A discussion about statistics
(With emphasis on the Bayesian approach)

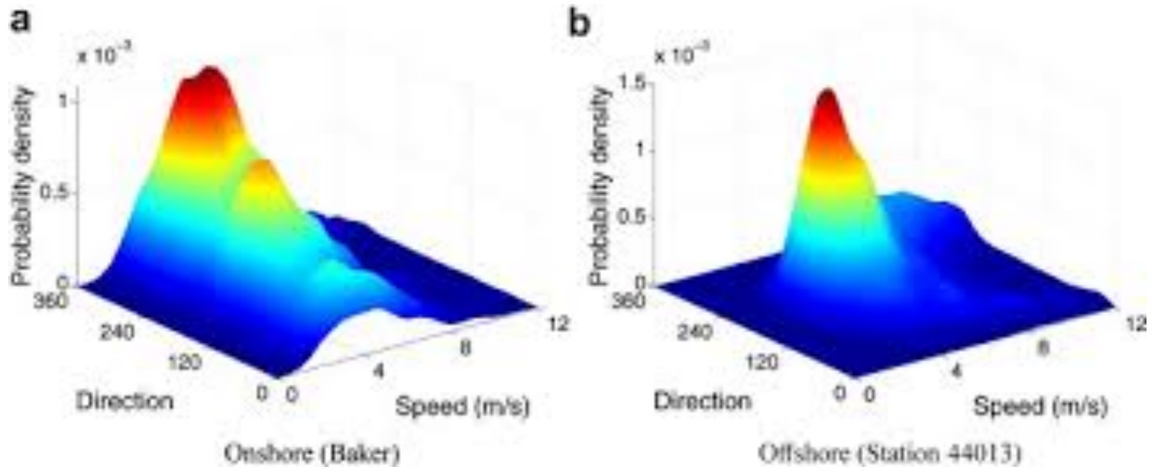
Radu Craiu
Department of Statistical Sciences
University of Toronto

Statistics

- Statistics is the science/art of extracting information from data.
- It relies on replication/repetition (observing only one patient won’t do much).
- It also relies on variation (observing the same kind of patient over and over again will not help with the population at large).
- Learning from data is embedded in our survival instincts but sometimes we get it wrong if we rely only on instinct (see ideas from behavioural economics).
- Statisticians are also good at/obsessed with reducing the complexity of a problem and finding simpler solutions whenever possible (“Occam’s razor”).

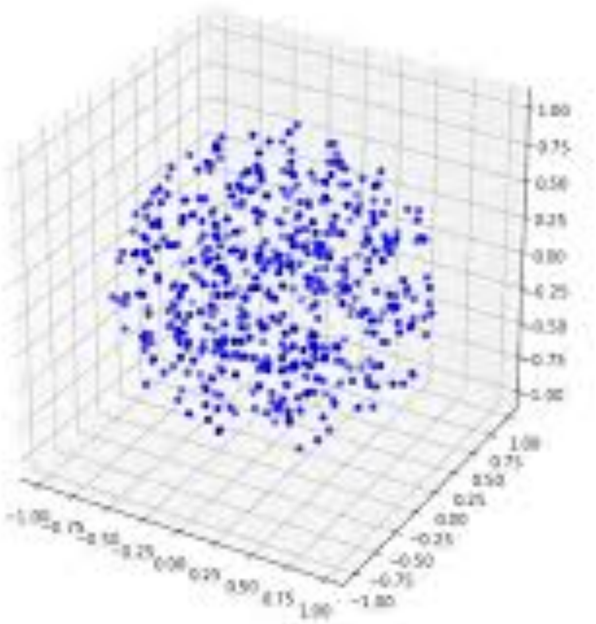
Central idea

- Central to statistics is the idea of a probability distribution.
- It represents variation but also frequency.
- It tells us how diverse is the population of outcomes and how likely a given outcome is.
- Most importantly, entire distributions are characterized by only a few parameters (once we estimate those we know everything).



Reducing complexity

- Data in high dimension (imagine 20 dimensions).
- Need to understand patterns, perhaps predict where a new observation comes in.
- Hopeless, unless we introduce some structure.
- If we assume that the data are multivariate Gaussian then all of a sudden we reduce an infinite dimensional problem to a finite manageable one.



Motivating article

- The paper discusses criticisms of frequentist statistics.
- Presents the Bayesian paradigm and its advantages.
- Today I will discuss some of the above, using some examples.

The Lancet
Volume 404, Issue 10457, 14–20 September 2024, Pages 1067-1076

Review

Bayesian statistics for clinical research

Ewan C Goligher MD [a b c], Anna Heath PhD [d e], Michael O Harhay PhD [f g h]

What we want from data

- Sometimes we want to understand the generative model — the mechanism through which nature produces outcomes.
- We want to approximate the generative model using distributions and we estimate their corresponding parameters along with uncertainty quantifications (e.g.,confidence intervals).
- We want to test hypotheses (‘drug A is better than drug B’).
- We want to predict the values of new outcomes (‘what will this patient look like in 1 month’).

Example: Pump failure data

- Pump failure data: number of failures in time t (in 1K hours)
- The model: $y_i \sim Poisson(\lambda \times t_i)$
- $\hat{\lambda} = 0.214$ with a std error of 0.025
- We expect 2.14 failures in 10,000 hours.

Pump	Failures (y)	Time (t)
1	5	94.32
2	1	15.72
3	5	62.88
4	14	125.76
5	3	5.24
6	19	31.44
7	1	1.05
8	1	1.05
9	4	2.10
10	22	10.48

Interpreting the uncertainty

- Theory tells us that if we were to repeat this experiment (infinitely) many times then 95% of times we will obtain $\lambda \in (0.164, 0.264)$ (estimate $\pm$ 2x std.err).
- This can allow managers to forecast the required stock of new pumps.
- But we do not have an infinite population of nuclear plants!
- Perhaps not all pumps are made by the same company so using the "same" $\lambda$for all of them is wrong
- Testing the null $H_0 : \lambda = 1$ yields the p-value 10^{-16} which is interpreted as "Assuming the null is true, the chance that we estimate λ to be at least as far away from 1 as 0.214 is 10^{-16} " — what a mouthful!
- Reject the null!

Common complaints against previous analysis

- Interpretability: confidence intervals and p-vals have awkward interpretations that lead to confusion and misuse.
- Replicability crisis in science.
- Ideally we want to be able to say "Probability the null is true is …".
- Or "The probability that the interval (a,b) contains λ is 95%".
- Both of these are offered by the Bayesian approach!

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets PPV = $(1-\beta)R/(R$

Bayesian approach

- Everything we have done up to now is frequentist statistics.
- Bayesian statistics is very different.
- Bayesians don't do confidence intervals and hypothesis tests.
- So what do they do? Bayesians treat parameters as random variables.
- To a Bayesian, probability is the only way to describe uncertainty.
- Things not known for certain - like values of parameters - must be described by a probability distribution.

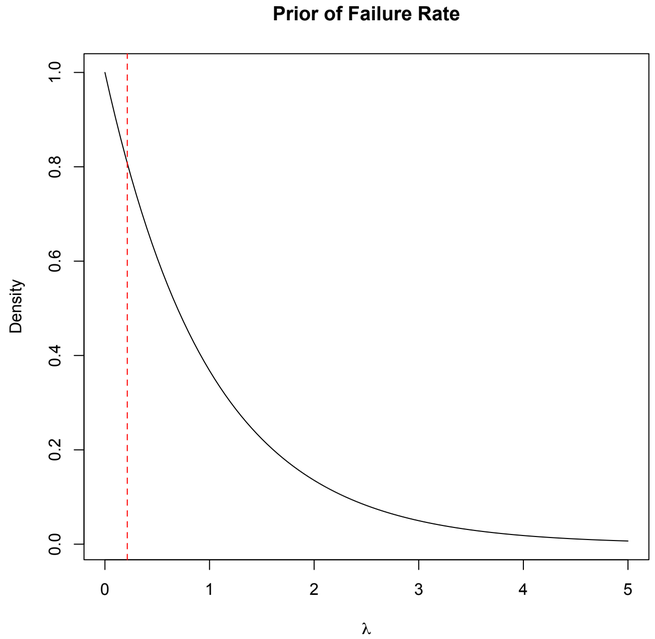
From prior to posterior

- Suppose you are uncertain about something.
- Then your uncertainty is described by a probability distribution called your prior distribution.
- Suppose you obtain some data relevant to that thing. The data changes your uncertainty, which is then described by a new probability distribution called your posterior distribution.
- The posterior distribution reflects the information both in the prior distribution and the data.
- Most of Bayesian inference is about how to go from prior to posterior.

Ingredients for Bayesian analysis

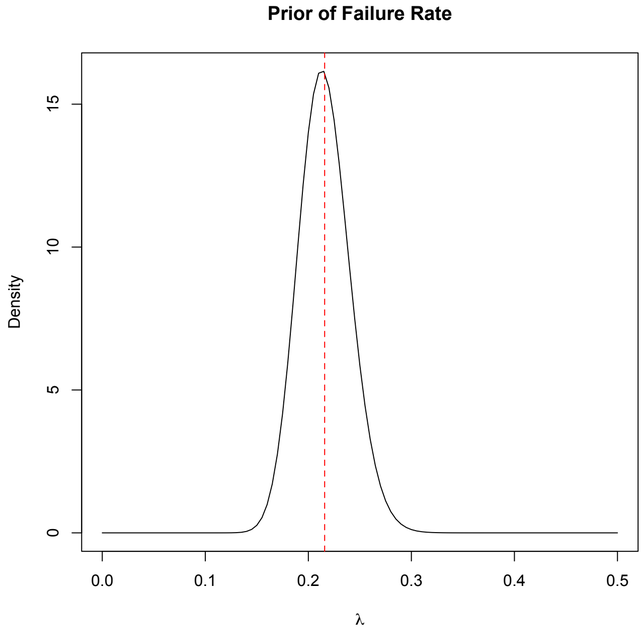
- The sampling distribution which describes the distribution of the data (this depends on parameters) - this one is used by frequentists too.
- In the pump example $f(y_i \mid t_i, \lambda) = Poisson(\lambda \times t_i)$
- The prior distribution summarizes what we know *a priori* (i.e. before looking at the data about the parameter)
- In the pump example the prior is $p(\lambda) = Gamma(1,1)$
- Posterior distribution of λ is

$$p(\lambda \mid y_1, \ldots, y_{10}) = \frac{p(\lambda)\prod_{i=1}^{10} f(y_i \mid \lambda)}{p(y_1, \ldots, y_{10})} = \frac{p(\lambda)\prod_{i=1}^{10} f(y_i \mid \lambda)}{\int p(\lambda)\prod_{i=1}^{10} f(y_i \mid \lambda) d\lambda}$$



Posterior distribution

- $p(\lambda \mid y_1, \ldots, y_{10}) = Gamma(76, 352.14)$
- Our knowledge about the parameter has changed substantially
- If we had to summarize the entire posterior by one point we could choose the mean (which is almost the same as the mode) $\hat{\lambda} = 0.21$
- The 95% credible interval is (0.168,0.265) very similar to the frequentist CI but with better interpretation.
- In this case the frequentist and Bayesian inferences are similar. This is not always the case.



What we didn’t discuss

- The choice of the model. For instance one could argue the pumps are different but similar:
  - $f(y_i \mid \lambda_i) = Poisson(\lambda_i \times t_i)$
  - $p(\lambda_i \mid \alpha, \beta) = Gamma(\alpha, \beta)$
  - $p(\alpha) = p(\beta) = Uniform(0,100)$
- Choice of priors: how does the prior influence the inference?
- How to choose between two models?
- How to judge whether a model fits well or not?
- How to predict (e.g., number of failures for a new pump)?
- How to test a hypothesis? (Short answer: Bayesian don’t really do it, they just compare the null hypothesis model that has, say $\lambda = 1,$ with the general one.

Parting thoughts

- If you need to do a statistical analysis, make sure that at least one statistician is being consulted (*at least* have them look at what you did)
- If you need to interpret a statistical analysis, you may need some help for complicated scenarios/models/analyses
- Here we scratched the surface, but we can go further once you identify some topics of interest
- If you want to see what I do: https://raducraiu.com
- Questions?