October 15, 2015

What is a Bayesian p-value?

1. The definition that I use the most is as follows. Let β be a univariate continuous *parameter* of interest, taking values on the real number line.

   The Bayesian p-value is the probability

   $$P(\beta > 0|\text{data}) \tag{1}$$

   or

   $$P(\beta < 0|\text{data}). \tag{2}$$

   Often we report which ever of (1) or (2) is smaller. These are defined for regression coefficients, for example, from linear Poisson, or logistic regression. These definitions are close to the frequentist one-sided p-value. I routinely calculate (1) (or 2) in WinBUGS or jags when my model has regression coefficients.

   When these p-values are close to zero (one) that means that the sign of the coefficient is well determined a posteriori. The frequentist interpretation is that β is significantly different from zero.

2. The frequentist p-value usually would double the calculation (1) or (2), which ever is smaller. And given the calculation, call it $p = P(\beta > 0|\text{data})$ from (1), if I'm in excel, I calculate

   $$2 * \min(p, 1 - p) \tag{3}$$

   and report that as a two-sided (classical) p-value.

3. We don't calculate (1) for the intercept β_0 in a regression model. Whether an intercept is positive or negative is not usually of great scientific interest. Suppose the outcomes y_i are all non-negative and we center the covariates x_{ij}. Then the intercept β_0 is the population mean of the y_i's and we know that β_0 is positive with probability one.

   If we now stop centering the x_i, the intercept changes, yet the science doesn't change in our analysis, yet the meaning of $\beta_0 = 0$ changes from

when we were centering the x_i. The meaning of $\beta_0 = 0$ depends on the definitions of the covariates in the model, that is, the meaning of $\beta_0 = 0$ depends on the centering, scaling, and units of all the predictors and the omission or inclusion of each individual predictor. Thus it is unlikely that $\beta_0 = 0$ or $\beta_0 > 0$ are important hypotheses to consider.

4. A precision parameter τ takes values on the positive real line $[0, \infty)$. Then logtau $= \log(\tau)$ takes values on the real number line, but the calculation $P(\text{logtau} < 0|\text{data})$ is not of interest, and is not a Bayesian p-value. Or maybe it is, but a silly one, as we're not usually interested in testing the hypothesis that a precision parameter is equal to 1, nor are we interested in calculating $P(\tau > 1|Y)$.

   In a location scale model, observations $y_i|\mu, \tau$ have density $\tau^{1/2} f(\tau^{1/2}(y_i - \mu))$ where μ is a location parameter and $\tau^{-1/2}$ is a scale parameter. The precision parameter τ has units that are the same as y_i^{-2}. By changing the units of y_i such as from meters to kilometers or from Fahrenheit to centigrade, we change the units of τ, and the meaning of 1 changes with the units. Hence it is implausible that testing $H_0 : \tau = 1$ is of much interest.

5. Another variation. Suppose my model has two similar parameters ω_1 and ω_2. I may be interested in identifying which parameter is bigger. The calculation

$$P(\omega_1 > \omega_2|\text{data}) \tag{4}$$

   is the posterior probability of the hypothesis $H_0 : \omega_1 > \omega_2$, and this probability is also a Bayesian p-value. This is again similar to a frequentist one-sided p-value.

6. A posterior predictive p-value (PPPV) is a quantity that Rubin (1984) and Gelman, Meng and Stern (1996) have suggested that we calculate.

   Let Y be the data that you have observed. Suppose our probability model is

$$f(Y|\theta)p(\theta) \tag{5}$$

   sampling density times prior. Calculate the posterior density

$$p(\theta|Y). \tag{6}$$

Generate new data Z by sampling Z from

$$f(Z|\theta)p(\theta|Y). \tag{7}$$

You can do this by first sampling $\theta*$ from $p(\theta|Y)$ then sampling $f(Z|\theta*)$. This f in $f(Z|\theta)$ or $f(Z|theta*)$ is the same f as was the sampling density for the original data Y. The hypothetical new data Z has a posterior predictive density $f(Z|Y)$.

7. Now, Rubin's original posterior predictive p-value was to calculate some *univariate* statistic $g(Y)$ from the data and calculate

$$P(g(Z) > g(Y)|Y) \tag{8}$$

8. Gelman, Meng and Stern's innovation was to let the function g be a function both of Y (or Z) and θ as well. The calculation in (9) is the same,

$$P(g(Z,\theta) > g(Y,\theta)|Y) \tag{9}$$

The cleverness is in picking a good choice of function $g(\cdot)$ so that is tells you something interesting about your data.

9. Comment. The development does not depend on Y or g or θ or β being discrete or continuous, but does rely on β in (1) or the function g in (9) being univariate.

10. Consider logistic regression, where y_i is either 0 or 1, for $i = 1, \ldots, n$. Let

$$\pi_i = P(y_i = 1|\theta) \tag{10}$$

Then π_i is a parameter (of function of the parameters theta) and it has a posterior. The posterior mean of π_i, $E[\pi_i|Y]$ is an estimate of your "surprise" at seeing y_i if $y_i = 1$. Alternatively, $1 - E[\pi_i|Y]$ is the surprise of seeing $y_i = 0$. This can be used as an outlier statistic for logistic regression. The discreteness of the data y_i is not a problem.

This outlier statistic $\mathcal{O}_i = y_i - \pi_i$ has simple properties, in that, given π_i, it only takes on two values, $0 - \pi_i$ and $1 - \pi_i$. We could try plotting the outlier statistic $\mathcal{O}_i$ to try to create a residual plot, perhaps $\mathcal{O}_i$ against x_{ij} for some covariate x_{ij}. But the funny properties of $\mathcal{O}_i$ mean that this plot is likely to be difficult to interpret.

11. The posterior probability that some function h of the parameters θ and/or data Y and/or predictions Z is in a set S is a probability

$$P(h(\theta, Y, Z) \text{ is in } S|Y). \tag{11}$$

    If this probability is large then there is no surprise. If this probability is small, then your model (or perhaps your computations) may have an error. This interpretation depends on having independent information that the set S is likely to occur a priori.

12. However you define small.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4), 733–760.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics, 12(4), 1151–1172.