Neil Klenk

The Problem with the P-Value

**Problem/Importance of the P-value**

The P-value is currently the defacto value sought after in determining if a null hypothesis can be rejected or if it is failed to be rejected. With the pervasiveness of the P-value’s use in academic and industry research, these values have an incredible amount of influence on the future of everyday society. The problem here is that many scientists are not also statisticians, and therefore do not understand that the P-value is not as robust or objective as they assume it to be. It was never meant to be a definitive test, only a way of telling if something was worthy of a second look or not. Depending on the initial plausibility of the results, the same p-value can result in many different odds for there having been any chance of real effect.1

Another problem lies in the perception of the P-value. It is not irregular for people to see a P-value of 0.01 and assume that this means that there is a 1% chance of that result being false. However, this is not the case. It is only able to summarize data assuming a specific null hypothesis. Not work backwards to make a statement about the true underlying reality. All the P-value represents in this case is that there is strong evidence that there is some effect, not that the effect itself is strong. This results in a situation where the more implausible the hypothesis, the more likely that an exciting finding is a false alarm. This characteristic has led to the practice of p-hacking, trying multiple hypotheses until you get the desired result.1

**Summary of the Book**

The objective of this project was to explore the alternative ways to determine statistical significance of findings, given the fallacies of the p-value. The potential for Bayesian inference to replace the p-value was investigated, and so *A First Course in Bayesian Statistical methods* by Peter D. Hoff was read. Bayesian inference differs significantly from the more traditional frequentist approach. The basis behind Bayesian inference is that it is a method of inductive learning. The individual performing the analysis already has a prior belief, then that prior is updated with the likelihoods present in the dataset. If new data becomes available then those likelihoods can be used to update the prior into a new, and more accurate posterior. The frequentist approach is focused on the probability of an event as proportion of outcomes. Any experiment can be thought of an infinite sequence of possible repetitions of that same experiment, where the results of each experiment are independent of one another.2

Bayes rule (fig. 1) describes the belief that θ is the true value having observed dataset y. The prior distribution is a description of the belief that θ is the true population characteristics before the evidence in dataset y can be taken into consideration. The sampling model Ƥ(y|θ) is a description of the belief that y would be the outcome of the study, assuming our prior θ is true. The posterior distribution Ƥ(θ|y) based on the marginal prior distribution (the prior belief) Ƥ(θ) and the sampling model Ƥ(y|θ) divided by the normalizing constant which is equivalent to Ƥ(y) after integration.

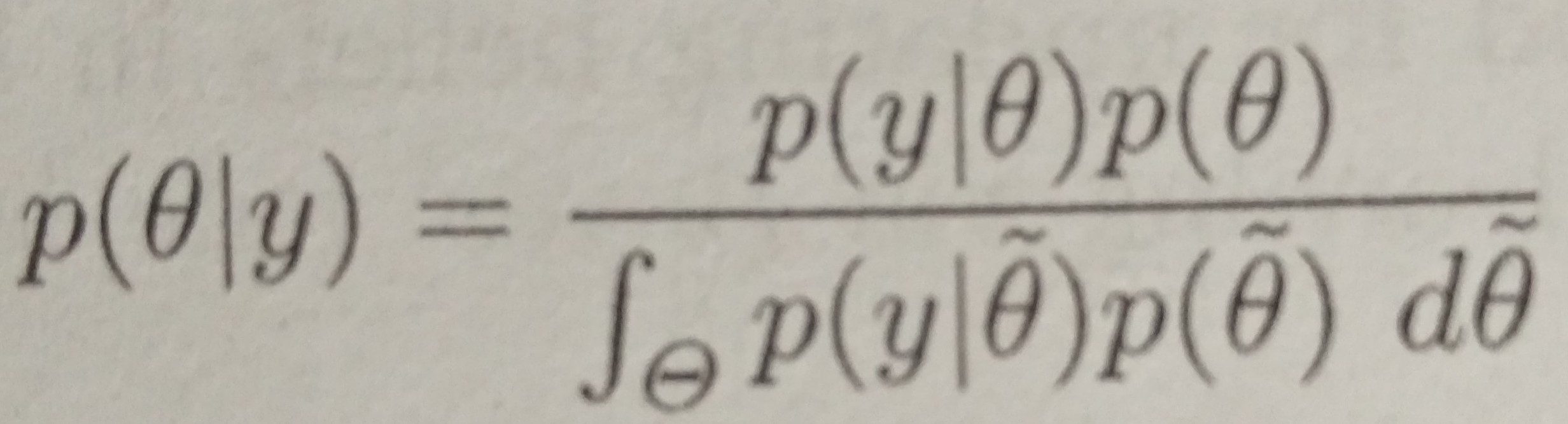


Fig 1. Bayes Rule2

One parameter models in Bayesian inference include the beta-binomial model and the gamma-poisson model, which are both instances of one-parameter exponential family models. The beta-binomial model has a beta(a,b) prior, binomial likelihood function, and a resulting beta(a+y, b+n-y) posterior where n is the sample size, y is the number of successes, and a and b are the prior information. These are characterized by success or fail type problems (eg. coin flipping). The gamma-poisson model is used for determining the mean of a random variable. It is described by the gamma(a,b) prior, poisson likelihood function, and a resulting gamma(a+ΣYi, b+n). The summation of Y is the number of successes, n is the sample size, and a and b the prior information.2

In both of these methods, it can be seen that as the sample size increases compared to the size of the prior a and b values, the posterior will be affected to a higher degree. This is throughout the Bayesian methods, and takes away a lot of the argument that the prior beliefs affect the credibility of the analysis. If a large enough sample is taken, the prior will be swamped even if it was bias to begin with. Bias can also be beneficial, especially with smaller samples (e.g. if a similar study has already been done, you can expect similar results).

Monte Carlo approximations are very powerful tools in Bayesian analysis because almost any aspect of a posterior distribution can be arbitrarily approximated using a Monte Carlo sample that is large enough. These approximations are based on random sampling of a given set of parameters (eg. gamma(a,b)), and are primarily performed computationally. They can also be used to generate both posterior predictive and prior predictive distributions, which can then be sampled from. A prior predictive distribution is a distribution which integrates over unknown parameters, but is not conditional on observed data. This differs from posterior predictive distributions which conditions on an observed dataset. Posterior predictive model checking gave rise to the notion of the posterior predictive p-value which is discussed later.2

The normal model is based off of the normal curve, the symmetrical curve with %99.7 of its data lying within 3 standard deviations of the mean, characterized by an independent sampling mean and sample variance. The importance of the normal curve comes from the central limit theorem, which states that “under very good conditions, the sum(or mean) of a set of random variables is approximately normally distributed”. Even if the data itself is not normal, the normal model can still be taken advantage of. Instead of the sampling distribution of the data points themselves being used, the sampling distribution of the sample mean will be used. The idea is that while the data itself may not be normal, the distribution of the means should be at least close to normal.2

Determining if the observed difference in group means is significant, relative to the sampling variability, is traditionally the job of the frequentist t-statistic. The t-statistic results in the two means being treated as either completely different, or completely the same. A better way includes using the information from one group to help estimate the mean in the other. Hierarchal models utilize hierarchal data (in which there is a hierarchy of nested populations). The Hieratical normal model describes the extent of the similarity of means and variances across several populations, and has normally distributed within-group and between-group sampling models. One of the advantages of hierarchal modeling is that information can be shared across groups. This results in shrinkage, where the means and variances of groups with low samples get pulled toward the overall average mean and variance. The advantage here is that now, smaller samples are less affected by the large potential variance due to their small sample size. Means and variances can also be modeled hierarchally, and sampled with the Gibbs sampler. Gibbs sampling generates posterior samples by iteratively going through each variable (or a block of variables) to sample from its conditional distribution while keeping all other values at their current value.2

**Alternative to the P-value**

Bayesian methods are a fitting alternative to the frequentist p-value. These methods much better at describing probability as the plausibility of an outcome, not just the frequency of it. The posterior predictive P-value is a particularly viable alternative to the frequentist P-value. Posterior predictive P-values are a result of the evaluation of goodness-of-fit models that were dependent on both parameters and predictive data. A Monte-Carlo simulation of the posterior predictive distribution of the empirical odds of the value in question will show what percentage of simulated datasets that were equal to or greater than the initial observed data. Significantly low percentages indicate that the model is flawed, and another needs to be used to determine the true probability distribution of the sampled values.2

Bibliography

1Nuzzo, Regina. "Scientific Method: Statistical Errors." *Nature* 506 (2014): 150-52. *Nature International Weekly Journal of Science*. Nature, 12 Feb. 2014. Web. 04 Apr. 2016.

2Hoff, Peter D. *A First Course in Bayesian Statistical Methods*. New York: Springer, 2009. Print.