesian approach ". . . with no prior assumption about the true value of the parameter." A more accurate description of their method is "approximately" Bayesian, and, rather than assuming no prior information, their approach has a very specific, but hidden, joint prior on parameters. This prior expresses certain assumptions about parameters and it might not be the prior that we want to use.

For simple problems involving inference about a mean from normally distributed data, the approach of Batterham and Hopkins is exactly Bayesian with use of the improper Jeffreys prior. A useful property in this case is that the posterior credible interval coincides exactly with the usual confidence interval, suggesting that this choice of prior provides a reasonable default for objective inference. For this particular problem, an excessively large sample size is not required before the posterior is insensitive to the choice of prior distribution. Usually a sample size of at least 20 will suffice. Unfortunately, this will not be true in general, particularly for data that are not normally distributed. For such problems, exact small-sample confidence intervals rarely exist and inference is usually based on asymptotic normality of the estimators.

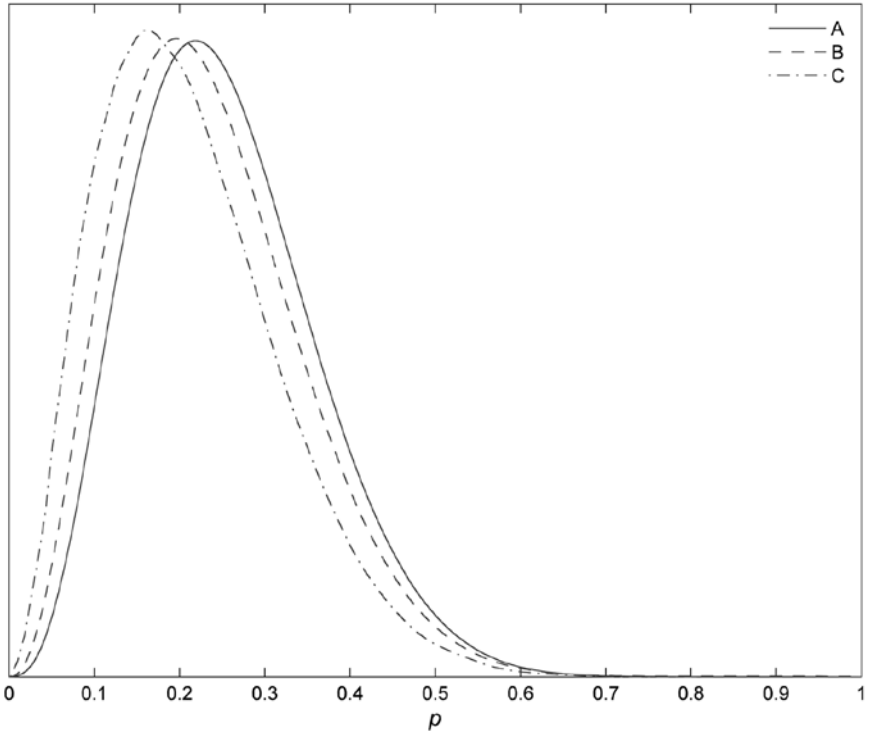

Figure 5 — Posterior density for the population proportion p, where $y = 3$ positive responses are observed in a sample of $n = 14$ individuals assuming a binomial $B(n, p)$ model for y and with three distinct priors. Prior A is a beta $Be(1, 1)$ distribution, prior B is a beta Be(1/2, 1/2) distribution, and prior C a logit-normal logitN(0, 1000000) distribution.

In the binomial example we consider above, the confidence interval usually used in practice, $\hat{p} \pm \sqrt{(\hat{p}(1-\hat{p}))/n}$, is based on the asymptotic normality of the maximum likelihood estimator $\hat{p}$. Even for reasonably large samples, this approximation can behave poorly and the true coverage rate of the confidence interval can be well below the nominal value.[8] Thus an adaptation of the Batterham and Hopkins approximation based on the asymptotic confidence interval will be unreliable, except in very large samples, in the sense that derived probabilities will not provide accurate statements about magnitudes of effects. However, once a prior has been specified, an exact solution is available using Bayesian methods as outlined in our example, which shows reassuring insensitivity to choice of prior despite poor performance of the frequentist estimator.

The advantages of adopting a full Bayesian approach rather than the approximate approach as outlined by Batterham and Hopkins are that nothing is hidden,

there are no approximations, and solutions are exact conditional on the assumptions in the model. Importantly, the type of inference based on magnitudes advocated by Batterham and Hopkins can only be carried out using Bayesian methods. We can also readily extend the method to more complicated data structures than inference about the mean of normally distributed data. For example, Bayesian methods of inference are also well suited to hierarchical modeling, which is useful for complex data modeling problems as well as for individual prediction.[11] A further advantage of Bayesian inference is that all uncertainty is represented by probability and it is easy to extend coherent statistical inference to coherent decision making in the presence of uncertainty.[10]

A difficulty with adopting Bayesian methods is that practical instruction in Bayesian inference does not appear in the statistics courses that most sport scientists are exposed to. The large-scale adoption of methods of Bayesian inference is a recent phenomenon in statistics, taking place mostly in the last 15 years. This change has not filtered down to introductory statistics courses. To accomplish the aims of the type of analysis suggested by Batterham and Hopkins, scientists need to be familiar with statistical modeling and Bayesian methodology. Two excellent texts we recommend are Gelman et al (2004)[3] and Lee (2004).[12] Another useful introductory text with emphasis on carrying out computations is the statistical package R in Albert (2007).[13]

Batterham and Hopkins[1] finished their article with a discussion of the need for methods of inference that can apply to individuals. Hierarchical Bayesian models allow us to investigate individual effects, provided the data exist to support the aims of the analysis. To make inference about a particular individual, data must be collected on that individual. In many cases, we can improve our inference about a particular individual through incorporating a postulated relationship between different individuals directly into the statistical model. For example, as of 30 January 2008, Asafa Powell was ranked number 1 on the all-time list for 100 m and Justin Gatlin, number 2. If we believe that Justin Gatlin's performances can help explain those of Asafa Powell, then we can obtain better predictions for Powell making use of data from Gatlin. However, to do this we need to describe how the performances of the two athletes are related. This connection is formalized through the statistical model.

References

1. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 2006;1:50–57.
2. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Sportscience.* 2005;9:6-13.
3. Gelman AB, Carlin JS, Stern HS, Rubin DB. *Bayesian data analysis.* 2nd ed. Boca Raton: Chapman and Hall/CRC; 2004.
4. Marshall SW. Commentary on Making Meaningful Inference about Magnitudes. *Sportscience.* 2005;9:43–44.
5. Senn S, Two cheers for P-values. *J Epidemiol Biostat.* 2001;6(2)193–204.
6. Moran JL, Soloman PJ. A farewell to P-values? *Crit Care Resuscitation.* 2004;6:130–137.
7. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Stat.* 1984;12:1151–1172. doi:10.1214/aos/1176346785

8. Agresti A, Coull BA. Interval estimation for a binomial proportion. comment *Stat Sci.* 2001;16:117–120.
9. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *J Am Stat Assoc.* 1996;91:1343–1370. doi:10.2307/2291752
10. Smith JQ. *Decision analysis: A Bayesian approach.* London: Chapman and Hall; 2004.
11. Barker RJ, Schofield MR. Classifying individuals as physiological responders using hierarchical modelling. *J Appl Physiol.* 2008; in press.
12. Lee P. *Bayesian statistics: An introduction*. 3rd ed. London: Hodder Arnold; 2004.
13. Albert J. *Bayesian computation with R.* New York: Springer; 2007.