Spreadsheets in Education (eJSiE)

Volume 5 | Issue 3 Article 3

10-3-2012

Teaching Bayesian Parameter Estimation, Bayesian Model Comparison and Null Hypothesis Significance Testing Using Spreadsheets

Christopher R. Fisher
Miami University, fisherc2@miamioh.edu

Christopher R. Wolfe
Miami University

Teaching Bayesian Parameter Estimation, Bayesian Model Comparison and Null Hypothesis Significance Testing Using Spreadsheets

Abstract
Learning statistics is often characterized by tedium and frustration. To make matters worse, pervasive misunderstandings of conditional probabilities often impede learning. We present an interactive spreadsheet designed to elucidate such misconceptions through a comparison three statistical approaches: Bayesian parameter estimation, Bayesian model comparison and null hypothesis significance testing. Learning is facilitated through the systematic exploration of each method and the use of graphical displays of the distributions. The conceptual underpinnings of each approach are described as well as their implementation in the spreadsheet. We conclude with some suggested pedagogical questions designed to elucidate the commonalities and differences between each approach.

Keywords
Null Hypothesis Significance Testing, Bayesian Statistics, Statistics, Inferential Statistics

Teaching Bayesian Parameter Estimation, Bayesian Model Comparison and Null Hypothesis Significance Testing Using Spreadsheets

Abstract

Learning statistics is often characterized by tedium and frustration. To make matters worse, pervasive misunderstandings of conditional probabilities often impede learning. We present an interactive spreadsheet designed to elucidate such misconceptions through a comparison of three statistical approaches: Bayesian parameter estimation, Bayesian model comparison and null hypothesis significance testing. Learning is facilitated through the systematic exploration of each method and the use of graphical displays of the distributions. The conceptual underpinnings of each approach are described as well as their implementation in the spreadsheet. We conclude with some suggested pedagogical questions designed to elucidate the commonalities and differences between each approach.

Introduction

Uncertainty is inherent in most domains of life, including science. Even the most well controlled experiment is susceptible to experimental error. One of the primary goals of inferential statistics is to manage such uncertainty by allowing researchers to distinguish between systematic and random variability in data. For this reason, the ability to reason statistically is crucial for scientists as well as consumers of scientific knowledge. Unfortunately, many students endorse misconceptions that compromise the correct interpretation of statistics [12]. Extensive evidence from the psychological literature reveals pervasive misunderstandings of the conditional probabilities upon which null hypothesis significance testing (NHST) and Bayesian statistics are based [1] [4] [16]. Concepts based on conditional probabilities, such as p-values, confuse not only students, but also experienced research psychologists and even statisticians [7].

One common misunderstanding of conditional probabilities is termed the conversion error, which occurs when P(A|B) is erroneously judged to be equal to P(B|A) [16] [17]. In general, P(A|B) ≠ P(B|A), as can be demonstrated with the following simple example: P(Cloudy Skies|Rain) ≠ P(Rain|Cloudy Skies). When it is raining, the chance of cloudy skies is 1.00 Conversely, the presence of clouds does not guarantee rain. Understanding the difference between P(Data|Hypothesis) and P(Hypothesis|Data) is critical for the correct interpretation of NHST and Bayesian Statistics (henceforth, H = Hypothesis and D = Data). A second common misunderstanding concerns the role of base rates (e.g. an average rate in a population) when reasoning with conditional probabilities [1] [4]. People tend to underweight base rates relative to individuating information, thereby

violating Bayes' theorem. This so-called base rate neglect has direct implications for the construction of prior distributions in Bayesian Statistics and the evaluation of improbable claims, such as extra-sensory perception [14].

Currently, there are few, if any, interactive, pedagogic resources available for comparing NHST and Bayesian statistics. Our goal is to fulfill this need with an interactive spreadsheet that allows direct comparisons between NHST and two Bayesian approaches—Bayesian parameter estimation and Bayesian model comparison. Comparing these approaches is important for two reasons—the first of which is pragmatic. The use of NHST is ubiquitous in science, especially the social sciences. Despite its popularity, considerable controversy and discontent surround the use of NHST [10] [13]. Bayesian statistics have been proposed as a viable alternative and could possibly supersede NHST in the future [6]. Thus, having knowledge of each approach is important. The second reason is pedagogical. We believe that misconceptions, such as conversion errors, can be ameliorated through this process of comparison because NHST provides results based on P(D|H) while the Bayesian approaches produce results based on P(H|D). Thus, an understanding of each approach may elucidate the distinction between P(D|H) and P(H|D). Each approach is demonstrated through a simple example of testing a coin for bias. One advantage of an interactive spreadsheet is that the distributions are displayed graphically and change in real time according to user input. In addition, the spreadsheet allows users to systematically explore the properties of each approach without the need for programming knowledge (e.g. Winbugs or R). Although the example used in the spreadsheet pertains to testing a coin for bias, it can be extended to any situation involving binary outcomes. Moreover, the underlying concepts presented herein form the foundation for more complex applications, including hierarchical models with multiple parameters.

The remainder of the article is structured as follows. First, we introduce key theoretical concepts for NHST and Bayesian statistics and explain how to use the spreadsheet. A ready-to-use version of the spreadsheet can be downloaded from the Spreadsheets in Education (eJSiE) website. In this article, we focus primarily on the Bayes factor in the tab labeled Bayes factor but discuss the other representations of evidence in the Bayesian Model Comparison Section. Next, we explain the implementation of the formulae in the spreadsheet for the interested reader. Lastly, we conclude with six problems designed to illustrate important concepts and differences between each statistical approach. Based on our experience in the classroom, we found that presenting the theory along with the six problems is highly beneficial. Explaining the details of the implementation does not appear to have a pedagogical benefit and may even obfuscate the concepts.

Null Hypothesis Significance Testing

NHST is based on the frequentist interpretation of probability as the relative frequency of an event in a large number of repetitions [13]. The history of NHST is fraught with contention between two theoretical approaches: the Fisherian approach and the

Neyman-Pearsonian approach [8] [13]. According to the Fisherian approach, the p-value (described below) serves as an index of evidence against the null hypothesis, such that smaller p-values indicate more evidence against the null hypothesis. By contrast, the Neyman-Pearsonian approach prescribes a binary decision procedure for rejecting the null hypothesis wherein type I and type II error rates are controlled across repeated experiments. Although a detailed discussion of each approach is beyond the scope of the present article, it should be noted that there is no consensus regarding the interpretation of the p-value [2] [3] [8]. For simplicity we adopt a unified theoretical approach [8], while noting it is not universally endorsed [2] [3].

In the simplest case, NHST begins with the specification of two complementary hypotheses —generally, a null hypothesis describing chance performance (i.e. no effect) and an alterative hypothesis describing either a directional or non-directional effect [10] [13]. For the present example, the null hypothesis is that the coin is fair, which is formally stated as $H_0 : \theta = .50$. θ represents a possible parameter value for the proportion of heads. The alternative hypothesis is that the coin is biased, which is formally stated as $H_1: \theta \neq .50$. A criterion, α, is specified to control type I errors—the conditional probability of falsely concluding an effect exists assuming the null hypothesis is true. By convention $\alpha = .05$. Once the data are collected, a test statistic is compared to a theoretical sampling distribution that represents all possible test statistics under the assumption that the null hypothesis is true. A p-value is computed by determining the relative position of the obtained test statistic within the sampling distribution. In other words, the p-value is the probability of obtaining a test statistic at least as extreme as the one obtained. The presumed null hypothesis is rejected if the p-value $\leq \alpha$. However, if the p-value $> \alpha$, judgment is suspended. Within this unified framework, the magnitude of the p-value also quantifies the degree of evidence against the null hypothesis.

Before proceeding it is important to emphasize two points. First, the p-value is not a posterior probability. Whereas the posterior distribution refers to the probability of a hypothesis after data have been collected, the p-value refers to probability of the data assuming the null hypothesis is true. Second, the p-value is similar to $P(D|\theta = .50)$ in that both are conditional on θ. However, the p-value is more accurately defined as $P(|T| \geq T_{obs} |\theta = .50)$, where T and T_{obs} denote test statistic and observed test statistic, respectively. In other words, the probability of a test statistic at least as extreme as the one observed assuming the null hypothesis is true. Thus, the p-value is not the same as $P(\theta = .50|D)$.

Data	
Sample Size	30
Proportion of Heads	0.60

Figure 1: Data entry Table.

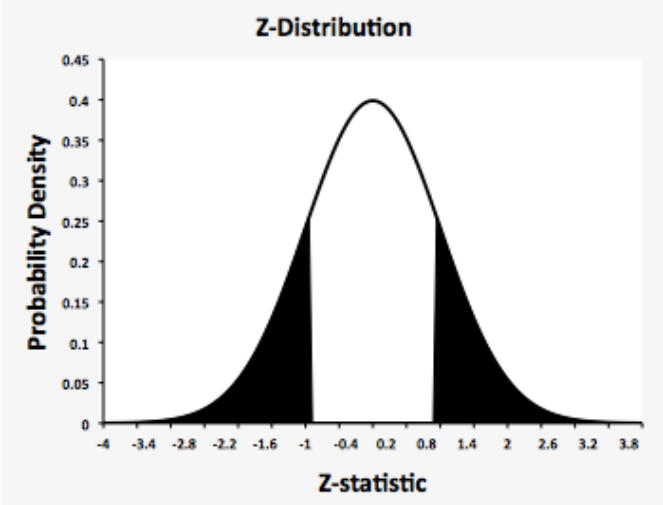

Figure 2: Sampling distribution of Z-statistics. The shaded area corresponds to the p-value.

Null Hypothesis Significance Testing	
Normal Approximation	
Z-statistic	0.91
p-value	0.36
Null value	0.50

Figure 3: Z-statistic, p-value and null parameter.

Normal Approximation to Binomial

A normal approximation was used rather than a binomial distribution for several reasons[1]. The use of graphical displays is integral to our approach for conveying concepts based on distributions and p-values. One problem with the binomial distribution is that it cannot be implemented in graphical form because the size of the referenced data changes according to sample size. The normal approximation has several other advantages in terms of pedagogy. For example, it is used commonly in practice because it is more tractable computationally for large sample sizes. Another reason is that it provides a seamless transition to distributions of continuous variables, which are more commonly used in practice. The normal distribution is a reasonable approximation to a binomial distribution when $n\theta_0 \geq 5$ and $n(1-\theta_0) \geq 5$, where n is the sample size and θ_0 is the parameter against which the null hypothesis is tested [9]. The corresponding Z-statistic is computed as follows:

$$(1)\ Z = \frac{(x - n\theta_0) \pm .5}{\sqrt{n\theta_0(1-\theta_0)}}$$

where x denotes the number of successes in n trials.

Hypothetical data may be entered into the spreadsheet in the table entitled Data, as shown in Figure 1. The table requires two inputs: the sample size (cell B2) and the proportion of observed heads (Cell B3), which are referenced by each of the three statistical approaches. Figure 2 displays the resulting sampling distribution. The shaded area of the sampling distribution is proportional to the p-value. Figure 3 displays the Z-statistic and p-value that correspond to the sampling distribution in Figure 2. The parameter for the null hypothesis can also be changed in this table (cell 34).

Bayesian Parameter Estimation

In the Bayesian framework, probability indexes subjective degree of belief that an event will occur [13]. Unlike the frequentist interpretation, probabilities may be assigned to non-repeatable events, such as a candidate winning a particular election. At a conceptual level, Bayesian statistics is intuitive. One begins with a set of beliefs regarding a particular phenomenon and updates those beliefs as data become available. However, the underlying computations and the interplay between the prior, likelihood and posterior distributions are complex, as detailed below. This process of updating beliefs is formalized through Bayes theorem. One key feature of Bayes' theorem is that it enables $P(\theta|D)$ to be inferred from $P(D|\theta)$. In general, Bayes' theorem is defined as:

$$(2)\ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$$

where $P(\theta|D)$ is the posterior probability of θ, $P(D|\theta)$ is the probability of the data given θ, $P(\theta)$ is the prior probability of θ and the denominator is a normalizing constant formed by integrating across all values of θ. Each of these components will be discussed in turn.

Prior Distribution

The primary goal of Bayesian parameter estimation is to identify the relative plausibility of candidate parameter values while taking into account prior beliefs and newly acquired data [6]. The first step is to specify a prior distribution that characterizes the relative plausibility of candidate parameters before additional data is acquired. As a theoretical construction, the prior distribution is unobservable unlike the data. However, the prior distribution can be based upon the posterior distribution from previous experiments. One important constraint in specifying a prior distribution is that it must be derived from existing theory and data and must ultimately pass the scrutiny of qualified reviewers.

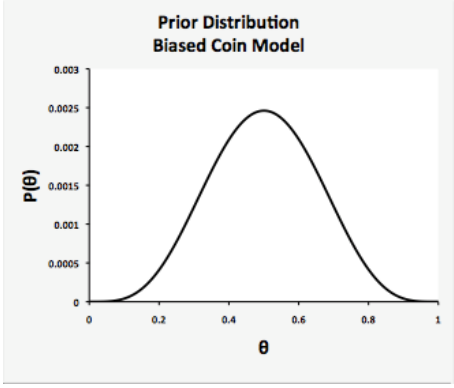

Figure 4: A symmetrical prior distribution with a mean = .50 and N = 10.

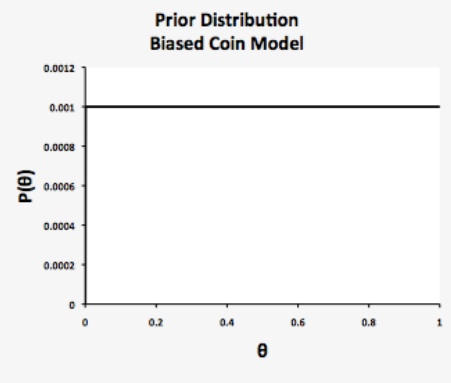

Figure 5: A flat prior distribution with a mean = .50 and N = 2.

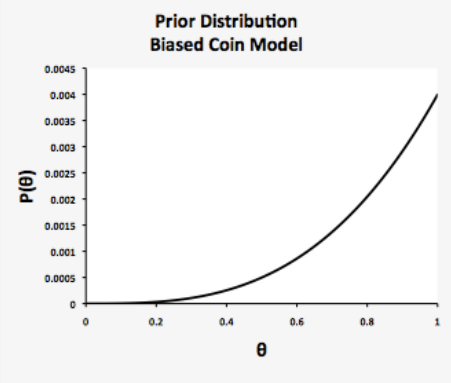

Figure 6: A skewed prior distribution with a mean = .80 and N = 5.

A beta distribution is used to characterize the prior distribution of θ because it is constrained to the interval [0,1] and can assume many qualitatively different shapes, as shown in Figures 4, 5 and 6. Another attractive feature of the beta distribution is that the prior and posterior distributions are conjugate when used with a binomial likelihood function [13]. This means that the resulting posterior distribution is also a beta distribution, and thus, has the same parametric interpretation. The beta distribution is governed by two parameters: α and β. In the present example, α represents the outcome heads and β represents the outcome tails. The ratio of α to β governs the central

tendency of the distribution. For example, α = 2 and β = 2 produces a beta distribution with a mean of .50. The magnitude of α and β governs the dispersion of the beta distribution, such that higher magnitudes produce less dispersion. The beta distribution in the spreadsheet is re-parameterized to be more intuitive [5]. The mean of the beta distribution is defined as mean = $\alpha/(\alpha + \beta)$, while the sample size is defined as N = $\alpha + \beta$, which governs the dispersion of the distribution. At a conceptual level, the mean represents the average value of θ, with higher values of N corresponding to higher degrees of certainty in the mean value. As shown below in Figure 7, these parameters may be entered in the table of the spreadsheet entitled Bayesian Parameter Estimation under the subheading Prior Distribution (cells B9 and C9).

Bayesian Parameter Estimation		
Prior Distribution		
	Mean	**N**
Parameters	0.50	10
Posterior Distribution		
	Lower Bound	**Upper Bound**
HDI	0.42	0.72
	Mean	**N**
Parameters	0.58	40.00
ROPE		
Percentage	**Lower Bound**	**Upper Bound**
12.91%	0.48	0.52

Figure 7: Table for Bayesian Parameter Estimation.

Likelihood Distribution

The likelihood distribution represents the degree to which values of θ are consistent with the observed data [6]. Formally, the likelihood distribution is generated by the binomial likelihood function:

$$(3)\ P(D|\theta) = \binom{n}{h}\theta^h(1-\theta)^{n-h}$$

where n is the sample size and h is the number of observed heads. As to be expected, the likelihood distribution peaks at the observed proportion of heads, with the likelihood of θ decreasing as it departs from the observed proportion of heads (see Figure 8).

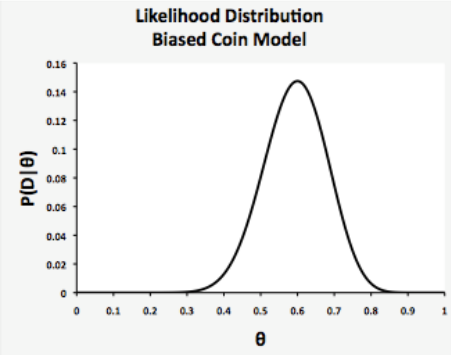

Figure 8: The likelihood distribution for sample size = 30 and proportion of heads = .60.

Posterior Distribution

Up to this point we described the specification of a prior distribution that describes pre-experimental beliefs and the likelihood distribution that results from observed data. The posterior distribution can be conceptualized as a resolution between the prior and likelihood distributions, formed by averaging the distributions together (see Figure 9). This conceptualization is reinforced by the fact that the mean of the posterior distribution is a weighted average of the observed proportion of heads and the mean of the prior distribution [5]:

$$(4)\quad \underbrace{\frac{h+\alpha}{N+\alpha+\beta}}_{posterior} = \underbrace{\left(\frac{h}{N}\right)}_{data}\underbrace{\left(\frac{N}{N+\alpha+\beta}\right)}_{weight} + \underbrace{\left(\frac{\alpha}{\alpha+\beta}\right)}_{prior}\underbrace{\left(\frac{\alpha+\beta}{N+\alpha+\beta}\right)}_{weight}$$

where h is the number of observed heads N is the sample size and α and β are the original parameters of the beta distribution. More formally, the posterior distribution is defined as:

$$(5)\ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where,

$$(6)\ P(D) = \int P(D|\theta)P(\theta)d\theta$$

Thus, P(D) is a normalizing constant formed by summing each likelihood of θ weighted by its prior probability. The posterior distribution is commonly summarized with respect to its parameters and a 95% highest density interval (HDI)—a Bayesian analog to the 95% confidence interval in NHST. As shown in Figure 7, the parameters of the posterior distribution (cells B14 and C14) and 95% HDI (cells B12 and C12) can be found under the subheading Posterior Distribution.

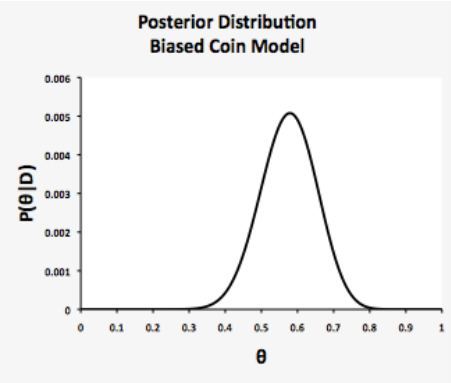

Figure 9: The posterior distribution of θ.

While the posterior distribution is rich in information, it is possible to simplify this information with a binary decision procedure similar to that used in NHST [6]. As in NHST, a researcher must first define null hypothesis—in this case, θ = .50. The null hypothesis is rejected if θ = .50 is not included in the 95% HDI. Otherwise judgment is suspended. Alternatively, one may define a region of practical equivalence (ROPE), an interval within which θ is deemed equivalent to the null parameter for practical purposes. Although the ROPE is determined on a case-by-case basis, Figure 7 shows this to be θ ∈ [.48, .52]. As before, the null hypothesis is rejected when the 95% HDI does not overlap with the ROPE. One advantage of using a ROPE, as opposed to a point null, is that it is possible to accept the null hypothesis because the 95% HDI converges on the true value as the sample size increases. The null hypothesis can be accepted if the 95% HDI falls entirely within the ROPE. In addition, the value labeled Percentage in Figure 7 indexes graded levels of support for the null hypothesis. It represents the proportion of the posterior distribution that overlaps with the ROPE.

Bayesian Model Comparison

The goal of Bayesian model comparison is to determine the relative plausibility of two or more models [6]. A model may refer to a simple hypothesis (e.g. the coin is fair) or a complex mathematical model. Comparisons can be made between nested or non-nested models. Continuing with the present example, the two models of interest are the fair coin model and the biased coin model. A prior distribution of θ must be specified for both models. By definition, the prior distribution is P(θ = .50) = 1.00 for the fair coin model, as shown below in Figure 10. However, the prior distribution of θ for the biased coin model can, in principle, assume many forms. To allow direct comparisons between both Bayesian approaches, the parameters for the biased coin model are entered in the table for Bayesian parameter estimation under prior distribution. As with Bayesian parameter estimation, the prior distribution must map onto the characteristics of the model. One reasonable model of a biased coin may use the parameters: mean = .50 and N = 10. Such a model assumes small degrees of bias are more likely than large degrees of bias and bias towards heads is as likely as bias towards tails.

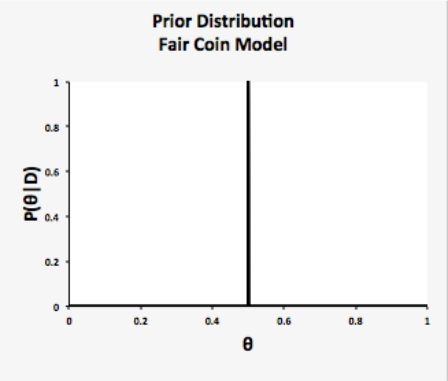

Figure 10: The Fair Coin Model. P(θ = .50) = 1.00.

Bayes' theorem for model comparison is defined below:

$$(7)\ P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{i=1}^{M} P(D|M_i)P(M_i)}$$

where M_i is the i^{th} model and

$$(8)\ P(D|M_i) = \int P(D|\theta, M_i)P(\theta|M_i)d\,\theta$$

Bayes Factor

At a conceptual level, each model is compared in terms of its concordance with the data. Concordance is indexed using the Bayes factor, which is computed from the ratio of likelihoods: $P(D|M_i)/P(D|M_j)$. In terms of the present example, the Bayes factor is:

$$(9)\ \frac{P(D|M_{Biased})}{P(D|M_{Fair})} = \frac{\int P(D|\theta,M_{Biased})P(\theta|M_{Biased})d\theta}{P(D|\theta{=}.50)}$$

The Bayes factor provides graded levels of support for one model relative to another. For example, a Bayes factor of 2 would indicate the data favor the biased coin model twice as much as the fair coin model. Thus, it is the factor by which one's prior beliefs should be adjusted in light of the evidence. Although the Bayes factor is a continuous measure, some suggested categories for interpretation are presented in Table 1. One difficulty in interpreting Bayes factors concerns its asymmetry. For example, a Bayes factor of 6 indicates moderate support for the biased coin model, while a Bayes factor of .167 indicates the same degree of support for the fair coin model. To overcome this problem, the spreadsheet displays the reciprocal of the Bayes factor when it is less than 1 and indicates which model received more support. Using the previous example, the spreadsheet would convert .167 to 6 and display "Supports Fair Coin Model" (see Figure 11).

Table 1: Categories for interpreting Bayes factors adapted from [15].

Bayes Factor (BF)	**Interpretation**
1	Equal evidence
1 < BF ≤ 3	Anecdotal evidence
3 < BF ≤ 10	Substantial evidence
10 < BF ≤ 30	Strong evidence
30 < BF ≤ 100	Very strong evidence
100 < BF	Decisive evidence

A similar configuration for the posterior odds is used because it can disagree with the Bayes' factor under certain circumstances, as demonstrated in Figure 11. Alternative methods for facilitating the interpretation of Bayes factors can be found in the tabs

labeled Log Bayes Factor and Probability. For example, the Bayes factor can be symmetrized using a logarithmic transformation. Another method is to convert the Bayes factors to probabilities.

Posterior Odds

It is important to note that the Bayes factor is a likelihood ratio rather than the posterior odds [6]. In order to obtain the posterior odds, the prior probabilities of each model must be specified, as shown in Figure 11 under Prior Probability of Model. Only the prior probability for the Fair Coin model must be specified (cell B23) because the prior probabilities are complementary. When the prior probability of each model is equal, the Bayes factor equals the posterior odds.

Bayesian Model Comparison		
Prior Probability of Model		
	Fair Coin	Biased Coin
Probability	0.20	0.80
Supports Fair Coin Model		
Bayes Factor	1.28	
Supports Biased Coin Model		
Posterior Odds	3.12	

Figure 11. Table for Bayesian Model Comparison.

Implementation

In this section, we outline the implementation of the formulae beginning with the normal approximation to the binomial distribution. In cell B32, the Z statistic is computed according to formula (1). The p-value is computed with the NORM.S.DIST function, as follows: B33 = =2*MIN(NORM.S.DIST(B32,1),(1-NORM.S.DIST(B32,1))). The minimum value is multiplied by two to reflect that it is a non-directional test. Correction for continuity is employed using the IF function in cell B32. If $x < n\theta_0$, then .5 is added to the numerator. If $x > n\theta_0$, then .5 is subtracted from the numerator. Otherwise the numerator will equal 0.

One difficulty in implementing Bayesian statistics in Excel is that it requires the use of integrals. This problem is resolved in the following manner: A continuous prior distribution is approximated by taking the difference between the cumulative distribution for values of θ and θ - .001 over the interval [0,1]. Column AJ enumerates θ in increments of .001, while column AL references θ in column AJ and computes the cumulative difference. To understand the logic of this approach, imagine the non-contiguous distribution that would be formed by computing the probability density for each .001 increment of θ. The resulting gaps can be filled by computing the cumulative difference for θ and θ - .001. Approximating the prior distribution in this fashion has several advantages. First, it provides a degree of precision that is sufficient for pedagogical purposes[1]. Second, the prior distribution sums to 1.00, as necessitated by the law of total probability. By extension, the posterior distribution also adheres to the law of total probability.

The likelihood distribution is produced in column AK by applying the binomial probability mass function across all values of θ: =BINOMDIST($B$2*$B$3,$B$2,AJ2,0). The posterior distribution is computed in two steps. First, corresponding cells in columns AK and AL are multiplied and the intermediate values are recorded in column AM. Second, these values are normalized by dividing each cell by the sum of column AM and storing them in column AN: AN2 = AM2/SUM($AM$2:$AM$1002). Two steps are used to approximate the 95% HDI. First, each value of θ in the posterior distribution is converted to percentiles with a cumulative sum in column AO: AO2 = SUM($AN$2:AN2). Next, the match function returns the rank order associated with the percentiles .025 (lower bound) and .975 (upper bound) in cells B12 and C12, respectively. For example, the lower bound is computed as follows: B12 = (MATCH(0.025,AO2:AO1002)-1)/1000. One is subtracted from the rank order and divided by 1000 to transform the rank order to the correct value of θ. The parameters of the posterior distribution (cells B14 and C14) are computed using formula (4). The ROPE is implemented in the spreadsheet by summing the cumulative probability values in column AO that correspond to the user defined ROPE interval in Cells B17 and C17: A17=SUMIFS(AN2:AN1003,AJ2:AJ1002,">="&B17,AJ2:AJ1002,"<="&C17).

The Bayes factor and posterior odds are implemented in the spreadsheet in the following manner: For interpretability the Bayes factor and posterior odds are transformed using the IF function such that both are ≥ 1. Cells A24 and A26 use the IF function and the untransformed Bayes factors and posterior odds indicate which model is supported. The Bayes factor in cell B25 is a likelihood ratio of the Biased and Fair Coin Models. The Biased Coin Model is equivalent to the normalizing constant P(D): SUM(AM2:AM1002). The Fair Coin Model is simply the likelihood of the null parameter in cell B34 of the Null Hypothesis Significance Testing Table. VLOOKUP returns the likelihood of the null parameter defined by the user: BA3 = VLOOKUP(B34,AJ2:AK1002,2)). The posterior odds in cell B27 are computed by multiplying the prior odds, defined as a ratio of cells C23 and B23, by the Bayes factor.

Problems

In this section, we provide problems with suggested answers. Although it is by no means an exhaustive collection of problems, we believe these problems underscore important features of each approach.

Problem 1: Calibration

A coin is flipped 120 times to determine whether it is fair or biased, resulting in 60% heads. Set the biased coin model to mean = .50 and N = 10. Interpret the p-value and the Bayes factor. How do they compare?

Suggested answer: According to the Neyman-Pearsonian perspective, the null hypothesis that the coin is fair should be rejected. The Bayes factor indicates "anecdotal"

or weak support in favor of the biased coin model. Although both methods agree, they are calibrated differently. See [15] for details on calibration.

Problem 2: Sensitivity to Model

Continuing with Problem 1, change the biased coin model to observe its effects on the Bayes factor and p-value. As a starting point you may want to try mean = .5, N = 2 and mean = .65, N = 30. What happened and why?

Suggested answer: The p-value was invariant to changes in the biased coin model because it only compares the observed result to sampling distribution based on the assumption that the null hypothesis is true. The Bayes factor is sensitive to different models because the fair and biased coin models are compared relative to each other. For example, when the biased coin model was flat (i.e. mean = .5, N = 2), it received less support because the data were inconsistent with extreme values of θ (i.e. $\theta = 0$, $\theta = 1$), which received equal weight in the model as moderate values (i.e. $\theta = .50$). Evidence for the biased model increased when mean = .65 and N = 30 because the data were more consistent with it than the fair coin model.

Problem 3: Robustness of Bayesian Parameter Estimation

Examine the effect of the two different priors used in Problem 2 on Bayesian parameter estimation. How does this compare to your answer in problem 2. Why is this the case?

Table 2: Comparison of two posterior distributions based on two prior distributions and a large sample size.

Prior Distribution		
	Mean = .5, N = 2	**Mean = .65, N = 30**
95% HDI	.51-.68	.52 - .69
Mean	.60	.61
N	122	130

Suggested answer: Compared to the Bayes factor, Bayesian parameter estimation is relatively stable in this case (see Table 2). Bayesian model comparison measures the evidence for one model *relative* to another, whereas Bayesian parameter estimation takes a weighted average of the prior distribution and likelihood distribution to form a posterior distribution. By comparing the data to the priors in terms of the parameters, it is evident that the data receive more weight and exert more influence on the posterior distribution (N = 120 for the data vs. N = 2 and N = 30 for the priors).

Problem 4: HDI's and Confidence Intervals

Create a flat prior distribution by setting Mean = .5 and N = 2. Record the 95% HDI. Compute a 95% confidence interval for NHST. Disregarding rounding error, how do they compare. What does this imply about NHST?

Suggested answer: Continuing with Problem 3, the 95% HDI and 95% confidence interval are equal in this special case. This implies that NHST weights each parameter equally. The degree to which the 95% HDI will differ from the confidence interval depends on the prior distribution. Moreover, based on the answer from Problem 3, the influence of the prior distribution on the 95% HDI is moderated by the data because the posterior distribution is a weighted average of the prior and likelihood distributions.

Problem 5: Prior Probability of Models

Suppose that you found a coin on the street and want to determine whether it is fair or biased. Construct a biased coin model and set the prior probabilities of each model. Further suppose, the coin was flipped 20 times and resulted in heads 55% of the time.

Now suppose a slightly different situation. This time your uncle gives you a coin and your uncle has a reputation for being a trickster. Use the same biased coin model and specify the prior probabilities of the models for this situation. Assume, again, that the coin was flipped 20 times and resulted in 55% heads. Explain how you chose the prior probabilities for the models and compare the Bayes factor and prior odds in both cases.

Table 3: A comparison of the Bayes Factor and prior odds based on two prior distributions.

	Bayes Factor	**Prior Odds**
Street: Prior fair = .80	2.56 Biased	1.56 Fair
Uncle: Prior fair = .20	2.56 Biased	10.26 Biased

Suggested answer: Although the results may vary from person to person, the qualitative pattern of results will match those depicted in Table 3. The prior probability of the fair coin model is higher when the coin was found in the street. This reflects the fact that the uncle has a reputation for being a trickster. The Bayes factors are the same in both situations because the results of the coin flipping experiment were the same. However, the posterior odds reflect the fact the prior probabilities are different across situations. For the coin that was found in the street, the prior odds show weak evidence in favor of a fair coin. The case is much different for the uncle's coin. The posterior odds indicate strong support of the biased coin model.

Problem 6: Disagreement Regarding Priors

It is possible that your peers may disagree with the prior probabilities you choose for the models. Propose some solutions for resolving such disagreements.

Suggested answer: One possibility is to perform a sensitivity analyses. The posterior odds based on different prior probabilities may support same model, but disagree slightly to moderately in magnitude—in which case, the disagreement may be minimal. However, if the posterior odds support different models or indicate large differences in magnitude, this reflects a state of uncertainty that may be resolved by simply collecting additional data.

Conclusion

The spreadsheet we presented is interactive and easy to use. It allows the user to explore the properties of NHST and two Bayesian approaches and make direct comparisons between them. The graphical displays of distributions may facilitate understanding of certain concepts. For example, by visual examination the distributions, it is easy to see that the posterior distribution is a weighted average of the prior and likelihood distributions. This fact may be obscured by the formalism of Bayes' theorem. The spreadsheet may reduce conversion errors by demonstrating that NHST is based on P(Data|Hypothesis), whereas the Bayesian approaches use prior distributions to infer the converse probability, P(Hypothesis|Data).

References

1. Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30, 241–254.
2. Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *American Statistician*, 59, 121-126.
3. Hubbard, R., & Bayarri, M.-J. (2003). Confusion over measures of evidence(p's)versus errors(15 's) in classical statistical testing. *American Statistician*, 57, 171-182.
4. Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.
5. Kruschke, J.K. (2010). Doing Bayesian data analysis: A tutorial introduction with R and BUGS. Burlington, MA: Academic Press.
6. Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, **6**(3), 299-312.
7. Lecoutre, M.-P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology, 38*(1), 37-45.
8. Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88, 1242–1249.
9. Mann, P. S. & Lacke, C. J. (2010). Introductory Statistics (7th ed.). United States of America, Wiley & Sons.
10. Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301.
11. Reyna, V. F. (2004). How people make decisions that involve risk. A dual-processes approach. *Current Directions in Psychological Science* 13:60–66.

12. Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98-113.

13. Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. Psychonomic Bulletin & Review, 14, 779-804.

14. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H.L.J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100, 426–432.

15. Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3), 291-298.

16. Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy-trace theory account. *Journal of Behavioral Decision Making*, 8, 85-108.

17. Wolfe, C.R., Fisher, C.R. & Reyna, V.F. (2012). Semantic Coherence and Inconsistency in Estimating Conditional Probabilities. *Journal of Behavioral Decision Making*: DOI: 10.1002/bdm.1756

Footnote

The spreadsheet is intended to be a pedagogical tool. P-values are generally in close agreement to those obtained from SPSS. Bayes Factors are generally accurate for sample sizes and effect sizes commonly observed in research. However, there is some loss of precision for very large Bayes Factors (i.e. > 1000). Users interested in comparisons to the binomial distribution may adapt the spreadsheet to their purposes using the BINOM.DIST function.