

Bayes' Theorem: Why Does it Matter?

Charles Kankelborg

Revised March, 2022

1 Introduction

Once upon a time, there were frequentist statistics, practiced mainly by old white men. That's all anybody had in those dark times. Then the prophet Bayes had a revelation. Bayes' theorem portended a revolution, bringing light and understanding to statistical inference. The Bayesians, disciples of their prophet, waged war against the frequentists. Although the frequentists were pretty dense, a few saw the light and converted to the new faith. Most simply aged out and retired. Actually, none of that is true. A short historical introduction appears in [the Wikipedia article on Thomas Bayes](#). Statistical inference under either a "frequentist" or "Bayesian" name can be done well or poorly. But even frequentist statistical work will be less prone to error if we understand the implications of Bayes' theorem.

What follows is a completely unoriginal introduction to Bayes' theorem using the [Prosecutor's Fallacy](#)¹ to show why there is more to life than p -values.

2 p -Values

Suppose that we are doing a correlation study. We have N ordered pairs of data, (x_i, y_i) . From the data, we calculate a correlation coefficient, τ . Perhaps we found τ very close to its maximum value of 1, which makes us think x and y are correlated. How can we quantify our confidence in this correlation?

We assign the name S to the statistical evidence actually obtained; in other words, S stands for the proposition that the correlation coefficient τ is greater than or equal to the value we measured. The

null hypothesis, H , is that the variables x and y are uncorrelated in the parent population from which my data were drawn. The virtue of this mode of thought is that H is a particularly simple assumption, so starting from this assumption is straightforward to calculate the probability of the evidence, S . That probability is called

$$\alpha = \Pr(S|H),$$

which we read as " α equals the probability of S given H ". Another name for α is the " p -value". We also use the word "confidence" to describe $1 - \alpha$. Hence, if the p -value is 0.001, we might say that "we reject the null hypothesis with 99.9% confidence." In other words, since the null hypothesis H is unlikely to have given rise to the evidence S that we actually observed, we cite this as justification to discard hypothesis H . We *do* accept the *negation* of the null hypothesis. In plain language, x and y are correlated. We celebrate, and we publish. That is representative of the frequentist approach. For years, this is all I thought about in the context of statistical significance, and I can't honestly say that it led me astray in any important way.

3 Bayes' Theorem

In the previous section, we made the tacit assumption that if the probability of the evidence S given the null hypothesis H is small, then the likelihood that H is true given the evidence is also small. Very roughly, we assumed

$$\Pr(H|S) \approx \Pr(S|H).$$

¹https://en.wikipedia.org/wiki/Prosecutor's_fallacy

Bayes' theorem is simply an equation that gives a rigorous relationship between the above two quantities. When "Bayesians" are up in arms, they assert that "frequentists" make an equation out of the above assumption. Rather than calling people names, let's derive the true relationship. Then in the next section, we'll see just how far south things can go when the above assumption is made in a cavalier or thoughtless manner.

Consider the *joint probability distribution* $\Pr(S, H)$. This is the probability that two propositions, S and H , are both simultaneously true. In the case we have been considering, the proposition S refers to our evidence, which consists of the fact that some statistical quantity is at a specified (measured) value or greater. Proposition H refers to a null hypothesis, crafted to make $\Pr(S|H)$ easy to calculate. The joint probability is

$$\Pr(S, H) = \Pr(S|H) \Pr(H).$$

In words, the probability that both S and H are true is equal to the probability of S given H times the probability of H by itself. $\Pr(H)$ is called the *prior*, which is short for "the *a priori* probability, ignoring the evidence S , that H is true." I take the above equation as obvious, but perhaps it is more profound than I think. Similarly,

$$\Pr(H, S) = \Pr(H|S) \Pr(S).$$

The quantity $\Pr(S)$ is, as you might imagine, the probability of the evidence S regardless of whether the null hypothesis H is true. Now, $\Pr(H, S)$ and $\Pr(S, H)$ mean the same thing by definition, so

$$\Pr(H|S) = \frac{\Pr(S|H) \Pr(H)}{\Pr(S)}.$$

This is called **Bayes' theorem**.

4 Prosecutor's Fallacy

Having obtained the exact relationship between $\Pr(S|H)$ and $\Pr(H|S)$, we wonder: how different can they be?

A prosecutor introduces forensic evidence from the scene of a murder that matches the accused with 99.9% confidence.² When asked by the defense exactly what this confidence means, the forensic pathologist describes it as one minus a p -value:

$$\Pr(S|H) = 0.001,$$

where S is the match of the forensic evidence, and H is the null hypothesis: that the accused is *not* the killer. This sounds very damning, especially after the prosecutor has told the jury every terrible detail of the crime. For the sake of argument, let us stipulate that the evidence was fresh, almost certainly associated with the perpetrator. Remembering that null hypotheses are formulated to be very straightforward, then the scenario of H is that the accused is an individual randomly chosen from the population at large, who had nothing to do with the crime. Thus, $\Pr(S|H)$ means the probability that any random person would just happen to match the forensic evidence. What does $\Pr(S|H)$ tell us about $\Pr(H|S)$, which our stipulation equates with the likelihood of guilt?

Suppose the crime occurred in a city of one million people. How does Bayes' theorem illuminate the problem? The *a priori* chance that a particular person is the perpetrator is one in a million:

$$\Pr(\bar{H}) = 10^{-6}.$$

The chance that the opposite is true, which we identify as the prior for the null hypothesis H , is

$$\Pr(H) = 1 - 10^{-6} \approx 1.$$

If we make no assumptions about H , then estimating the probability that a randomly chosen person matches the forensic evidence requires marginalizing over two scenarios: a 0.001 probability of a random match, and a 10^{-6} probability that the person happens to be the perpetrator:

$$\begin{aligned} \Pr(S) &= \Pr(S|H) \Pr(H) + \Pr(S|\bar{H}) \Pr(\bar{H}) \\ &= (0.001 \times 1) + (1 \times 10^{-6}) \\ &= 0.001001. \end{aligned}$$

²Hair? Fingerprints? I don't know. You've watched CSI; use your imagination.

When we put all the factors together, we find

$$\Pr(H|S) = \frac{0.001 \times 1}{0.001001} = 0.999.$$

In other words, unless the prosecutor has additional evidence connecting the accused to the crime, there is a 99.9% chance that the accused is innocent. Carefully accounting for all the possibilities, Bayes' theorem has taken us all the way from 99.9% probability of guilt to 99.9% probability of innocence.

You don't have to be a Bayesian to see the fallacy: There are $0.001 \times 10^6 = 1,000$ people in this city who match the forensic evidence. 999 of those 1,000 people are not the murderer. If the forensic evidence is all we have to go on, then there is a 99.9% chance that the crime was committed by one of the 999 people who were not charged with murder. Put this way, it sounds painfully obvious, but there is a reason this is called the Prosecutor's Fallacy. Plenty of people have gone to prison (or worse) on comparably flimsy evidence.

A word of caution is in order here. If the detectives went wandering around the city collecting samples from random people until they found a suspect, or if they simply picked the first match they found in a massive database, then the above calculation holds and the prosecutor's case is baseless. However, if the prosecutor has shown that only three people were in the building at the time, and one of them is a perfect match, that is a very different case indeed. If we stipulate that one of the three is the perpetrator, we would have $\Pr(H) = 0.667$ and $\Pr(S) = 0.001 + 0.333 = 0.334$. The result then would be $\Pr(H|S) = 0.002$. This is twice the raw p -value presented by the prosecutor, but it is still small. We are left with a 99.8% chance of guilt. Of course, a responsible investigator should also test the two other possible suspects! The lesson of Bayes' theorem is that context matters a lot, and we should carefully take into account *all* the information at our disposal.

5 Conclusion

We started in section 2 with a simple argument to interpret one minus the p -value as a confidence that

the null hypothesis can be rejected. In section 4, we reviewed the classic Prosecutor's Fallacy, in which the p -value by itself could send the wrong message.

In today's parlance, people call the quoting of p -values "frequentist", but the content of the relevant probability theory has not changed much since the 1763 publication of Thomas Bayes' eponymous theorem.³ The slur "frequentist" derives from "frequency distribution," a longstanding term of art for all the quantities in this essay that begin with Pr. In my view, the prosecutor's reasoning is not frequentist, but *sloppy*. Call yourself a Bayesian if you want to be cool, or a frequentist if you prefer to be contrarian. More importantly, think carefully before making assertions based on the p -value. If you have no useful prior knowledge ($\Pr(H)$), then the p -value speaks for itself. If not, make a credible attempt to work out what your evidence *really* implies.

Suppose you have performed a study with a very low p -value, but accounting for prior knowledge puts $\Pr(H|S)$ at an unremarkable 15%. What should you do? If your evidence has perceptibly moved the needle compared to previous work, then please, for the love of Bayes, publish. Do not expect to convince anyone to reject the received wisdom on your account, but the community deserves to know that H has been pushed into ambiguous territory. Science is harmed by a systematic bias against publishing contrary results.

³https://en.wikipedia.org/wiki/Bayes%27_theorem#History