DOI: 10.51845.38.1.19

Let Go Your Wee P!

by William M. Briggs

***Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science**, Aubrey Clayton, Columbia University Press, 2021, pp. 368, $27.00 hardcover.*

Ask a psychologist what the chances are that a person will walk slower after reading from a list of words having to do with old age than reading from a neutral list. He won't tell you. He can't tell you. What he can tell you is that in his model of walking time, after "controlling" for a number of items including the word list, the "parameter" representing something to do with walking time was highly "statistically significant," with something called a "*p* value" that was boastfully small.

You see a drug commercial on TV and are impressed by the cavorting of the actors. You want to cavort. So you go to the doctor and ask him if Profitol is right for you. You ask him the chance the pill will let you cavort. He won't tell you. He can't tell you. What he can tell you is that he read about an experiment using the pill, and that if the "null hypothesis" comparing that pill to another pill was true, the probability of seeing data that was not seen in the experiment was pretty low.

This answer being incomprehensible, you seek a second opinion. The next doctor gives you a test for cavortitis, the malady which causes an inability to cavort. The test is positive. So you ask the doctor, "Does that mean I got it?" He says, "Well, in those patients with cavortitis, the test comes back positive ninety-five percent of the time. And it's even better for those without the disease: the test comes back negative ninety-nine percent of the time." He writes you an exorbitantly expensive prescription for Profitol. Suddenly you don't feel so good.

And you shouldn't. Because the second doctor didn't answer your question. Neither did the first. Neither did the psychologist. Neither can *anybody* who uses classical statistical procedures. Because those are designed *not* to answer questions put to them in plain language.

Take the second doctor. You asked him (implicitly) what the probability is that you have the disease after testing positive. Let's call your having the disease your "hypothesis." He instead tells you the probabilities having to do with the test. The test is "data," so he gives you probabilities of the data instead of the probability of the hypothesis. Worse, he acted as if the probability of the data *were* the probability of the hypothesis. So did the first doctor and the climate scientist.

So does everybody who uses classical statistical procedures.

The conflating of the probability of the data as if it were the probability of the hypothesis is called, as Aubrey Clayton tells us in the book of the same name, Bernoulli's Fallacy. Named from Jacob Bernoulli, the seventeenth century mathematician who gave us *Ars Conjectandi*, the Law of Large Numbers, and, though he never realized it, the fallacy given in his name. Reasoning based on this fallacy is the foundation of, in Clayton's words, the "frequentist jihad" that has captured and overrun statistical practice. A jihad that has done great bloody work on science, a menace that continues to this day. Hence his subtitle *Statistical Illogic and the Crisis of Modern Science*.

Clayton's is the latest in a long line of works laboring, so far all in vain, to explain that those using "frequentist" statistics are basing their decisions on fallacies—captivating, but hard-to-suppress fallacies—so that when people are right when using them they are right only accidentally, and not because they followed procedure and performed the right calculations. And that people are wrong far, far more often than they know. So wrong that many fields are suffering a replication crisis because of blind slavish adherence to an incoherent methodology.

These are mighty claims. And they have been proved, many times, in many places. Clayton does so again, taking a different tack than earlier works, by stepping through the early history of statistics. His idea is to carefully show where the errors originated, what passions drove them, and how they became ensconced in ordinary scientific practice. He succeeds in this. Succeeding, alas, does not mean he will win new converts. Because, as I said, trying to convince people of the errors of frequentism has been tried often, and no attempt has yet won the field. So, we might expect, neither will Clayton's book.

Now I played a little trick on you, dear reader. If you can agree with the last sentence in the previous paragraph, you are not a frequentist, which is a person beholden to the fallacy. If you see the logic of the paragraph, even if you know nothing of the history, you have already proven to yourself, though you might not yet know it, that something other than frequentism is needed to explain or quantify uncertain propositions, such as whether Clayton's book will defeat his enemies. That something is logic.

Clayton advocates, and a small band of us agree, to cease teaching and using classical statistical methods, and substitute them with logical probability. Sometimes this is called objective Bayesianism, or just Bayes. But there is a branch of Bayesianism that holds with something called subjective probability, so one has to be careful when using or reading the term.

"Bayes" is from the Reverend Thomas Bayes who in the mid-eighteenth century first worked out a formula in logical probability which relates the uncertainty in one proposition when accepting as true another one. For instance, how uncertain is it that "Clayton's book will win few or no converts" assuming it is true that "No work in this line has yet succeeded"? That's logical probability. This example isn't mathematical, but it gets mathematical fast when the accepted-as-true proposition comes in parts, such as observational data does. That math, however, is not the point. The idea is everything.

Bayes's motivation, perhaps, was to poke David Hume. Hume thought he had created a "problem" for induction, and forms of induction are used in probabilistic reasoning. In Clayton's summary, "we have no way of knowing experience is a guide for valid conclusions about the future because if we did, that claim would be based only on past experiences." The induction from (using an example from Hume) "All the many flames observed before have been hot" to "this flame will be hot" is inductive, the conclusion not to be trusted because of the circularity in reasoning. Yet, as far as I know, Hume never stuck his hand into a flame twice. Hume also infamously rejected evidence of miracles *using inductive reasoning* (and was subsequently posthumously critiqued, Hume's work on this not having been published until he had gone to his reward). Bayes, naturally, was all for miracles. And his math is true.

But true math does not necessarily imply applicable math. An equation does not represent what it is claimed to represent solely because someone says so. That has to be proved, and throughout the long and circuitous, and *unfinished*, history of uncertainty, it was. Clayton quickly surveys the work of men like Laplace, Venn, of diagram fame, and George Boole, from whose name we have Boolean logic. These and others discovered ways of putting uncertainty into mathematical form; i.e., probability.

Probability then, as now, has two main interpretations: that the world contains probability and so can be measured, like height or weight, or that probability is purely a matter of our thoughts, like logic. Statistics rode piggyback on probability, and in its early days adopted the first interpretation. Which, under the tutelage of men like Karl Pearson, who originated the "correlation coefficient," Francis Galton, who gave us "regression to the mean," and the combative Ronald Fisher, who savagely beat down all critics. It was Fisher who gave us the "*p* value," which is explained below.

Fisher was a mathematician and geneticist and a brilliant experimenter on things like crops and fields, so it was natural for him to think probability grew out of the soil. His book *Statistical Methods for Research Workers* found its way into the hands of nearly everybody. The interpretation Fisher preferred was eventually called *frequentism*, based on the belief that not only does probability exist as a real thing, but that every uncertain measurement must and could be embedded in infinite sequences of measurable parts of reality. The infinities are needed to prove the math.

It was Fisher more than any other figure who said that if you plug your numbers into these formulas, scientific success can be yours. To say he triumphed is like saying the sun is hot. Nearly everybody now uses the kinds of methods Fisher developed and advocated. Fisher had his detractors, like Pearson, and a small cadre of physicists of whom more below, but none could stand up to his pugilistic probabilism. Clayton does a good job detailing the hot feuds between the main players, showing us with careful mathematical examples each sad step in the descent down the frequentist ladder.

Frequentist procedures have two centers: model parameters and hypothesis testing. If you've read any scientific literature that uses statistics, you've seen both. The most common entry employs these procedures to justify correlations, and pretend these correlations have proved a causal link. Clayton knew of no way to write his book, and I know of no way to adequately review and do it justice, without walking through an example. Since his are rather mathematical, I'll pick something more qualitative.

Suppose you suspect that the fluctuating distance between Jupiter and the Sun is causative of the number of secretaries employed in Alaska, because of an obscure astrological theory. You collect data on each and calculate the "correlation coefficient" between the two sets. Correlations run between -1 and +1, given by the letter r, with more extreme numbers indicating greater correlation. Your Jupiter-Alaska correlation turns out to be 0.95.

Next step is to form a "null" hypothesis, which says the correlation is 0. Then you find a "test statistic" which is "unbiased" or "uniformly most powerful" or possesses any of a number of obscure desirabilities (Clayton reviews the history of each). That test statistic is then calculated for your observed data. With me so far, dear reader? I hope so, because the next step is where the miracle happens.

Clever mathematicians have figured how to calculate the probability that test statistics like yours would *exceed* the one you calculated, if you were to repeat your "experiment" an infinite number of times, but *only* if the null hypothesis is *true*. This number is called the "*p* value." If yours is less than the magic number, then you are entitled by long custom to claim your result is "statistically significant"—where "significant" (I promise you) means having

a p-value less than the magic number. I do not need to tell you the value of this magic number. You have seen it hundreds, even thousands of times.

It turns out for the Jupiter-Alaska data, the p-value is less than 0.01. That is significant. Your next step is to write a paper in the *Journal of Astrological Amazingologies* to discuss how your correlation, now become a causal claim, has proved your theory. Because while every scientist knows that correlation does not (logically) imply causation—they also believe that correlation turns into causation when backed by a wee *p* value.

You will, I hope, have noticed that you rejected your null because the probability of seeing data *you didn't see* is small, just like the first doctor. You may laugh if you like, but the example is a real one, and found at the Tyler Vigen's "Spurious Correlations" website. He has a lovely plot of the Jupiter-Sun distance and Alaskan secretaries, the overlap of the two datasets being almost complete. This is one of a slew of preposterous connections. You will have your own, but my favorite is Popularity of the Name Killian and Air Bag Recalls: $r = 0.939$, $p < 0.01$. *Kill*ian—get it?

The statistical procedure followed in this example is the *exact* one used in serious research. The conclusion we reached, the leap tying our hypothesis test to our causal claim, is identical. Yet we know our conclusion is absurd. Why? Partly because, as Clayton says, "hypothesis testing is meaningless without alternatives against which to test. When the hypothesis that a population correlation is exactly 0 ... is tested against its simple negation—that the correlation is not 0—the null hypothesis will always lose if the amount of data is large enough." Large data gives wee *p* values like colleges give "degrees," a weakness which everybody knows but forgets (like correlation becoming causation) when it is *their* wee *p* value.

A null of precisely 0 correlation is ridiculous, but believed because of another fallacy, a reverse of the one which says that "significant" correlations become causes. It is supposed that a correlation of 0 implies no causal connection between two things. That's false. A correlation is just a mathematical formula that spits out values between -1 and +1. Cause isn't in it, anywhere. Even if the correlation is -1 or +1, it could be that two sets of numbers which gave rise to these extremes are both themselves caused by some third thing. There are many possibilities.

If you have access to software that can compute correlations, input numbers from things you know have no causal connection between them. See how often you get exactly 0. Never is a good guess. Do it again for numbers you know are connected. See how often you get exactly -1 or +1. Never wins again.

In frequentism you're not allowed to use "prior" information, like your knowledge of causal connections, to put a measure of uncertainty on the hypothesis the correlation is 0, or on *any* other hypothesis, of any kind. Yet

given our prior belief that there is no causal connection between Jupiter and Alaskan secretaries, no amount of observation is likely to convince us one exists. If it looked like there was a connection, as one looks like it exists in the example, we put it down to coincidence—a decision based on a logical (Bayesian) probability.

Clayton uses two excellent examples here, the first borrowed from ET Jaynes, whom we meet below, on whether we should believe results in parapsychology based only on frequentist hypothesis testing. His second example is of some notoriety, about the Cornell psychologist Daryl Bem who was sure his *p* was wee enough to convince the world psychic powers were real. Clayton, with complete mathematical detail, shows us Bem followed to the letter frequentist procedures that are scientifically *de rigueur*, and concludes "Bem's paper was a wake-up call that it was time to rethink those methods."

Now if do not believe in astrology nor psychic powers, even though these fields produce results which pass frequentist statistical tests, but you do believe other mainstream results because they pass these same tests, you are being inconsistent. You are breaking frequentist rules, opting for logical probability for some results, and believing frequentism for others. Yet frequentism is a theory for how *all* probability works, not just some. You don't get to pick and choose.

The proper conclusion, since you do not believe the obviously absurd examples, even though they are mathematically correct, is not to believe *any* result based on frequentist methods. Including the ones in "top" journals.

Do not forget that tests of correlation use *the wrong hypotheses*. They give us probabilities of data, when what we want to know is the chance that there are connections between things. What we want to know is, given the data we collect, and everything else we know about the matter, what is the probability the connection is real? You cannot get that using hypothesis testing, which is a form of Bernoulli's Fallacy.

Clayton has a list of what hypothesis testing fails to do, but which everybody thinks it can. Like "A *p* value is a measure of the degree of confidence in the obtained result." No. It isn't. Neither do confidence intervals (if you've heard of them) grant this. Another: "A *p* value is the probability the null is true (or false) given the data." It is not. It is a long, dreary list. Many times such warnings have been published. *Many.* They never stick. Never. A Bayesian would say it's probable this one won't, either.

Believing in hypothesis testing is what gave us the replication crisis. This is where the best results fail to reproduce—and they fail, says Clayton, because the results were wrong in the first place, but certified by fallacious reasoning.

Here is just one of a large and growing collection of similar stories, repeated in every field which relies upon frequentist statistical methods:

> A replication project for economics research conducted by a group of 18 collaborators found they were able to replicate 11 out of 18 experiments (61 percent) published in the *American Economic Review* and the *Quarterly Journal of Economics* in 2011-2014.... On average, the effects they did find to be significant were 66 percent of the originals. A 2017 study conducted by [John] Ioannidis suggested "the majority of the average effects in the empirical economics literature are exaggerated by a factor of at least 2 and at least one-third are exaggerated by a factor of 4 or more."

This is found over and over, in sociology, medicine, everywhere. Researchers attempting to replicate psychology results "found they were able to replicate only 35 of ... 97 results (36 percent) ... Of the effects they *did* replicate, they found the average size of the effect to be about half the original." The example beginning this review was a genuine, well-touted result that failed to replicate. Clayton's chapter on the crisis reveals a dismal, depressing record. But one of exceeding importance when we are routinely asked to "Follow the Science!" Why should we trust results based on fallacies? Answer: we should not.

The replication crisis will be with us as long as frequentism is. There is no way to be careful with a fallacy or think (as some say) *p* values have some uses. Fallacies have no uses, except as instruction of the young. The entire classical statistical apparatus has to go.

Summing up the old approach, Clayton observes, "Bernoulli's Fallacy is buried deep in modern scientific practice. It was planted there by the early frequentist statisticians, who were especially motivated to think of statistics as a completely objective discipline free from interpretation or prior judgement." Frequentists say, "Let the data speak for itself!"

Which data might that be? The kind gathered with a specific hypothesis in mind? A hypothesis subjectively chosen? A hypothesis someone had to know something about, or no one would have known which data might be relevant to subjectively chosen observations? That data? Well, let's listen to it. But how? Through the lens of a statistical model that was subjectively chosen *ad hoc*? Through a use of Bernoulli's Fallacy?

The alternative is simplicity itself. It is probability as a measure of uncertainty, which is in our thoughts, not in things. Gather all the evidence probative of some proposition that interests you and calculate the uncertainty of that proposition conditional on that evidence. That's logical probability. You want the probability the Jupiter-Sun distances creates or destroys Alaskan secretarial positions conditional on all you know of physics, in addition to the screwy data, which does not exist in isolation.

Subjectivity turns out to be a false charge; or, rather, a true one, but one made with just as much justice in frequentism. It is not a bug but a feature that when the evidence changes, the probability changes. It should. Think of the cop ruling a suspect out after he learns of his alibi. That's logical proba-

bility, or Bayes, in action. All questions of uncertainty can be put in this painless, easy-to-understand, plain-language way. It also solves the parameter problem, which I hinted at above, but which is too involved to explain here, except to say all results should be put in terms of measurable observables (like the opening questions), which is what science is supposed to be about.

Besides the math, which is harder in Bayes but which is anyway shunted off into hidden code, a weakness of the approach is that it doesn't make decisions for you. I mean this seriously. Frequentism lightens your mental load, which is a comfort to most. But it makes bad decisions, as the replication crisis and Bernoulli's Fallacy proves. There is no magic number in Bayes, and so the feeling that one has followed proper ritual, as psychologist Gerd Gigerenzer rightly charges of frequentist procedures, doesn't exist. Results are left as only vague probabilities. This uncertainty is too disquieting for many. One wants solid answers. Yet logic cannot give them where the evidence is insufficient.

It is a remarkable fact that most of the development of the logical, epistemological-based approach to probability came not from mathematicians or statisticians, but from outsiders, especially physicists. Starting with Laplace, and, closer to us, Harold Jeffreys, who showed how probability is found when information is at a minimum, and Richard Threlkeld Cox, who proved how uncertainty could be made into math (no small thing). Missing from Clayton's list is the late Australian philosopher David Stove, who used logical probability to prove there is no problem with induction.

The most fruitful of this bunch, and the man Clayton (and many of us) longs to bring attention to, is the late physicist Edwin T. Jaynes, especially through his book *Probability Theory: The Logic of Science*. If you work in probability or statistics and have never read it, your education cannot be said to be complete.

Jaynes does it all, building probability from plain, commonsense propositions, building a towering edifice of mathematics. It is a gorgeous text. But only the mathematically adept will make it all the way through. Clayton, perhaps to his disadvantage, produces a snippet of calculation that Jaynes found trivial but that most will struggle with, to show how he, Clayton, had no choice but to simplify his own book. This might work against Clayton in frightening casual readers who will see these equations, and the many Clayton has in his text to work out his own examples, and flee. Still, Clayton does the world a service in pointing people to Jaynes.

The reader also quickly realizes Clayton is very, very, most exceptionally nervous that people will discover that the originators of many frequentist statistical methods would not have been able to fill out their DEI statements satisfactorily, and thus think he, Clayton, might not be able to, either. He never misses a chance to distance himself from history. I lost count of the number of times

"racist," "eugenics," and similar tedious boo-words appeared. Godwin's Law was in full effect, and, though not part of the story, the man with the abbreviated moustache makes his obligatory appearance to frighten the children.

So rattled is he by the past that Clayton is adamant that all uses of classical statistical terms "carry an ethical instruction, calling on us to hunt down deviance and punish impurity, affirming a particular kind of ableist … imperialist, white-supremacist, capitalist, patriarchal cultural violence." Dude. Think about who you're punishing the next time you speak of your *p* value.

In spite of this skittishness, the book is worth reading. The writing can be dry, the math trying, and the details somewhat exhausting, but that is a function of the material discussed. One never livens up a party by announcing, "The statisticians have arrived!"

William M. Briggs *is an independent writer, scientist, and consultant. He has taught statistics at Cornell University, where he earned an M.S. in atmospheric science and a Ph.D. in statistics; matt@wmbriggs.com. He is the author of* Uncertainty: The Soul of Modeling, Probability & Statistics *(Springer, 2016), yet another failed attempt to co nvince the world to give up "frequentism." Briggs wrote "Math: Old, New, and Equalitarian," a review essay on Christopher J. Phillips's* The New Math *and Paul Ernest, Bharath Sriraman, and Nuala Ernest's* Critical Mathematics Education.
