Classical Frequentist Statistics: Point
by Ernest B. Hook

Editor's Note: The Spring 2025 issue of Academic Questions *contained a review by William M. Briggs of Aubrey Clayton's* Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science. *In the review, Mr. Briggs, backing Clayton's thesis, suggested that social scientists discard "frequentist" statistics, which in his view misinterpreted "the probability of the data as if it were the probability of the hypothesis." In response to Briggs' review, we present Dr. Ernest B. Hook's defense of frequentist statistics and his rejection of Briggs' call for their discontinuance. Following this essay, Mr. Briggs responds to Hook.*

All statistics are divided into two schools. The "frequentist," classical approach, taught in introductory courses, defines probability as the relative frequency of an event achieved after many repetitions. The other school, the Bayesian, named after its progenitor, the eighteenth century English minister Thomas Bayes, regards probability as a measure of degree of belief in some hypothesis.

Most statisticians use either approach, depending on the available data and ease of application of one method or the other. A subgroup however, will use only Bayesian methods, and some of these will insist that frequentists are irretrievably doomed to errors and meaninglessness.

Among the latter is William M. Briggs, who recently offered ostensible examples of the uselessness of classical statistics ("Let Go Your Wee P!" *AQ,* Spring 2025). He cites a patient who tests positive for "cavortitis," with a test known to be positive in 95 percent of those with the disease, and negative in 99 percent of those without it. He claims in essence that no one using "classical statistical procedures" can tell the patient the chance he or she is affected. But he omits a pertinent variable. The chance depends not just on the cited values of the sensitivity and specificity of the test, but *also* on the prevalence rate of the disease in the population, a value not provided. If the disease is common—say it affects 10 percent of the population, the chance, or probability if you will, of having the disease with this test if one tests positive is about 90 percent. If the disease is rarer, say 1 in 1,000, then the chance is only about 9 percent.

Briggs' example, provided in a review of Aubrey Clayton's *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science* (2021), which endorses a similar viewpoint, hardly provides grounds for rejecting classical statistical procedures. Yet motivated by this and other examples he informs us that he, as well as Clayton, have ceased teaching and using them. But as my counterexample should caution: reader beware!

'Bernoulli's Fallacy'

Briggs states something termed "Bernoulli's Fallacy" is at the heart of the failure of classical statistics. This results from conflating the probability of a hypothesis with the probability of the data offered in its' support. Unfortunately, anyone who is not a statistician is unlikely to understand what this means.

Let me try to explain what he is getting at. Suppose someone publishes data on an experiment involving extra-sensory perception (ESP) and reports the probability of a chance result as less than one percent. This implies the probability is over 99 percent that the trend in the observed *data* or a stronger one did not occur by chance alone. Does this mean that the probability of the associated *hypothesis*, e.g. that the subjects employed ESP, is over 99 percent? Of course not. There are many other hypotheses that explain the data, e.g. secret direct communication of the subjects of which the author was unaware, selective publication of one of many chance results, data manipulation, etc. These and similar grounds may explain a failure by others to replicate the observation (see below). The data *are consistent* with the hypothesis. But they do not *prove* it. Certainly, one may misinterpret the implications of the calculated probability value, but that does not mean the method itself was erroneous.

The Replication 'Crisis'

As Briggs notes, many have claimed that, at least in some fields there has been a perceived "replication crisis," a failure by many workers to replicate observations reported by earlier authors. Briggs claims that the use of classical frequentist statistical methods accounts for this failure of replication. It is an "incoherent methodology," he states. He implies that if we abandon classical statistics and embrace its alternative, the problem should diminish if not disappear.

There are, however, many reasons why an investigator may not confirm another's finding of a "statistically significant" association which have nothing to do with the use of classical, frequentist statistics.

First, s/he may not have replicated the conditions precisely. At least one unrecognized contributory variable may differ. This has been well documented multiple times. For example, one British lab could not confirm an American biochemist's report of an enzymatic reaction because the ambient room temperatures in the two laboratories differed.

Secondly, by chance alone rare events do occur, which will not be observed on replication.

Thirdly, professional success depends upon publication, and publication often depends on finding "statistically significant" results. Investigators may cherry pick data they analyze and report on, believing perhaps innocently, that they are justified because the excluded result is at odds with the general trend. Years ago, a colleague in a first draft of a report on a project omitted an outlier in data, strongly differing from the trends in the rest. It was only with difficulty I could convince my coworker that we should include the outlier, which indeed diminished the strength of our published findings.[1]

Fourthly, is a well known classical statistical principle termed "regression to the mean." One can conceptualize this principle by recognizing that all phenomena are affected by *both* chance and "deterministic" events. Imagine, an extreme experience, for example, an exquisite meal at a restaurant. If the chef is always meticulous in following a recipe and selecting fresh ingredients, then you are likely to replicate the experience when you return. But if the chef has been sloppy but lucky on your first visit, you are much less likely to have the same experience again. Chance contributes to all outcomes, and the more extreme the outcome, the more likely chance has contributed and replication will yield a less extreme outcome.

Fifthly, is the issue of unadjusted multiple comparisons, resulting from what may be termed "data dredging." This is similar to the "Texas sharpshooter fallacy," the latter named after a man who fires a gun at the wall of a barn, paints a bull's-eye around the bullet holes, and claims he has demonstrated he is a sharpshooter.

Lastly, is the problem of reverse causation. For example, observational studies may find healthy people tend to take multivitamins and conclude taking vitamins makes them healthier. But if those who exercise more and eat a healthy diet also choose to take a vitamin then an experimental study will not replicate the observed trend, finding no difference in the health of matched groups who are given or not given vitamins.

None of these are caused by the use of classical statistics.

Bayesianism

In the Bayesian statistical paradigm which Briggs endorses, *before* examining the data, one specifies what one believes the probability of some hypothesis to be investigated. This is one's "prior" probability. Then after evaluating the likelihood of the data, one revises or "updates" this prior probability in the light of the observations to derive what is termed the "posterior" or final probability.

The calculations are much more complex than those of classical statistics. And how does one choose the prior probability of a hypothesis before one starts calculation? Does one choose a single prior or a range of values?

Some years ago, contemplating a Bayesian analysis of Down syndrome prevalence, I asked David Blackwell, a noted Bayesian statistician, how to choose a prior probability. "Ask the experts," was his advice when I inquired. "But in this field I *am* an expert," I reflected, "and if I don't know how to pick a prior on this matter, who does?"

Moreover, *how* one chooses one's prior probability divides the Bayesian world into at least two schools, subjectivists and objectivists, and there are subdivisions even within those.

Many Bayesians dismiss the importance of choice of different prior probabilities because ostensibly, whatever the values chosen, the derived posterior probabilities converge to one value as more data accumulate. With enough data, the precise choice of the prior does not matter. But often we do not have enough data and the probabilities derived using different priors are not the same. Moreover, if one's prior probability is zero, that is one believes a hypothesis is *impossible*, then no matter how much data one has, as zero multiplied by the likelihood of any data set is still zero, so the derived posterior probability will always be zero.

For example, I assign a zero prior probability to any hypothesis that involves the possible existence of ghosts, zombies, astrological phenomena, and extra-sensory perception (ESP), among others. *No data whatsoever* will convince me these are "real."

Are my priors of zero justified? Or am I a close-minded unscientific bigot?

I defend my stance by noting ESP and the other phenomena are what Gunther Stent has defined in essence as "premature" or, what one may better term "disconnected" from generally accepted scientific knowledge. They cannot be connected to canonical knowledge by a simple series of logical steps. It is only reasonable to ignore such claims.[2]

On what grounds would I change my prior of zero about say, ESP? If investigators discover a human sense organ that transmits and detects hitherto unknown signals with requisite properties, I would revise my prior for the existence of ESP to a positive value. For I could now *connect* the hypothesis for ESP to canonical knowledge. (I can't imagine any evidence that would lead me to do so for ghosts, zombies, astrology, or related phenomena.)

A Bayesian analysis of the validity of frequentist statistics? Finally, consider a Bayesian analysis of Briggs' many interrelated claims, regarding them as part of one overarching "hypothesis": that classical statistics *must* lead to bad decisions.

What is your prior probability Briggs is correct, and on this issue R.A. Fisher, Jerzy Neyman, and other eminent statisticians have been wrong? I will give Briggs the benefit of the doubt and a non-zero prior, of say 0.1 percent. To derive pertinent data, I would search the published literature for all the statistical analyses of data that Fisher, Neyman, etc. had published and note how frequently they had made "bad decisions" because they drew incorrect

inferences from their calculations. (I am unaware of any such instances.) Likewise for Briggs. (I have already noted my view about a few conclusions he has reached.)

I will leave this as an exercise for the reader but predict that a Bayesian analysis will demonstrate with high probability that Briggs is in error.

Ernest B. Hook, M.D., *is professor emeritus at the School of Public Health, University of California, Berkeley;* ebhook@berkeley.edu

1. V.D. Mottironi, E.B. Hook, A. M. Willey, I. H. Porter, R. V. Swift, and N. H. Hatcher. "Decreased HLA heterogeneity in parents of children with Down syndrome," *American Journal of Human Genetics* 35, no. 6 (1983): 1289-1296.

2. Gunger S. Stent, "Prematurity in Scientific Discovery," in Ernest B. Hook, ed. *Prematurity in Scientific Discovery: On Resistance and Neglect*, (Berkeley & Los Angeles: University of California Press, 2002.) 23-34; Ernest B. Hook, "A Background to Prematurity and Resistance to 'Discovery.'" in Hook, ed. *Prematurity in Scientific Discovery*, 3-21.