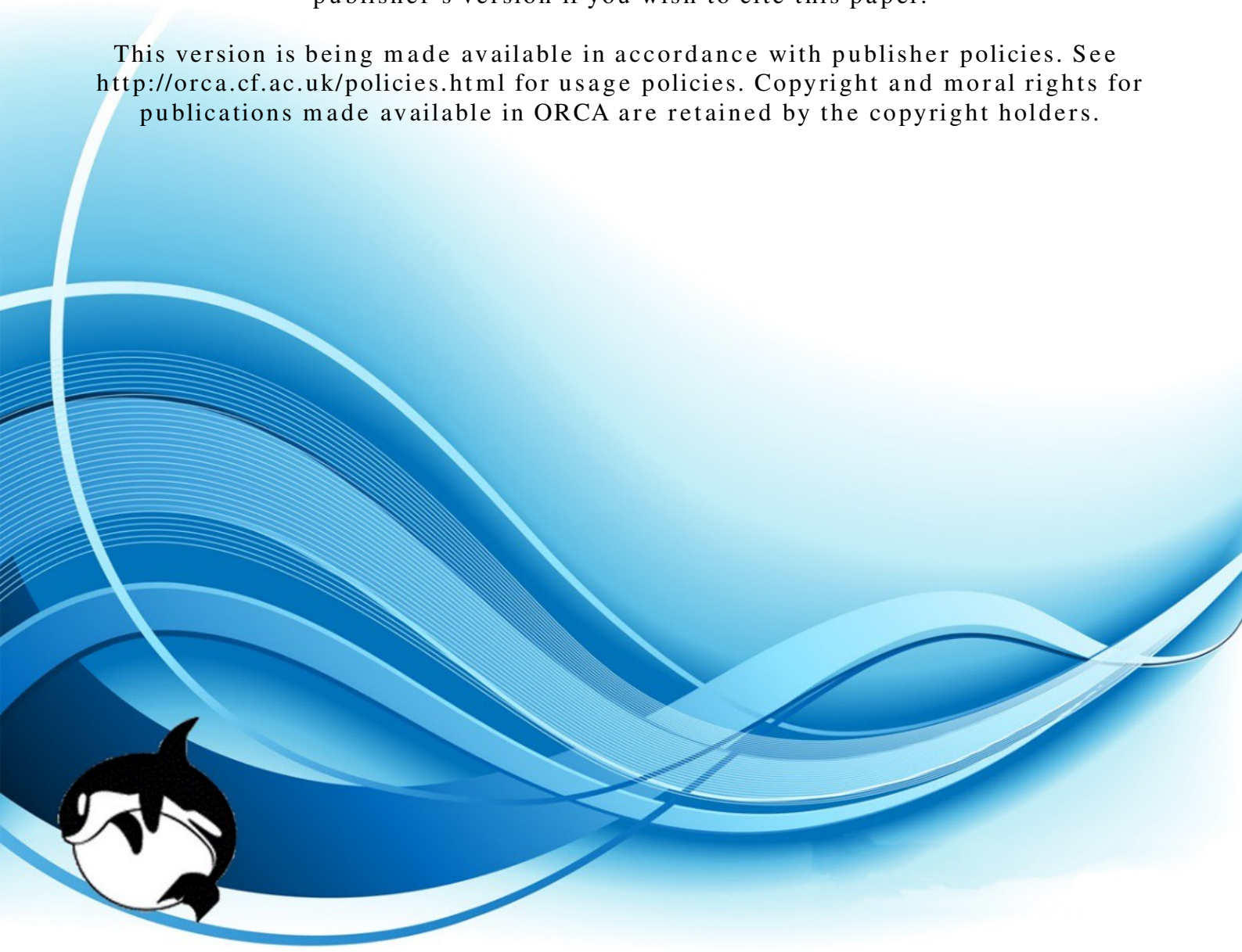


Running head: TEACHING BAYES' THEOREM

Teaching Bayes' Theorem: Strength of Evidence As Predictive Accuracy

Jeffrey N. Rouder

University of Missouri

Richard D. Morey

Cardiff University

Abstract

Although teaching Bayes' theorem is popular, the standard approach—targeting posterior distributions of parameters—may be improved. We advocate teaching Bayes theorem in a ratio form where the posterior beliefs relative to the prior beliefs equals the conditional probability of data relative to the marginal probability of data. This form leads to an interpretation that the strength of evidence is relative predictive accuracy. With this approach, students are encouraged to view Bayes' theorem as an updating mechanism, to obtain a deeper appreciation of the role of the prior and of marginal data, and to view estimation and model comparison from a unified perspective.

Teaching Bayes' Theorem: Strength of Evidence As Predictive Accuracy

As Bayesian statistics increases in popularity, it is essential to have effective ways of teaching Bayes' theorem. In this note, we present our approach that we find suitable for advanced undergraduates and beginning graduate students in an introduction to mathematical statistics or in a Bayesian analysis topics course. There has been much work on best methods for teaching conditional probability in introduction courses, especially with use of frequencies and intersections(Albert, 1997; Berry, 1997; Gigerenzer & Hoffrage, 1995, e.g.,). Bayes theorem, however, should be treated separately as Bayesians use Bayes theorem not only to update probabilities but as a basis of statistical inference. We assume for the following development that students are familiar with concepts of conditional, joint, and marginal probabilities and probability distributions. Our focus is on the implications of Bayes theorem as a measure of statistical evidence.

The Proportional Form of Bayes Theorem

In the vast majority of texts, Bayes' theorem is stated as,

$$\pi(\theta|Y) = \frac{p(Y|\theta)\pi(\theta)}{p(Y)}. \tag{1}$$

where $\pi(\theta|Y)$ and $\pi(\theta)$ denote the posterior and prior distributions of the (possibly multivariate) parameter θ, and $p(Y|\theta)$ and $p(Y)$ are the likelihood and marginal likelihood of the (possibly multivariate) data Y. For a continuous parameter, the marginal likelihood is

$$p(Y) = \int_{\Theta} p(Y|\theta)\pi(\theta)\, d\theta,$$

where Θ represents the parameter space of θ. For discrete parameters, the integration is replaced by summation.

After providing (1), many texts introduce a *proportional form* of Bayes theorem:

$$\pi(\theta|Y) \propto l(\theta; Y) \times \pi(\theta), \tag{2}$$

where the likelihood l is the probability of the observed data Y – that is, $p(Y|\theta)$ – as a function of θ. This form has a handy mnemonic—"the posterior is proportional to the likelihood times the prior." The proportional form may be illustrated with graphs of priors, likelihoods, and posteriors such as those in Figure 1. Here it may be seen that the posterior reflects the pull of both the likelihood and prior, and that there is no posterior mass where there is no prior mass.

When the proportional form is used, instructors will often (correctly) state that because $p(Y)$ is not a function of θ, it is simply a normalizing constant with an unknown value. Commonly-used posterior sampling methods require only knowledge of the posterior distribution up to a constant of proportionality; thus, $p(Y)$ may be ignored. Because many courses on Bayesian statistics make heavy use of such sampling methods (e.g., Markov chain Monte Carlo), it is perhaps not surprising that the proportional version is the one predominantly stressed in texts including Gelman, Carlin, Stern, & Rubin (2004), Jackman (2009), and Kruschke (2012).

We find, however, that when the proportional version is stressed, our students miss out on critical Bayesian elements. First, they do not develop an intuition about $p(Y)$, the marginal probability of data. This marginal is a uniquely Bayesian concept and intuition about it is critical for understanding Bayesian model comparison and Bayesian model criticism. Second, they tend to stress the output, the posterior, rather than the process of updating. Third, students take a dim view of priors. According to them, priors are

subjective while likelihoods are objective. Consequently, students value flat priors. Fourth, because $p(Y)$ is unused in estimation and is critical in model comparison, students see estimation and model comparison as separate rather than unified.

A Ratio Form Of Bayes Theorem

To address these difficulties and to promote a deeper understanding of Bayes' theorem, we follow a line of argument perhaps first presented by Carnap (1962). We augment the proportional form with the following *ratio form*:

$$\frac{\pi(\theta|Y)}{\pi(\theta)} = \frac{p(Y|\theta)}{p(Y)}. \tag{3}$$

Though (3) is simply a rearrangement of (1), this form makes clear some important implications of Bayes' theorem. We teach it as follows:

The left-hand-side of (3) concerns probabilities over parameters, or beliefs. The ratio describes how beliefs about values of θ are updated in light of data. Figure 2A shows an example where θ is the parameter in a binomial model. The datum in this case is 7 heads in 10 flips. The prior is a beta (2.5,1) that slightly favors larger values of θ; the posterior is beta (9.5,4). Two example points are provided, $\theta = .75$ and $\theta = .3$. For $\theta = .75$, the posterior and prior density is 3.24 and 1.62, respectively, and the updating factor is 2.0. Here the data have increased the plausibility of the point. For $\theta = .30$, the updating factor is .07, indicating the data have decreased the plausibility of the point markedly. The left-hand-side of (3), the updating factor, is shown as a function of θ in Figure 2B. As side exercises we ask students to find intervals where the data have decreased the plausibility by more than 10-1. We also ask students to explore how the prior affects the updating by plotting the left-hand side of (3) for different priors. This exercise can be extended to improper priors, say a beta (0,0) prior, where the updating

factor must be infinitely large. Such a peculiar state is not obvious in the proportional form, and motivates the need for caution when using improper priors.

We follow Jeffreys (1961) call this updating factor the *strength of the evidence* from the data about θ. Evidence from data is how the data license an update of beliefs.

The right hand side concerns of (3) concerns probabilities of observed data, and the term $p(Y)$ is particularly new. Students sometimes have the mistaken intution this term should have value 1.0 to reflect that the observed data were observed. To combat this intuition, we find it helpful to refer introduce probability mass functions over outcomes as *predictions*. We start with $p(Y|\theta)$, the numerator, as it is most accessible. If θ is specified, say at $\theta = .75$, then $p(Y|\theta)$ provides a probability distribution over outcomes (Figure 2C). Here, we can ask how well the observed datum, 7 heads in 10 flips, was predicted (see the starred point). We can compare this prediction to predictions from other models, and Figure 2D shows the case for $\theta = .30$. We also can introduce other prediction patters at this point, and show Figure 2E as an example. We point out to students that these whatever these patterns are, they must sum to 1.0. The implication is that if the prediction for one outcome is increased, the prediction for the others must be decreased to maintain the sum. With these three figures, students can compute ratios of how much better one pattern predicted the observe datum than another. We stress understanding these plots as predictions before seeing data, and that the comparison at the observed data is a measure of relative predictive accuracy.

Once students have been introduced to the concept of comparing predictions, we switch over to the concept of conditional and marginal predictions. Figures 2C and 2D are conditional predictions, conditional on specific values of θ. Figure 2E is marginal over all these predictions. Here we introduce the denominator, $p(Y)$, and its computation through the Law of Total Probability, namely, $P(Y) = \int_\theta P(Y|\theta)\pi(\theta)\, d\theta$. The term $p(Y)$ is termed the marginal prediction of the data (when weighted by the prior $\pi(\theta)$). The terms $p(Y|\theta)$

and $p(Y)$ are plotted as a function of θ in Figure 2F and labeled "conditional" and "marginal," respectively. The right-hand side of (3), the ratio, is shown in Figure 2G. We term this ratio the gain in predictive accuracy for θ.

Bayes rule states the equality of the left and right sides of (3), which can be seen by noting the equality of Figures 2B and 2G. The updating factor for a value of θ, the strength of evidence from the data, is how well the data are predicted when conditioned on this value relative to the marginal prediction. In words, we say that "strength of evidence for a parameter value is precisely the relative gain in predictive accuracy when conditioning on it" (see also Morey, Romeijn, & Rouder, 2016). We may even use the short-hand mnemonic, "strength of evidence is relative predictive accuracy." We find that allowing students to make this connection between evidence and prediction provides them with a deeper insight into Bayes theorem than afforded by the proportional form.

Unified Estimation and Model Comparison

To unify estimation and model comparison, we find it useful to introduce the concept of relative strength of evidence for competing parameter values. As an exercise we ask students to compare the relative evidence for two values θ_0 and θ_1:

$$\frac{\frac{\pi(\theta_1|Y)}{\pi(\theta_1)}}{\frac{\pi(\theta_0|Y)}{\pi(\theta_0)}} = \frac{\frac{p(Y|\theta_1)}{f(Y)}}{\frac{p(Y|\theta_0)}{f(Y)}} = \frac{p(Y|\theta_1)}{p(Y|\theta_0)}.$$

Here the relative strength of evidence is the ratio of probabilities of data, or gain in predictive accuracy. This development is an example of model comparison. We are comparing one model with a specific point value of θ_1 vs. another model with a specific point value of θ_2. This example may be leveraged to introduce model comparison more generally. In general, the relative strength of evidence is the *Bayes factor*, and the above example shows the Bayes factor for these two constrained models.

One advantage of the ratio form is that it seamlessly unifies parameter estimation and model comparison. The inclusion of $p(Y)$ in the RHS of (3) indicates that even parameter estimation yields only relative evidence. All specific θ values can be thought of as restrictions on a general model with a prior across all θ. Each of these restrictions is being implicitly compared to the model in which they are nested, whose likelihood is $p(Y)$.

To show students the unification, we let $\mathcal{M}_A$ be our previous model on θ, the probability of heads on a flip of a certain coin, defined by

$$\begin{aligned} Y \mid \theta &\sim \text{Binomial}(\theta, N), \\ \theta &\sim \text{Uniform}(0, 1). \end{aligned}$$

We integrate out θ to obtain $\Pr(Y) = .0909$ when $Y = 7$ for 10 flips. In fact, $\Pr(Y) = .0909 = 1/11$ for all values of Y for this uniform prior. Figure 3A shows the probability of data, the predictions, for all the outcomes of the ten-flip experiment.

Suppose we wish to compare this general model to a fair-coin model, denoted $\mathcal{M}_B$. The fair-coin model is

$$Y \sim \text{Binomial}(.5, N).$$

The predictions of the model are given by $\binom{N}{Y}(.5)^{10}$ and shown in Figure 3B.

The strength of evidence for each model is given respectively by

$$\frac{\pi(\mathcal{M}_A|Y)}{\pi(\mathcal{M}_A)} = \frac{p(Y|\mathcal{M}_A)}{p(Y)}$$

and

$$\frac{\pi(\mathcal{M}_B|Y)}{\pi(\mathcal{M}_B)} = \frac{p(Y|\mathcal{M}_B)}{p(Y)}.$$

The marginal density of data now is marginal over all considered models. Let M be the class of I models indexed $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_I$. Then,

$$p(Y) = \sum_i p(Y|\mathcal{M}_i)\pi(\mathcal{M}_i).$$

The example helps students understand that the probability (or density) of data can be expressed three ways: conditional on a particular model and parameter value; conditional on a particular model but marginal across all parameters in the space for that model, or marginal across several models. For example, the probability of 4 heads may be conditional on a specific value of θ, say $\theta = .5$ in the general model, and this probability is found from a simple binomial calculation (.205). If also may conditional on the general model but marginalized across all parameters. This calculation involves an integral across simple binomial calculations, and the value is .0909. Finally, the probability can be marginalized across uncertainty in whether the appropriate model is the fair one or the general, the appropriate calculation is the weighted average of .0909 and .205, where the weights reflect $\pi(\mathcal{M}_i)$.

The usefulness of relative strength of evidence for models is now a straightforward extension of the previous development. The relative strength of evidence, the Bayes factor, is the relative predictive probabilities (densities):

$$\frac{\frac{\pi(\mathcal{M}_A|Y)}{\pi(\mathcal{M}_A)}}{\frac{\pi(\mathcal{M}_B|Y)}{\pi(\mathcal{M}_B)}} = \mathrm{BF}_{AB} = \frac{p(Y|\mathcal{M}_A)}{p(Y|\mathcal{M}_B)}. \tag{4}$$

The ratio of these predictions, the Bayes factors, is shown Figure 3C. Had we observed a low number of heads, say $Y = 2$, we could note that the observation is better predicted under model $\mathcal{M}_A$ (with probability .0909) than under model $\mathcal{M}_B$ (with probability .044). The ratio is 2-to-1 in favor of the general model. Conversely, had we observed a moderate number, say $Y = 5$, then the observation is better predicted by the fair-coin model, model

$\mathcal{M}_B$ than by the more general model $\mathcal{M}_A$. The ratio here is 2.7-to-1 in favor of the fair-coin hypothesis.

The definition of $p(Y|\theta)/p(Y)$ as predictive accuracy allows students to easily see one of the oft-touted benefits of Bayes factors: they automatically reward parsimonious models (e.g., Jefferys & Berger, 1991). More parsimonious models are ones that can make specific predictions. The laws of probability require that a model with diffuse data predictions will offer a low probability (or density) for any observed data set; hence, it will be harder to obtain high amounts of evidence for a less parsimonious model unless the more parsimonious model predicts the data even less accurately.

Conclusion

In summary, we find three main advantages to teaching the ratio form of Bayes' theorem. First, students tend to see Bayes' theorem as a way of updating beliefs. This leads to a focus on the updating itself as much as on the resultant posterior. With such a focus, it is easier to show students how prior specification affects updating, the role of the prior in model specification, and the difficulties with improper priors. Second, students learn to reason about the strength of statistical evidence, which may not be a concept they have encountered except informally. Expressing strength of evidence as the degree to which a set of propositions can accurately predict data is particularly intuitive. Finally, students can unite estimation and model comparison as both are seen to flow from the very same form. Students' decision to use model comparison versus parameter estimation can be driven by the question at hand, and not by blanket recommendations to avoid one or the other.

References

Albert, J. H. (1997). Teaching Bayes rule: a data-oriented approach. *The American Statistician*, *51*, 247-253.

Berry, D. A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician*, *51*, 241-246.

Carnap, R. (1962). *Logical foundations of probability.* Chicago: The University of Chicago Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition).* London: Chapman and Hall.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*, 684-704.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* Chichester, United Kingdom: John Wiley & Sons.

Jefferys, W. H., & Berger, J. O. (1991). *Sharpening Ockham's razor on a Bayesian strop.* Technical Report. Retrieved from `http://quasar.as.utexas.edu/papers/ockham.pdf`

Jeffreys, H. (1961). *Theory of probability (3rd edition).* New York: Oxford University Press.

Kruschke, J. K. (2012). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General.*

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, -.

Retrieved from

http://www.sciencedirect.com/science/article/pii/S0022249615000723

Author Note

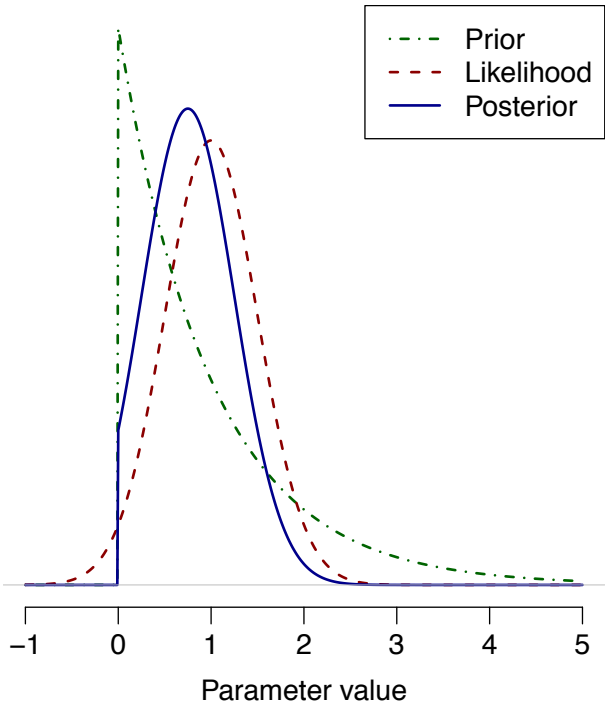
Jeff Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211, rouderj@missouri.edu. This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

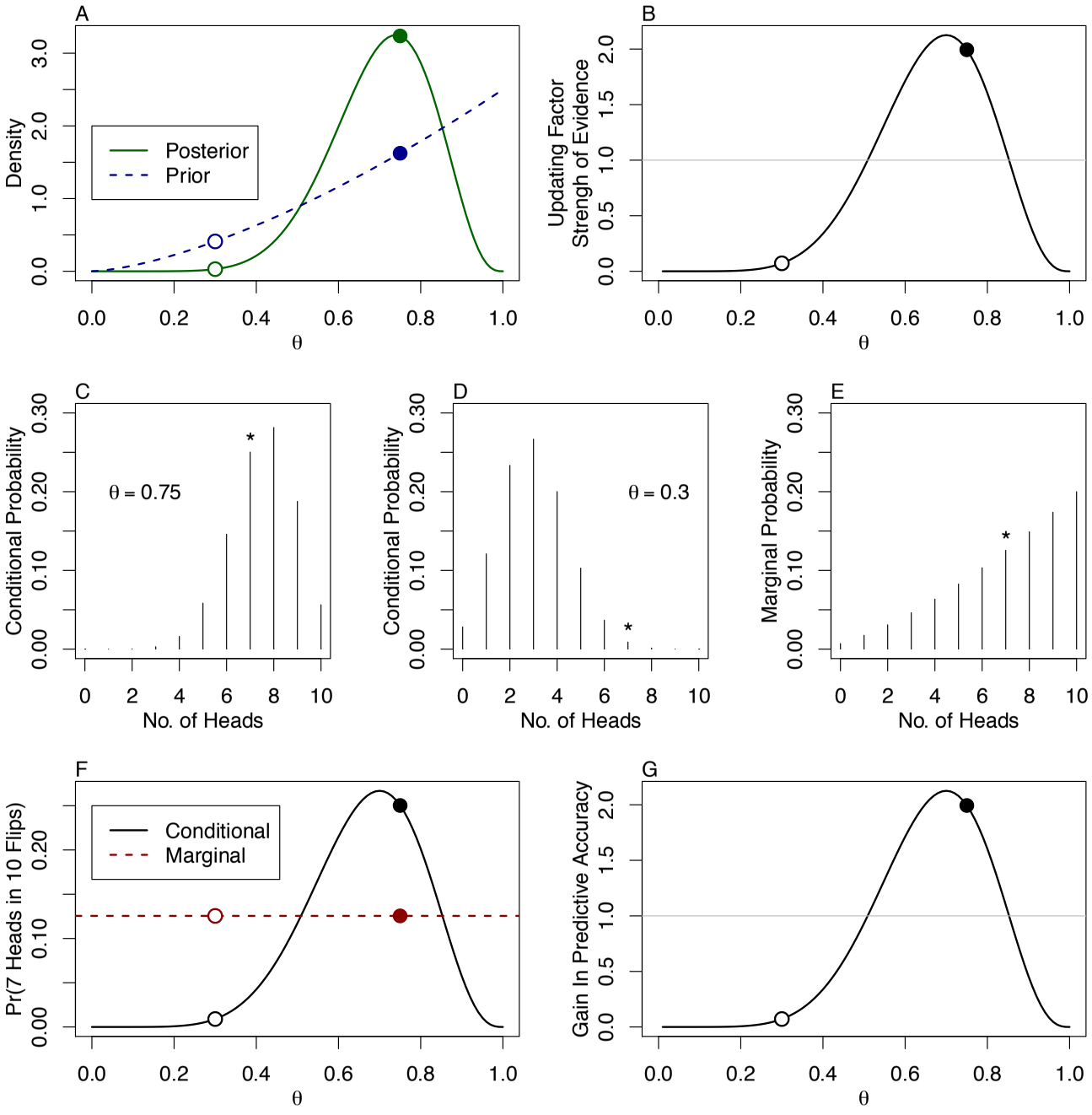
Figure Captions

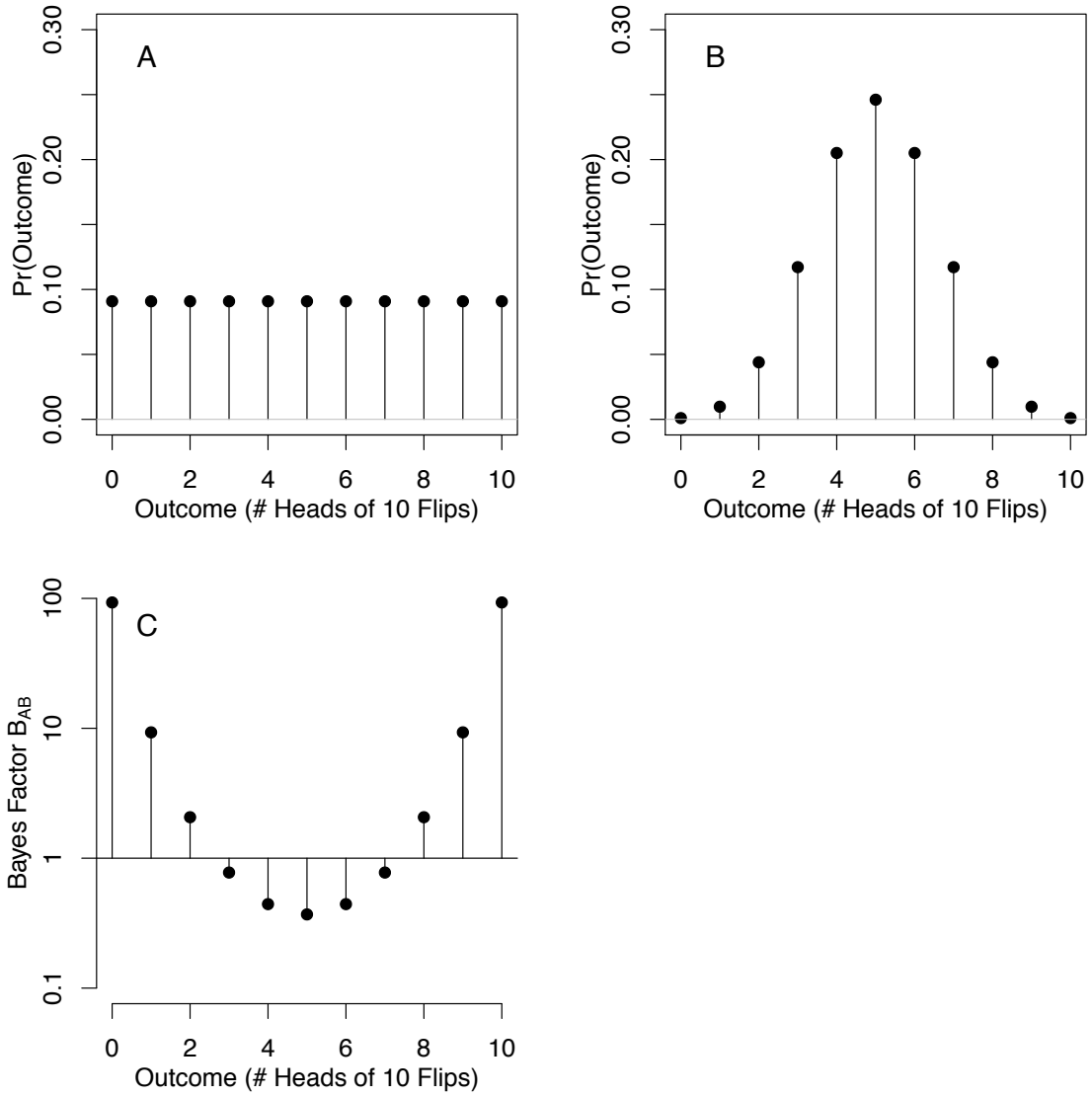
Figure 1. The relationship between the prior, likelihood, and posterior. The relationship is based on proportionality, and values on the y-axis need not be included.

Figure 2. Updating with Bayes' theorem. **A**. Prior and posterior distributions for 7 heads in 10 flips. **B**. The left-hand side of (3), which is the updating factor and may be defined as the strength of evidence from the data for values of θ **C-D**. Probability of outcomes for $\theta = .75$ and $\theta = .30$, respectively. The starred point is the observed value of 7 heads in 10 flips. We find that students best understand these as "predictions" about where data should be observed. **E**. Marginal probability of outcomes with marginalization across θ with respect to the prior. These may be called the marginal predictions. **F**. $P(Y|\theta)$ and $P(Y)$ as a function of θ for 7 heads in 10 flips. **G.** The ratio $P(Y|\theta)/P(Y)$, or the gain in predictive accuracy for values of θ. Bayes rule is a statement of the equality of plots **B** and **G**.

Figure 3. **A**, **B**: The probability of outcomes under the general model, $\mathcal{M}_A$ and under the fair-coin model, $\mathcal{M}_B$, respectively. **C**: The ratio of these probabilities is the Bayes factor between the models.


Prior
Likelihood
Posterior
−1
0
1
2
3
4
5
Parameter value


A
Density
Posterior
Prior
θ
B
Updating Factor
Strengh of Evidence
θ
C
Conditional Probability
θ = 0.75
No. of Heads
D
Conditional Probability
θ = 0.3
No. of Heads
E
Marginal Probability
No. of Heads
F
Pr(7 Heads in 10 Flips)
Conditional
Marginal
θ
G
Gain In Predictive Accuracy
θ


A
B
C
Pr(Outcome)
0.00
0.10
0.20
0.30
Bayes Factor B_{AB}
0.1
1
10
100
0
2
4
6
8
10
Outcome (# Heads of 10 Flips)