Title: Bayes' Theorem in the Twenty-First Century

Author: Bradley Efron[1]*

Affiliations:

[1]Stanford University.

*Correspondence to: brad@stat.stanford.edu

One Sentence Summary: Bayes' Theorem, now celebrating its 250th birthday, is playing an increasingly prominent role in statistical applications but, for reasons both good and bad, it remains controversial among statisticians.

Main Text: "Controversial theorem" sounds like an oxymoron, but Bayes' Rule has played this part for two and a half centuries. Twice it has soared to scientific celebrity, twice it has crashed, and it is currently enjoying another boom. The Theorem itself is a landmark of logical reasoning and the first serious triumph of statistical inference, yet is still treated with suspicion by a majority of statisticians. There are reasons to believe in the staying power of its current popularity, but also some worrisome signs of trouble ahead.

Here is a simple but genuine example of the Rule in action (1). A physicist couple I know learned, from sonograms, that they were due to be parents of twin boys. "What is the probability our twins will be Identical rather than Fraternal?" was their question. There are two pieces of relevant evidence: only one-third of twins are Identical; on the other hand, Identicals are twice as likely to yield twin boy sonograms, since they are always same-sex while Fraternals are 50/50. Putting this together, the Rule correctly concludes that in this case the two pieces balance out, and that the odds of Identical are even. Note: the twins were fraternal.

Bayes' Theorem is an algorithm for combining prior experience ("one-third Identicals") with current evidence (the sonogram). Followers of Nate Silver's FiveThirtyEight column got to see it in spectacular form during the presidential campaign: the algorithm updated prior poll results with new data on a daily basis, nailing the actual vote in all 50 states. "Statisticians beat pundits" was the verdict in the press. See (2) for a nice discussion of Silver, Bayes, and predictions in general.

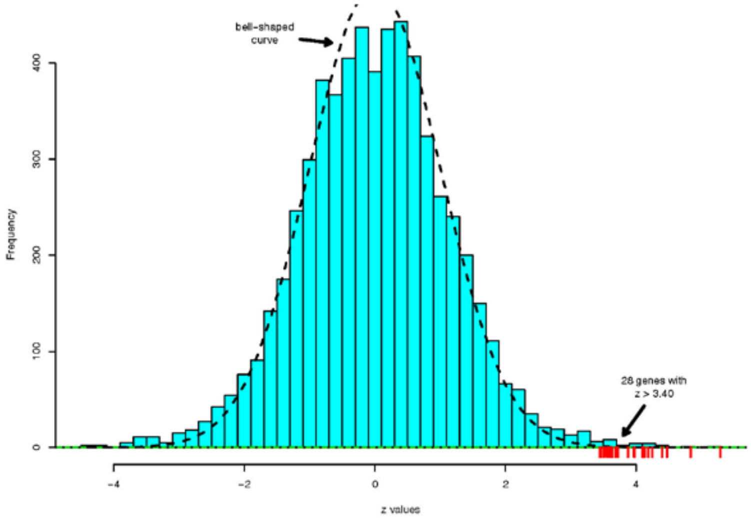
Bayes' 1763 paper was an impeccable exercise in probability theory. The trouble and the subsequent busts came from overenthusiastic application of the Theorem in the absence of genuine prior information, with Laplace as a prime violator. Suppose that in the twins example we lacked the "one-third Identicals" prior knowledge. Laplace would have casually assumed a uniform distribution between zero and one for the unknown prior probability of Identical (yielding 2/3 rather than 1/2 as the answer to the physicists' question.) In modern parlance, Laplace would be trying to assign an "uninformative prior," one having only neutral effects on the output of Bayes' rule (3).

Frequentism, the dominant statistical paradigm over the past hundred years, rejects the use of uninformative priors, and in fact does away with prior distributions entirely (1). In place of past experience, frequentism considers future behavior: an optimal estimator is one that performs best in hypothetical repetitions of the current experiment. The resulting gain in scientific objectivity has carried the day, though at a price in the coherent integration of evidence from different sources, as in the FiveThirtyEight example.

The Bayesian-frequentist argument, unlike most philosophical disputes, has immediate practical consequences: after a seven-year trial on human subjects, a research team announces that drug A has proved better than drug B at the .05 significance level. The team's leader, asked why the trial took so long, replies "That was the first time the results reached the .05 level." FDA regulators reject the team's submission, on the frequentist grounds that interim tests of the data could raise the false alarm rate to (say) 15% from the claimed 5%.

A Bayesian FDA (there isn't one) would be more forgiving. The Bayesian posterior probability of drug A's superiority depends only on its final evaluation, not whether there *might* have been earlier decisions. Similarly, Bayesian evidence that "caffeine causes strokes" isn't diminished if the investigators tried but failed to implicate salt, sugar, smoking, or twenty other suspects. All of this is a corollary of Bayes theorem, convenient but potentially dangerous in practice, especially when using prior distributions not firmly grounded in past experience.

I recently completed my term as editor of an applied statistics journal. Maybe 25% of the papers employed Bayes' theorem. It worried me that almost all of these were based on uninformative priors, reflecting the fact that most cutting-edge science doesn't enjoy FiveThirtyEight-level background information. Are we in for another Bayesian bust? Arguing against this is a change in our statistical environment: modern scientific equipment pumps out results in fire hose quantities, producing enormous data sets bearing on complicated webs of interrelated questions. In this new scientific era, the ability of Bayesian statistics to connect disparate inferences counts heavily in its favor.



An example will help here. The figure concerns a microarray prostate cancer study (4): 102 men, 52 patients and 50 healthy controls, have each had his genetic activity measured for 6033 genes. The investigators, of course, are hoping to find genes expressed differently in patients as opposed to controls. To this end a test statistic *z* has been calculated for each gene, which will have a standard normal ("bell-shaped") distribution in the Null case of no patient/control difference, but will give bigger values for genes expressed more intensely in patients.

The histogram of the 6033 *z*-values doesn't look much different than the bell-shaped curve (dashed line) that would apply if all the genes were Null, but there is a suggestion of interesting non-null genes in its heavy right tail. We have to be careful though: with 6033 cases to consider at once, a few of the *z*'s are bound to look big even under the Null hypothesis.

False Discovery Rates (5) are a recent development that takes multiple testing into account. Here it implies that the 28 genes whose *z*-values exceed 3.40, indicated by dashes in the figure, are indeed interesting, with the expected proportion of False Discoveries among them being less than 10%. This is a frequentist 10%: how many mistakes we would average using the algorithm in future studies. The determining fact here is that if indeed all the genes were Null, we would expect only 2.8 *z*-values exceeding 3.40, that is, only 10% of the actual number observed.

This brings us back to Bayes. Another interpretation of the FDR algorithm is that the Bayesian probability of nullness given a *z*-value exceeding 3.40 is 10%. What prior evidence are we using? None, as it turns out! With 6033 parallel situations at hand, we can effectively estimate the relevant prior from the data itself. "Empirical Bayes" is the name for this sort of statistical jujitsu, suggesting, correctly, a fusion of frequentist and Bayesian reasoning.

Empirical Bayes is an exciting new statistical idea, well-suited to modern scientific technology, saying that experiments involving large numbers of parallel situations carry within them their own prior distribution. The idea was coined in the 1950s (6), but real developmental interest awaited the data sets of the Twenty-First Century.

I wish I could report that this resolves the 250-year controversy, and now it is safe to always employ Bayes' Theorem. Sorry. My own practice is to certainly use Bayesian analysis in the presence of genuine prior information; to use empirical Bayes methods in the parallel cases situation; and otherwise to be cautious when invoking uninformative priors. In the last case, Bayesian calculations cannot be uncritically accepted, and should be checked by other methods, which usually means frequentistically.

References and Notes:

1. Efron, B. (2013). A 250-year argument: Belief, behavior, and the bootstrap. *Bull. Amer. Math. Soc.* **50**, 129−146. doi: 10.1090/S0273-0979-2012-01374-5
2. Wang, S. and Campbell, B. (2013). Mr. Bayes goes to Washington. *Science* **339**, 758−759.
3. Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1**, 385−402 (electronic).
4. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203−209. doi: 10.1016/S1535-6108(02)00030-2.
5. Benjamini, Y. and Hochberg, Y. (1995).Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289−300.
6. Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954−1955, Vol. I,* 157−163. University of California Press, Berkeley and Los Angeles.