Bayesian Statistics
(a very brief introduction)

Ken Rice

Epi 516, Biost 520

1.30pm, T478, April 4, 2018

Overview

Rather than trying to cram a PhD's-worth of material into 90 minutes...

- What is Bayes' Rule, a.k.a. Bayes' Theorem?
- What is Bayesian inference?
- Where can Bayesian inference be helpful?
- How, if at all, is it different to frequentist inference?

Note: the literature contains many pro- and anti-Bayesian polemics, many of which are ill-informed and unhelpful. I will *try* not to rant, and aim to be accurate.

Further Note: There will, unavoidably, be some discussion of *epistemology*, i.e. philosophy concerned with the nature and scope of knowledge. But...

Overview

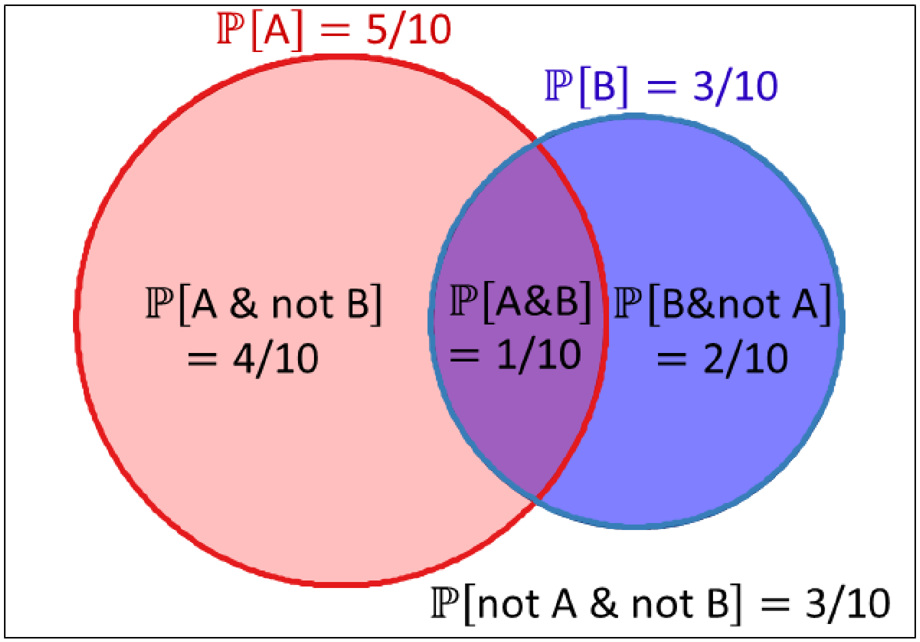
Using a spade for some jobs and shovel for others does *not* require you to sign up to a lifetime of using only Spadian or Shovelist philosophy, or to believing that *only* spades or *only* shovels represent the One True Path to garden neatness.

There are different ways of tackling statistical problems, too.

Bayes' Theorem

Before we get to inference: Bayes' *Theorem* is a result in conditional probability, stating that for two events A and B...

$$\mathbb{P}[\,A|B\,] = \frac{\mathbb{P}[\,A \text{ and } B\,]}{\mathbb{P}[\,B\,]} = \mathbb{P}[\,B|A\,]\frac{\mathbb{P}[\,A\,]}{\mathbb{P}[\,B\,]}.$$



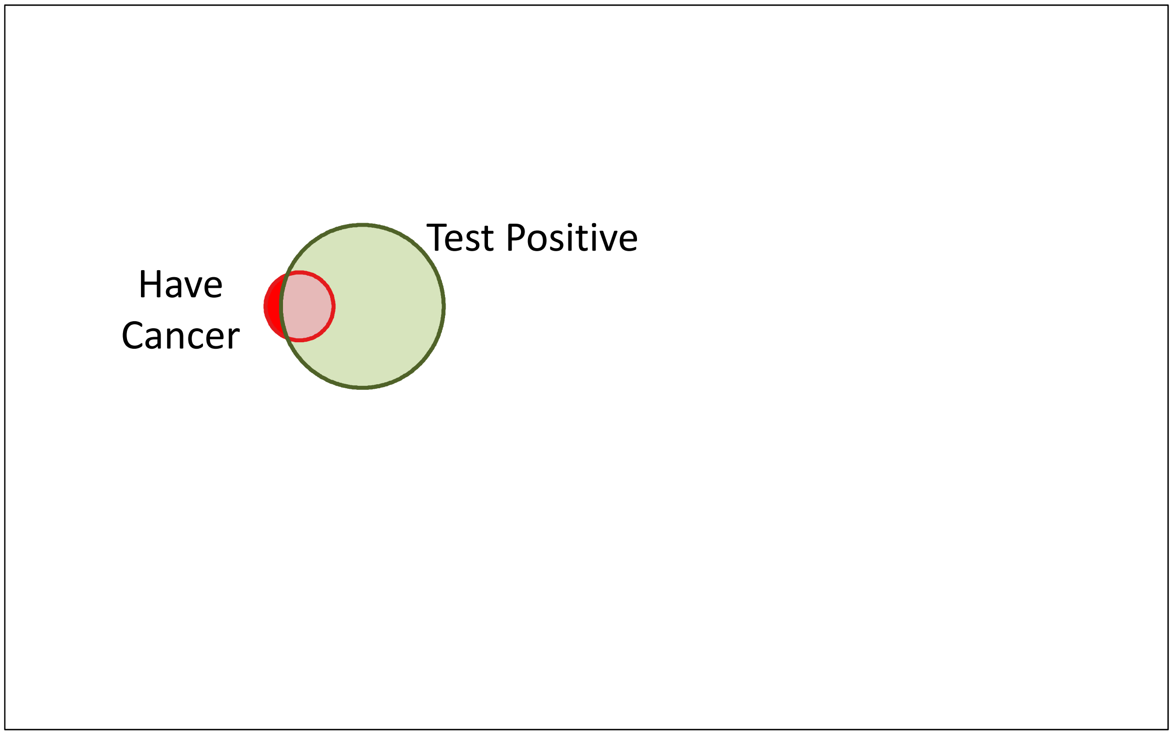
In this example;

- $\mathbb{P}[\,A|B\,] = \frac{1/10}{3/10} = 1/3$
- $\mathbb{P}[\,B|A\,] = \frac{1/10}{5/10} = 1/5$
- And $1/3 = 1/5 \times \frac{5/10}{3/10}$ ($\checkmark$)

In words: the conditional probability of A given B is the conditional probability of B given A scaled by the *relative* probability of A compared to B.

Bayes' Theorem

Why does it matter? If 1% of a population have cancer, for a screening test with 80% sensitivity and 95% specificity;



$$\begin{aligned} \mathbb{P}[\,\text{Test +ve}|\text{Cancer}\,] &= 80\% \\ \frac{\mathbb{P}[\,\text{Test +ve}]}{\mathbb{P}[\,\text{Cancer}\,]} &= 5.75 \\ \mathbb{P}[\,\text{Cancer}|\text{Test +ve}\,] &\approx 14\% \end{aligned}$$

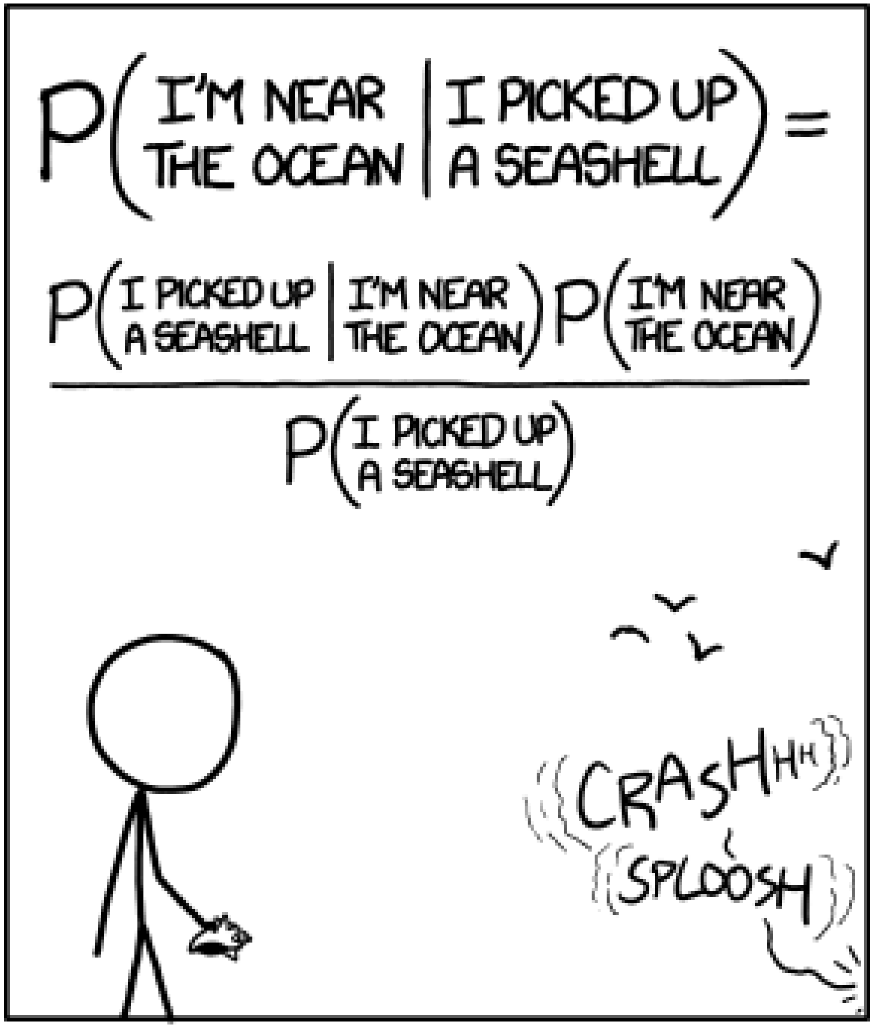
... i.e. most positive results are actually false alarms

Mixing up $\mathbb{P}[\,A|B\,]$ with $\mathbb{P}[\,B|A\,]$ is the *Prosecutor's Fallacy*; a small probability of evidence given innocence need NOT mean a small probability of innocence given evidence.

Bayes' Theorem: Sally Clark

- After the sudden death of two baby sons, Sally Clark (above, center) was sentenced to life in prison in 1999
- Among other errors, expert witness Prof Roy Meadow (above right) had wrongly interpreted the small probability of two cot deaths as a small probability of Clark's innocence
- After a long campaign, including refutation of Meadow's statistics, Clark was released and cleared in 2003
- After being freed, she developed alcoholism and died in 2007

Bayes' Theorem: XKCD at the beach



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

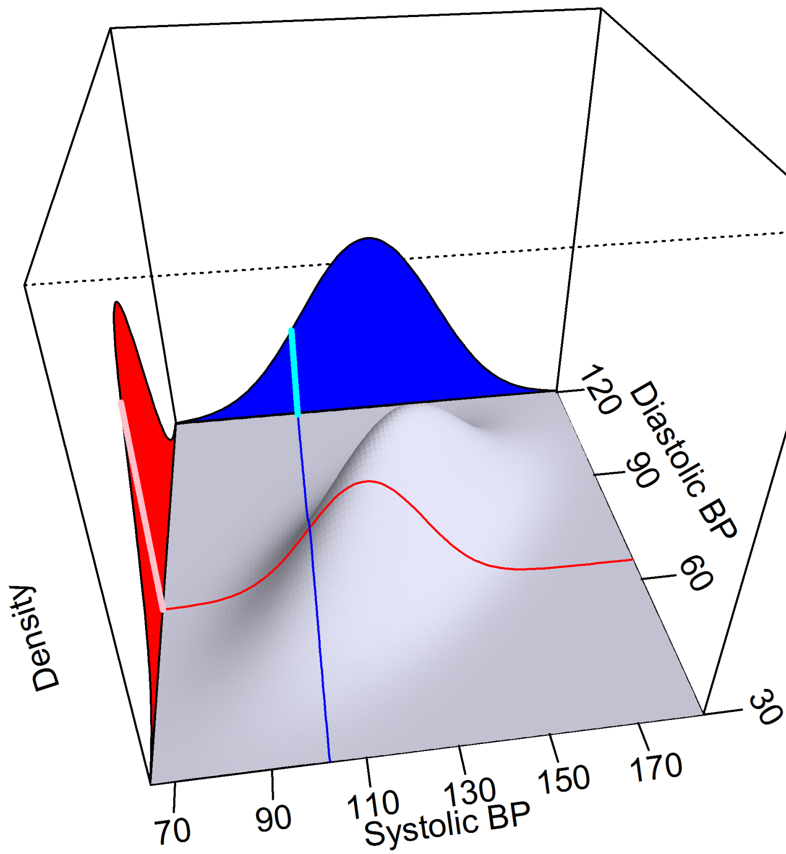
This is roughly equal to

$$\frac{\textit{\# of times I've picked up a seashell at the ocean}}{\textit{\# of times I've picked up a seashell}},$$

...which in my case is pretty close to 1, and gets much closer if we're considering only times I didn't put it to my ear.

Bayes' Theorem

Bayes' theorem also applies to continuous variables – say systolic and diastolic blood pressure;



The conditional densities of the random variables are related this way;

$$f(x|y) = f(y|x)\frac{f(x)}{f(y)}$$

...which we can write as

$$f(x|y) \propto f(y|x)f(x).$$

This proportionality statement is just a re-wording of Bayes' Theorem.

Note: Like probabilities, densities are ≥ 0, and 'add up to 1'.

Bayesian inference

So far, nothing's controversial; Bayes' Theorem is a rule about the 'language' of probabilities, that can be used in any analysis describing random variables, i.e. any data analysis.

Q. So why all the fuss?
A. Bayesian *inference* uses more than just Bayes' Theorem

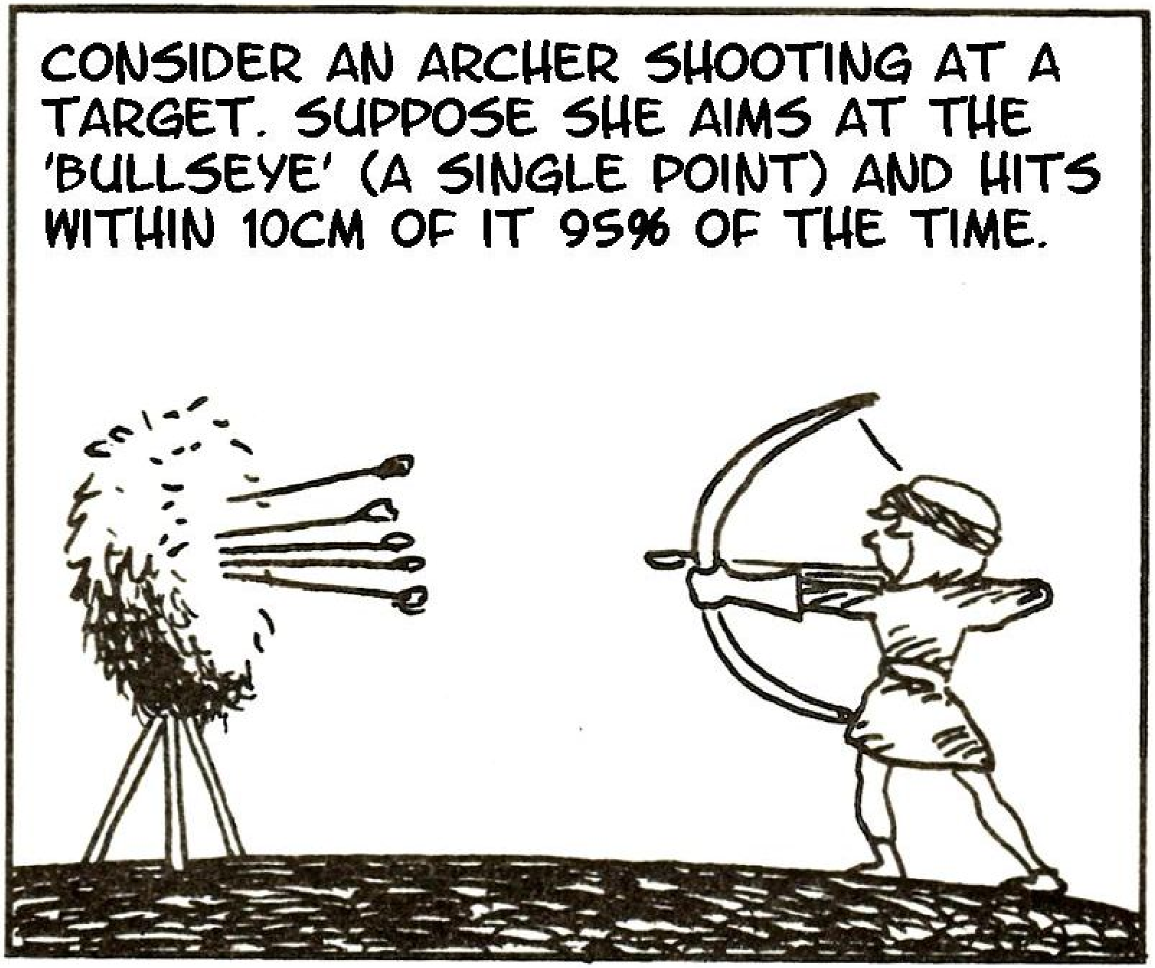
> In *addition* to describing random variables, Bayesian inference uses the 'language' of probability to describe what is known about parameters.

Note: Frequentist inference, e.g. using p-values & confidence intervals, does *not* quantify what is known about parameters.*

*many people initially *think* it does; an important job for instructors of intro Stat/Biostat courses is convincing those people that they are wrong.

Freq'ist inference (I know, shoot me!)

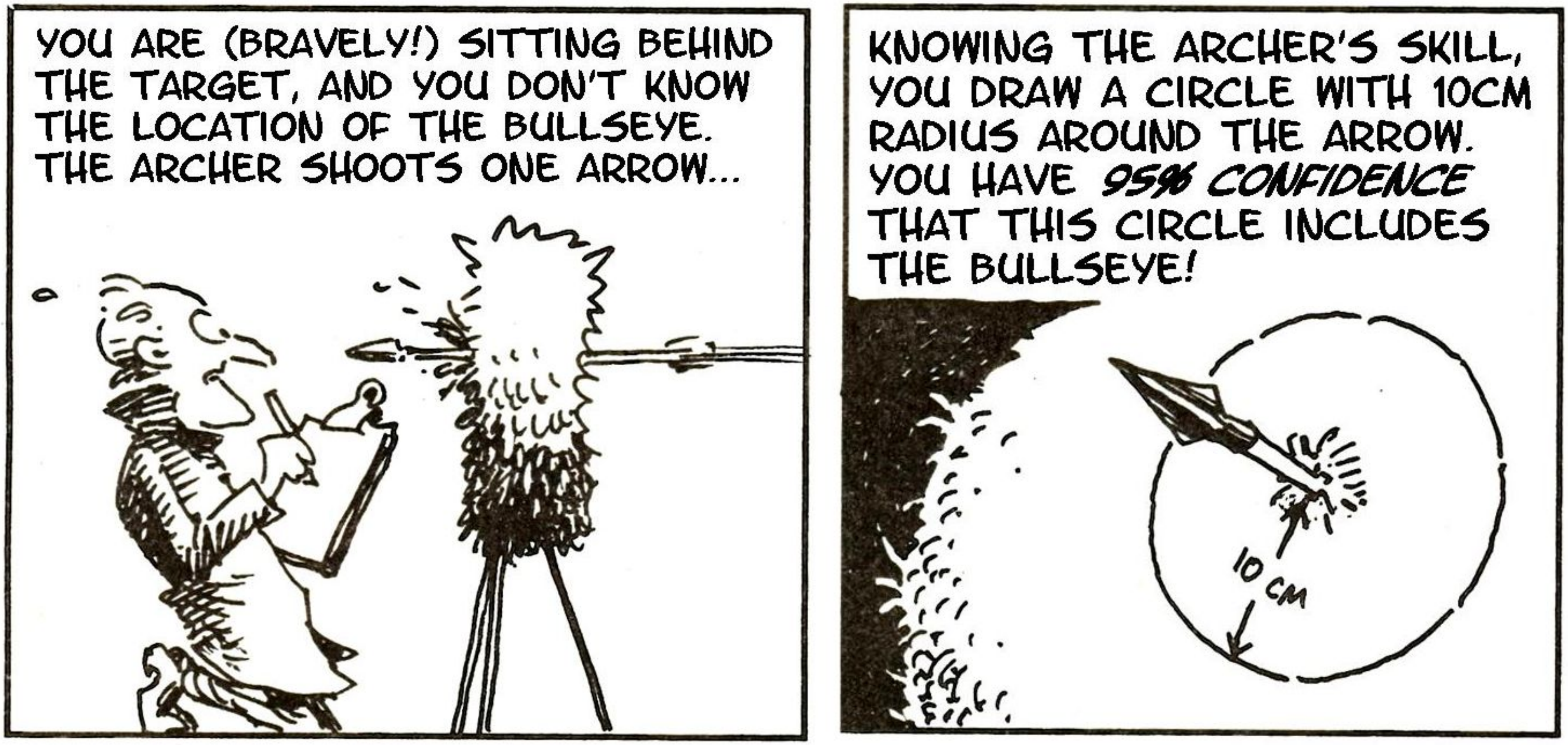
Frequentist inference, set all a-quiver;



Adapted from Gonick & Smith, *The Cartoon Guide to Statistics*

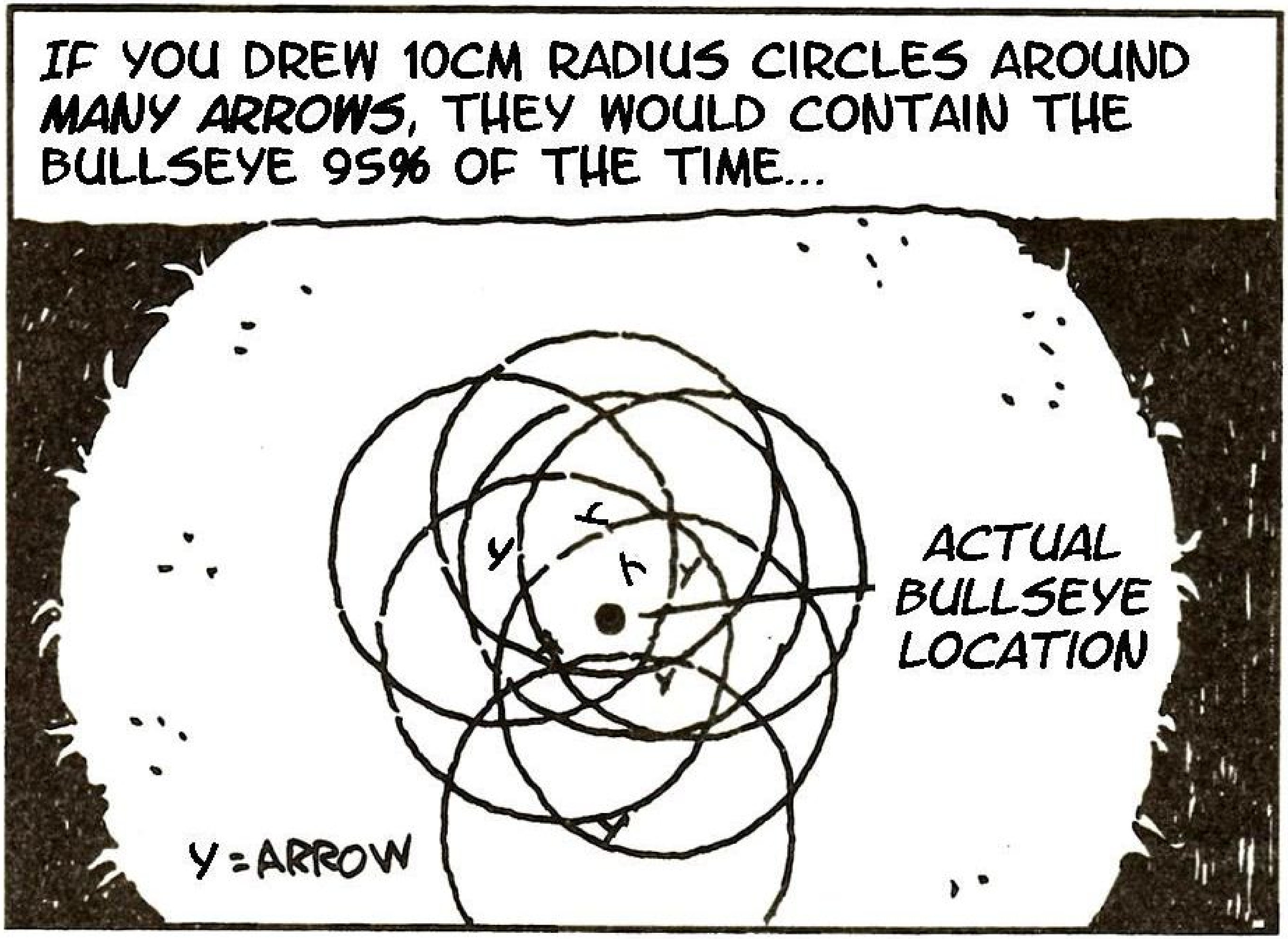
Freq'ist inference (I know, shoot me!)

Frequentist inference, set all a-quiver;



We 'trap' the truth with 95% confidence. Q. 95% of what?

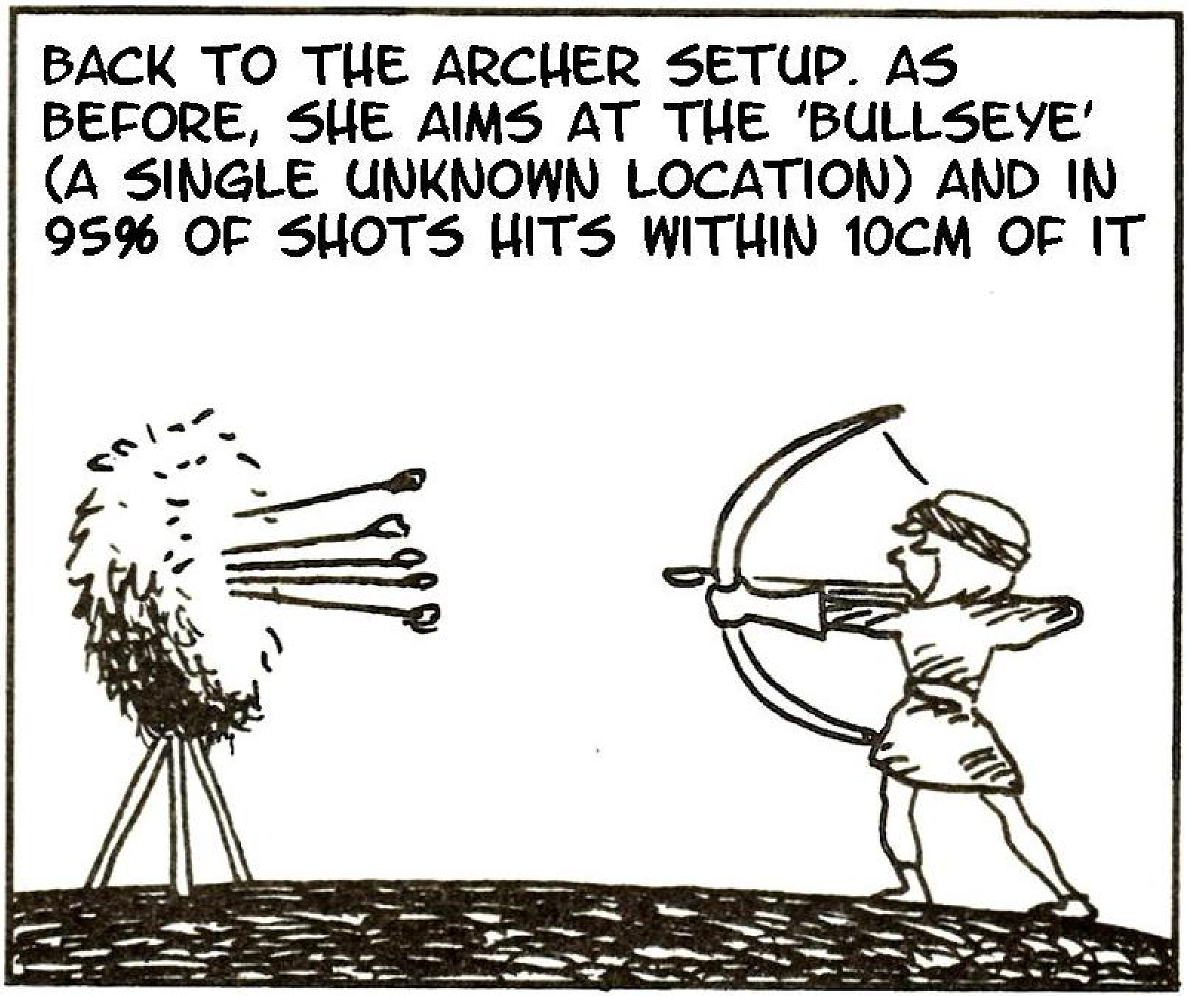
Freq'ist inference (I know, shoot me!)



The interval traps the truth in 95% of experiments. To define anything frequentist, you *have to imagine* repeated experiments.

Freq'ist inference (I know, shoot me!)

Let's do some more 'target practice', for frequentist testing;

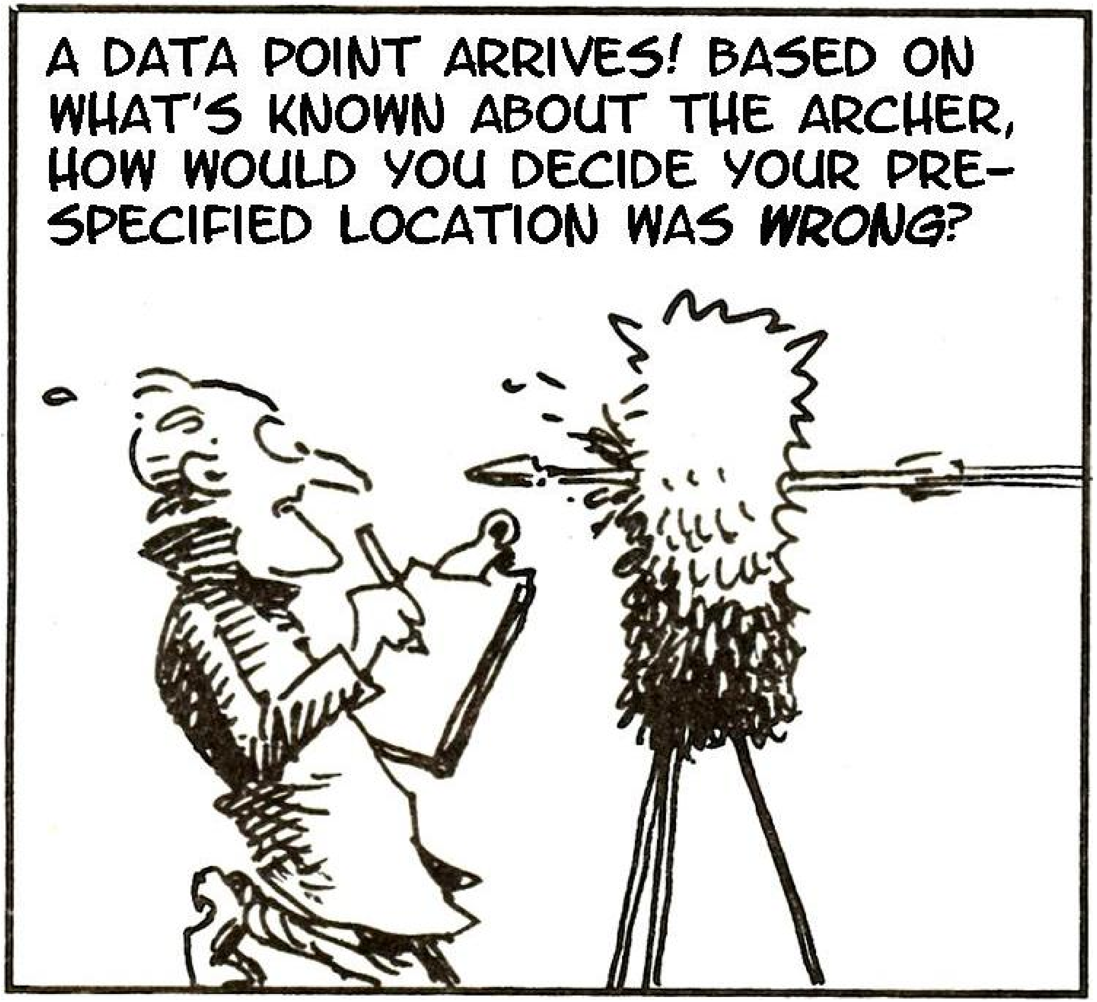


Freq'ist inference (I know, shoot me!)

Let's do some more 'target practice', for frequentist testing;

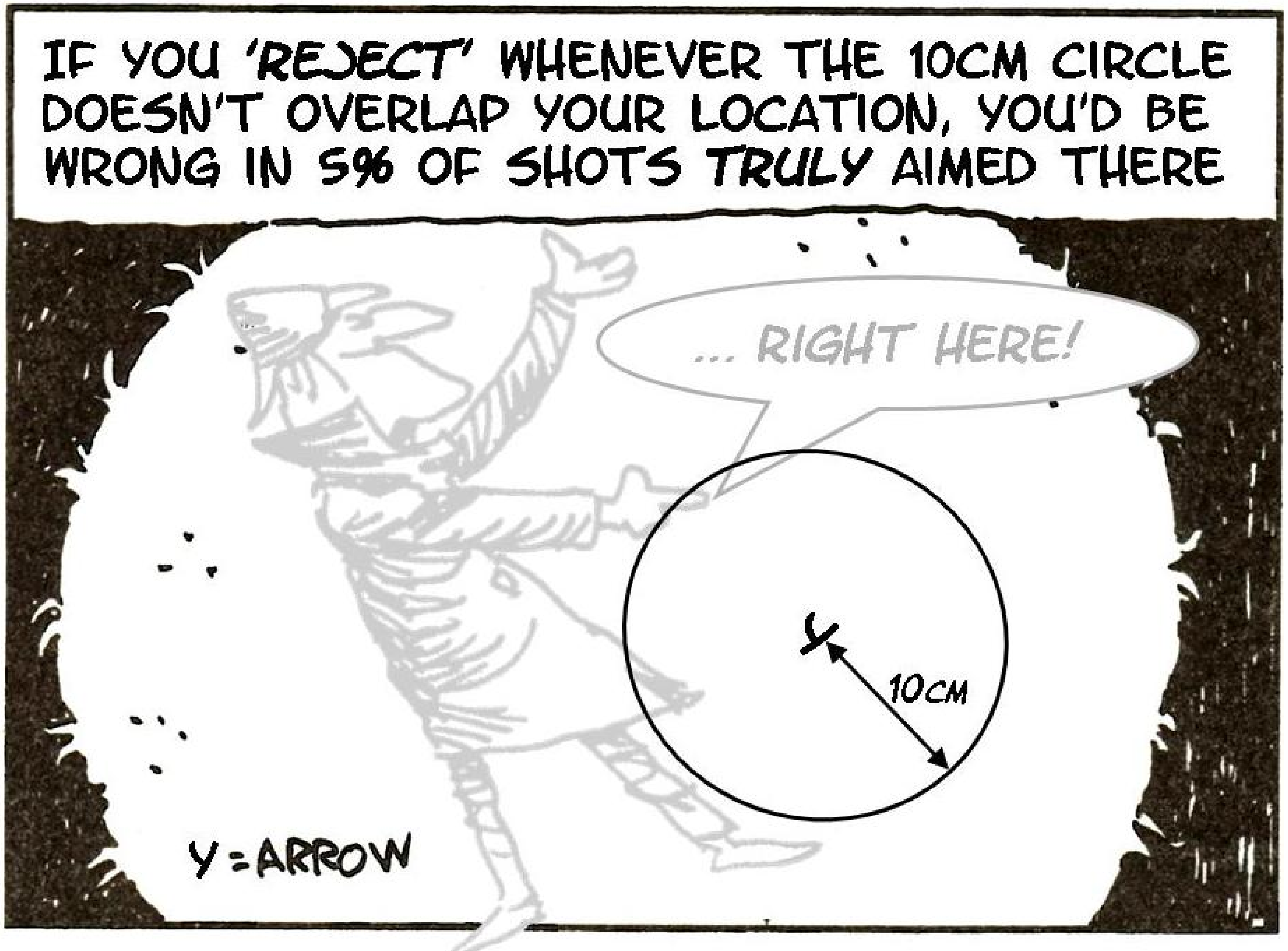
Freq'ist inference (I know, shoot me!)

Let's do some more 'target practice', for frequentist testing;



Freq'ist inference (I know, shoot me!)

Let's do some more 'target practice', for frequentist testing;



Freq'ist inference (I know, shoot me!)

For testing or estimating, imagine running your experiment again and again. Or, perhaps, make an argument like this;

On day 1 you collect data and construct a [valid] 95% confidence interval for a parameter θ_1. On day 2 you collect new data and construct a 95% confidence interval for an unrelated parameter θ_2. On day 3 ... [the same]. You continue this way constructing confidence intervals for a sequence of unrelated parameters θ_1, θ_2, ... 95% of your intervals will trap the true parameter value

Larry Wasserman, All of Statistics

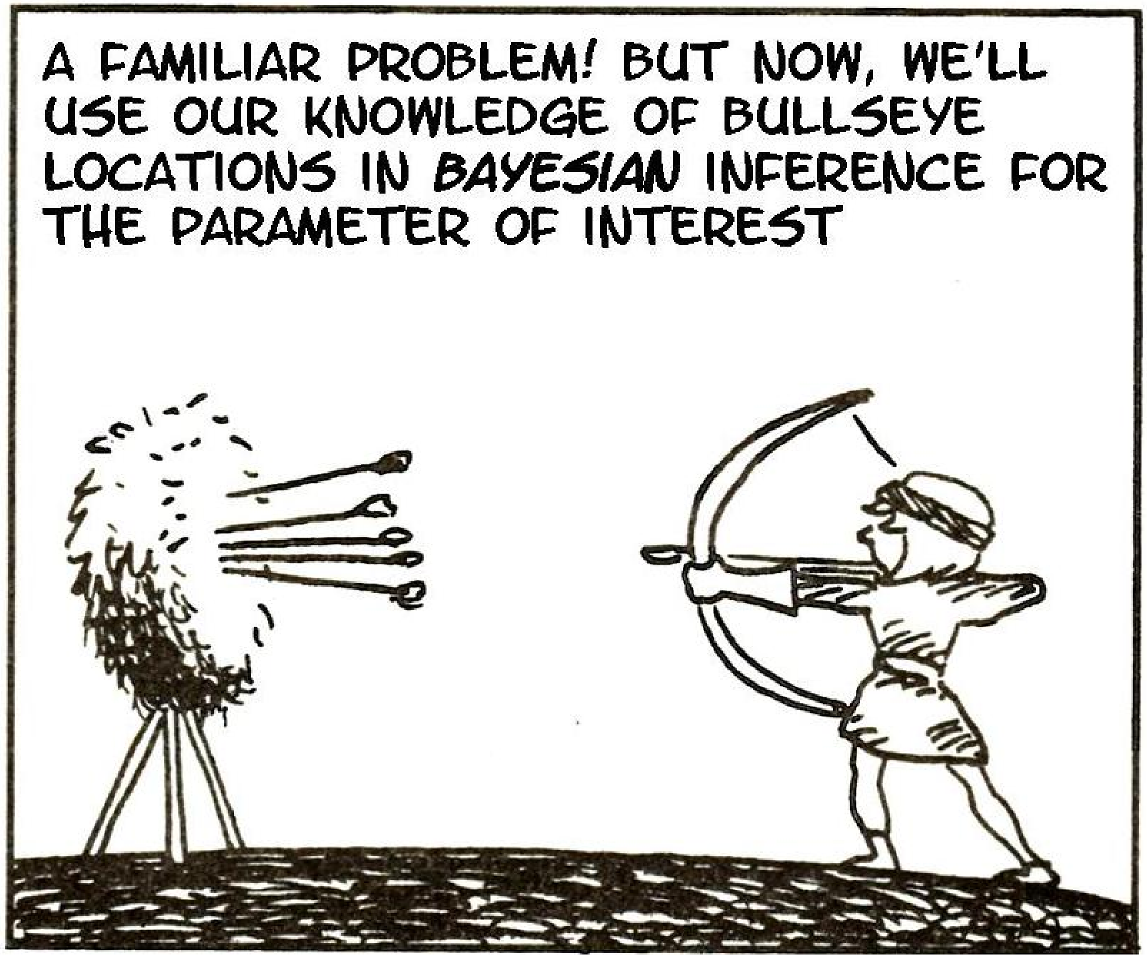
This alternative interpretation is also valid, but...

- ... neither version says anything about whether your data is in the 95% or the 5%
- ... both versions require you to think about many other datasets, not just the one you have to analyze

How does Bayesian inference differ? Let's take aim...

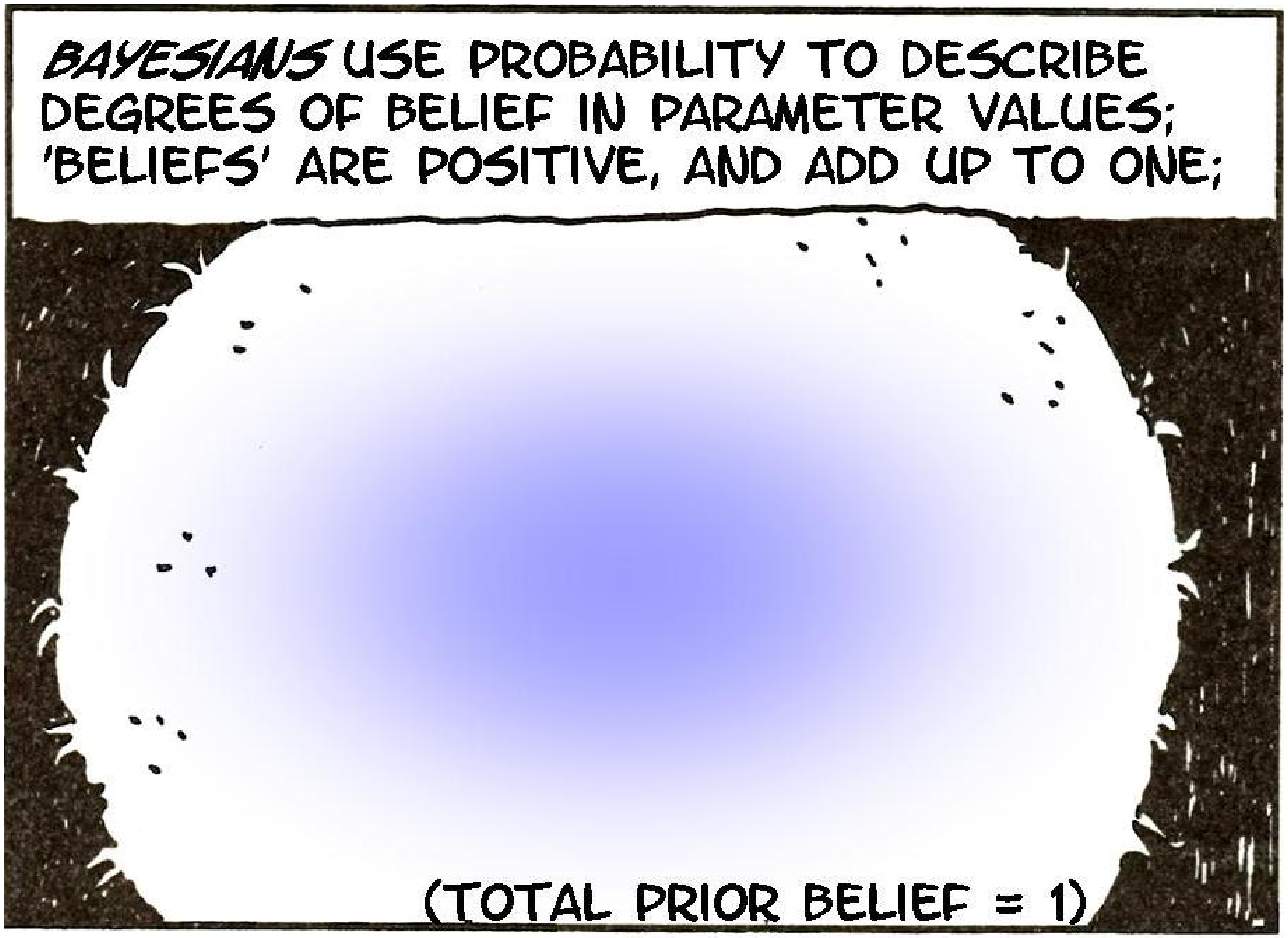
Bayesian inference

[Appalling archery pun goes here]



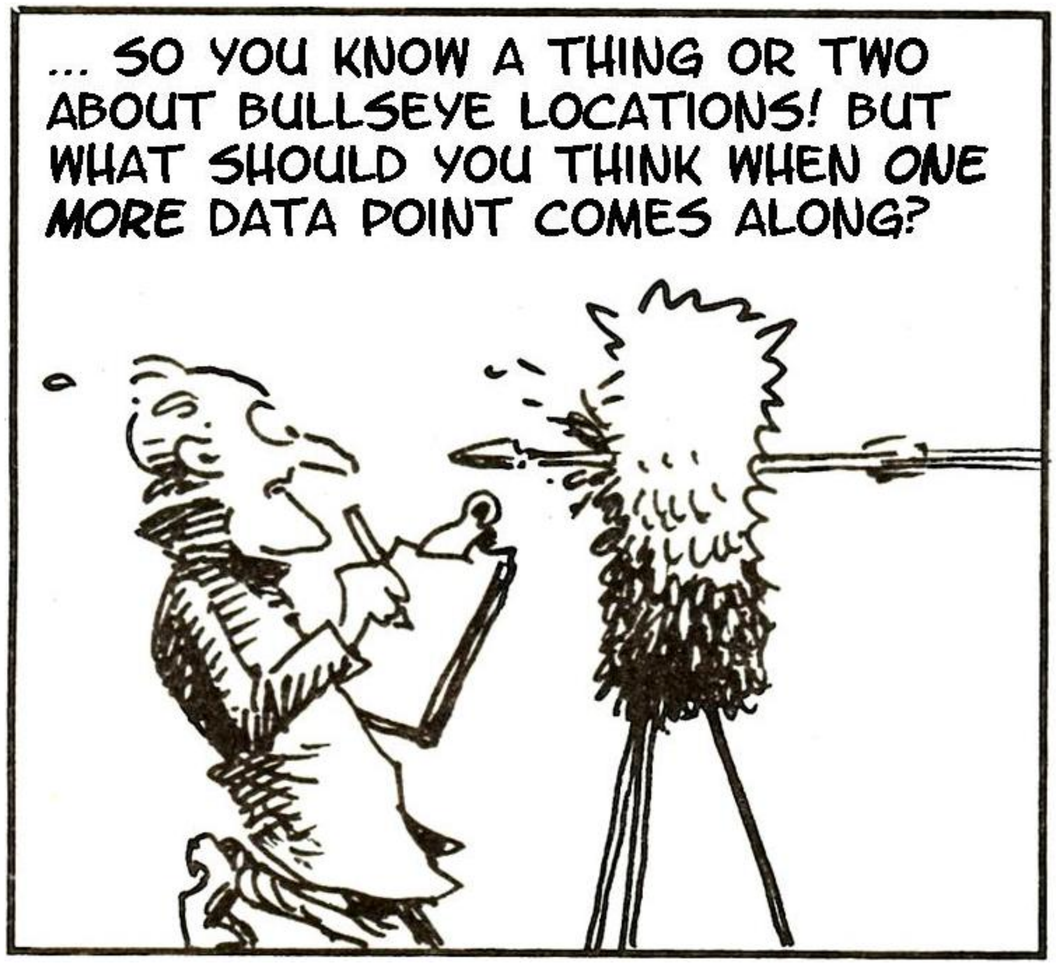
Bayesian inference

[Appalling archery pun goes here]



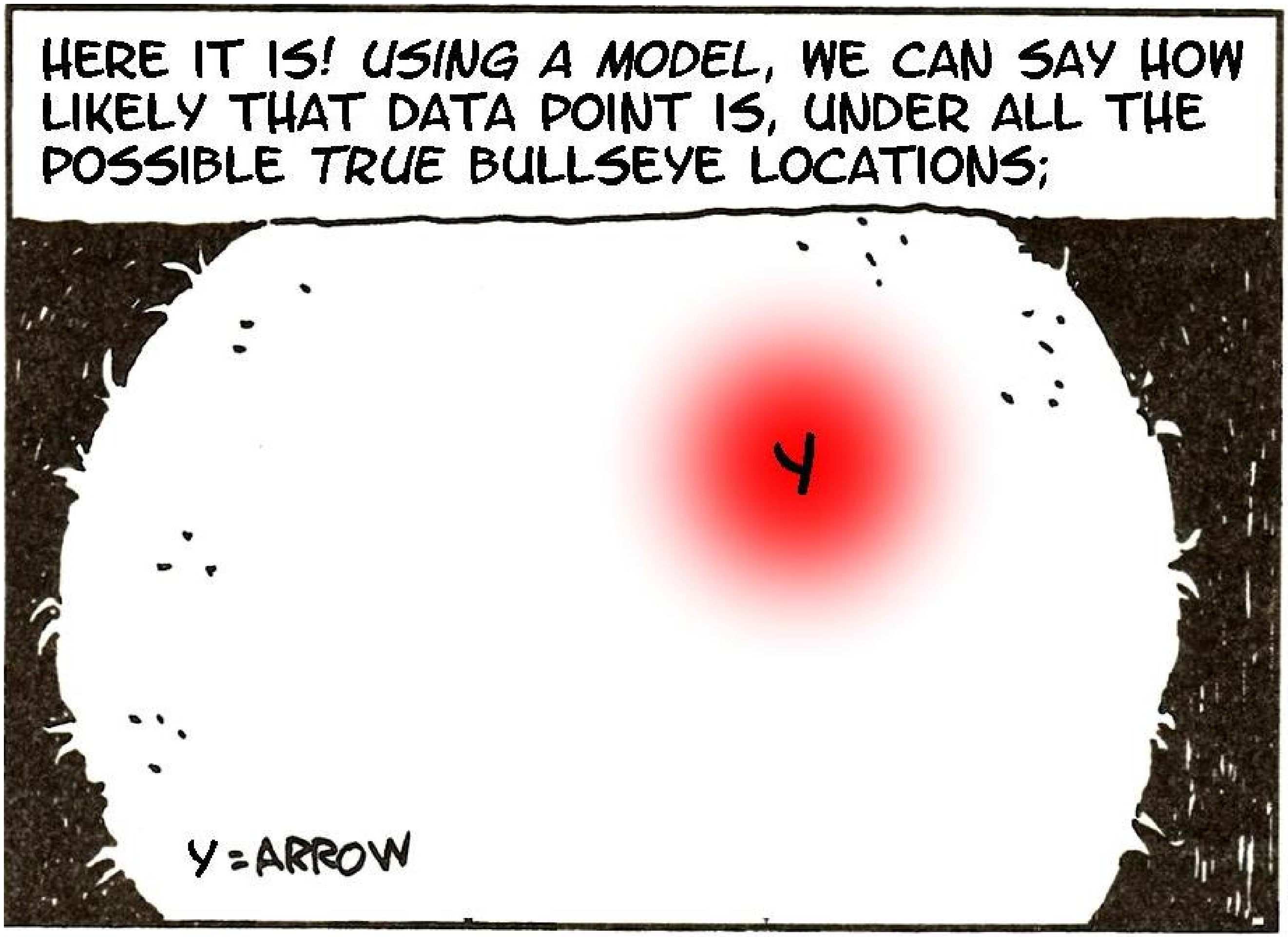
Bayesian inference

[Appalling archery pun goes here]



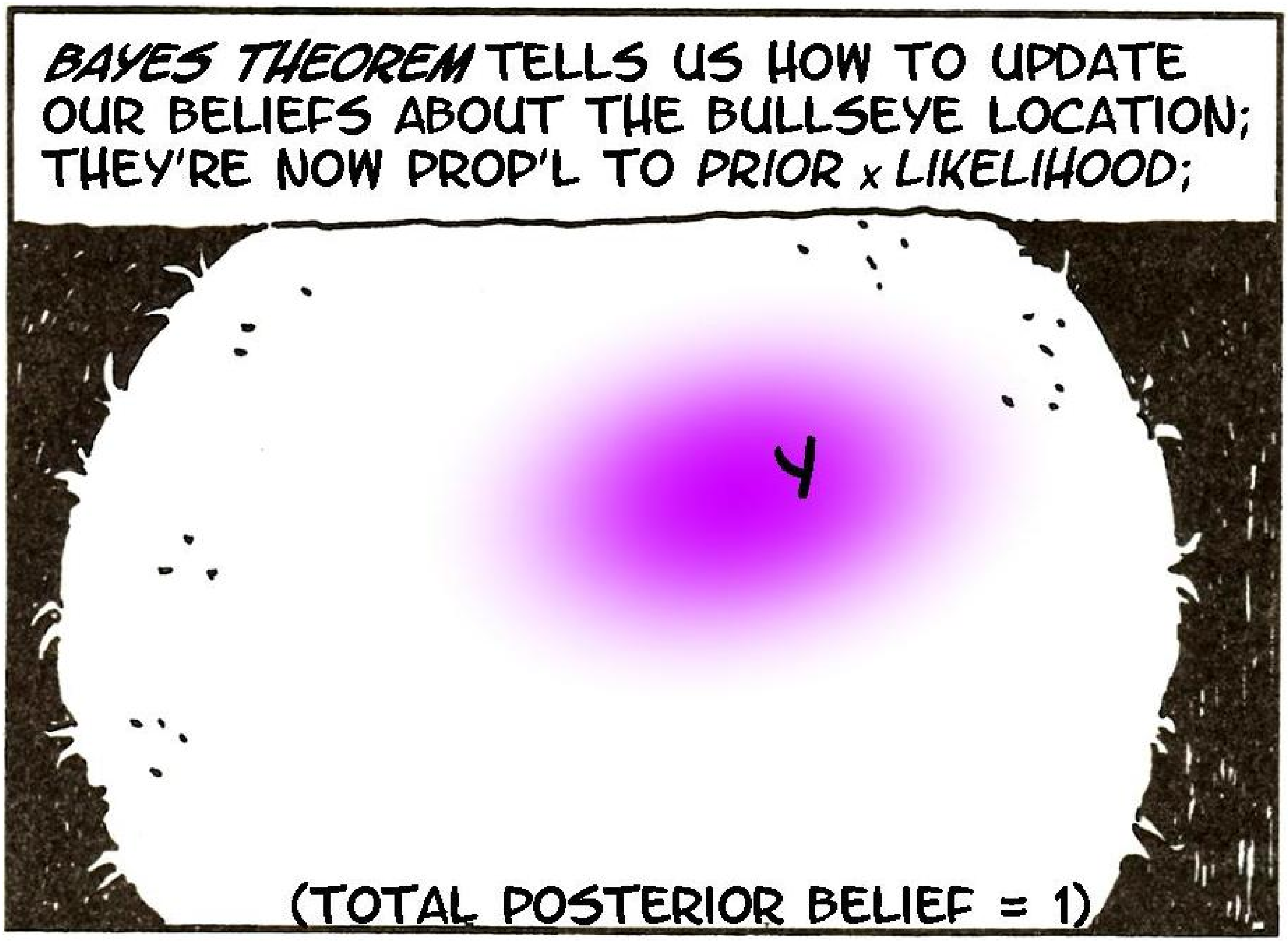
Bayesian inference

[Appalling archery pun goes here]



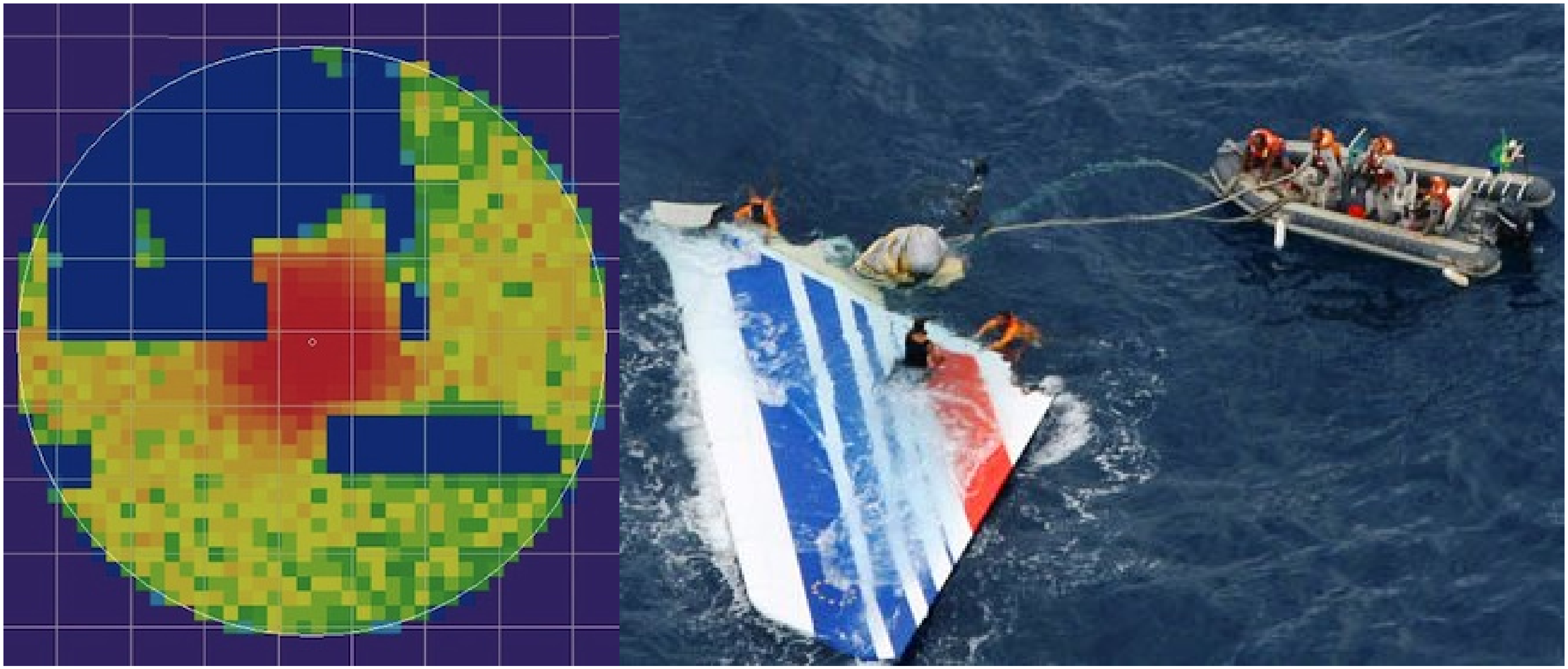
Bayesian inference

[Appalling archery pun goes here]



Bayesian inference

Here's *exactly* the same idea, in practice;



- During the search for Air France 447, from 2009-2011, knowledge about the black box location was described via probability – i.e. using Bayesian inference
- Eventually, the black box was found in the red area

Bayesian inference

How to update knowledge, as data is obtained? We use;

- **Prior distribution:** what you know about parameter $\boldsymbol{\beta}$, excluding the information in the data – denoted $\pi(\boldsymbol{\beta})$
- **Likelihood:** based on modeling assumptions, how [relatively] likely the data $\mathbf{Y}$ are *if* the truth is $\boldsymbol{\beta}$ – denoted $f(\mathbf{Y}|\boldsymbol{\beta})$

So how to get a **posterior distribution:** stating what we know about $\boldsymbol{\beta}$, combining the prior with the data – denoted $p(\boldsymbol{\beta}|\mathbf{Y})$? Bayes Theorem *used for inference* tells us to multiply;

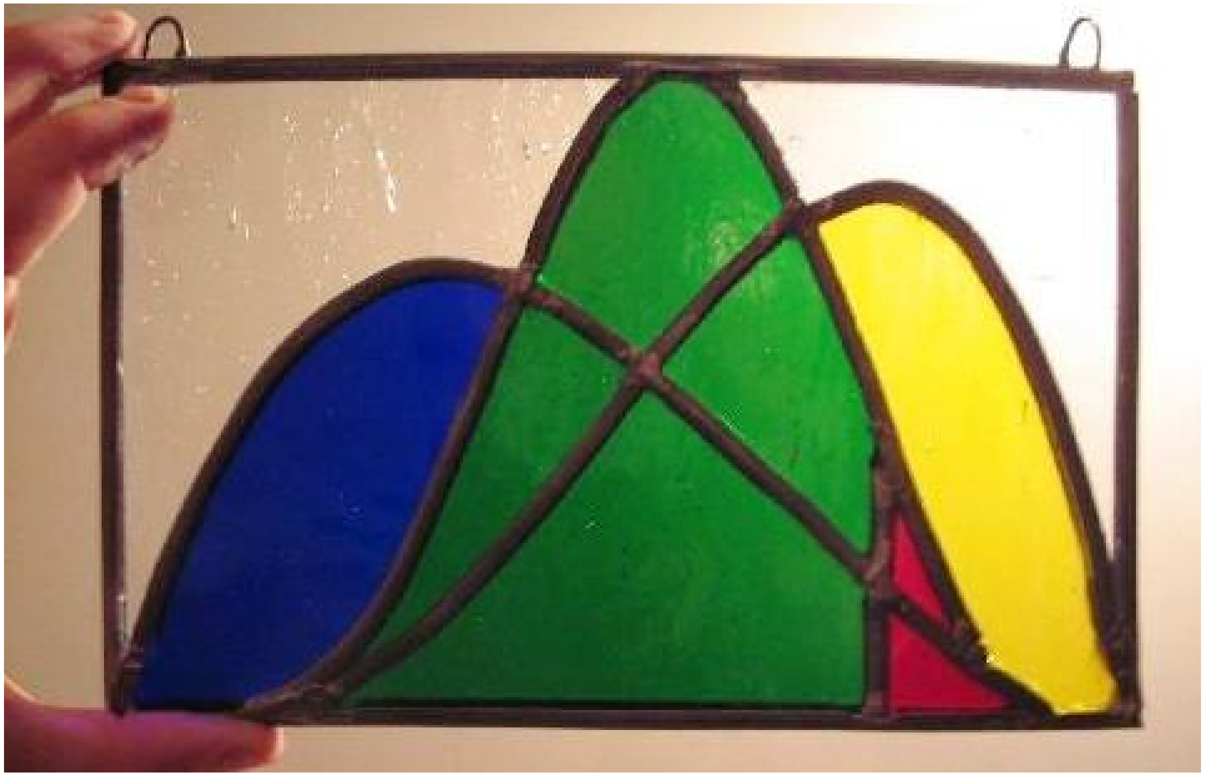
$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{Y}) &\propto f(\mathbf{Y}|\boldsymbol{\beta}) \times \pi(\boldsymbol{\beta}) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior.} \end{aligned}$$

... and that's it! (essentially!)

- No replications – e.g. no replicate plane searches
- Given modeling assumptions & prior, process is automatic
- Keep adding data, and updating knowledge, as data becomes available... knowledge will concentrate around true $\boldsymbol{\beta}$

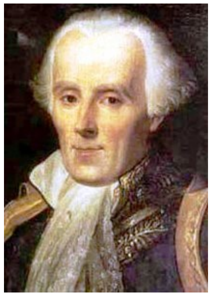
Bayesian inference

Bayesian inference can be made, er, transparent;



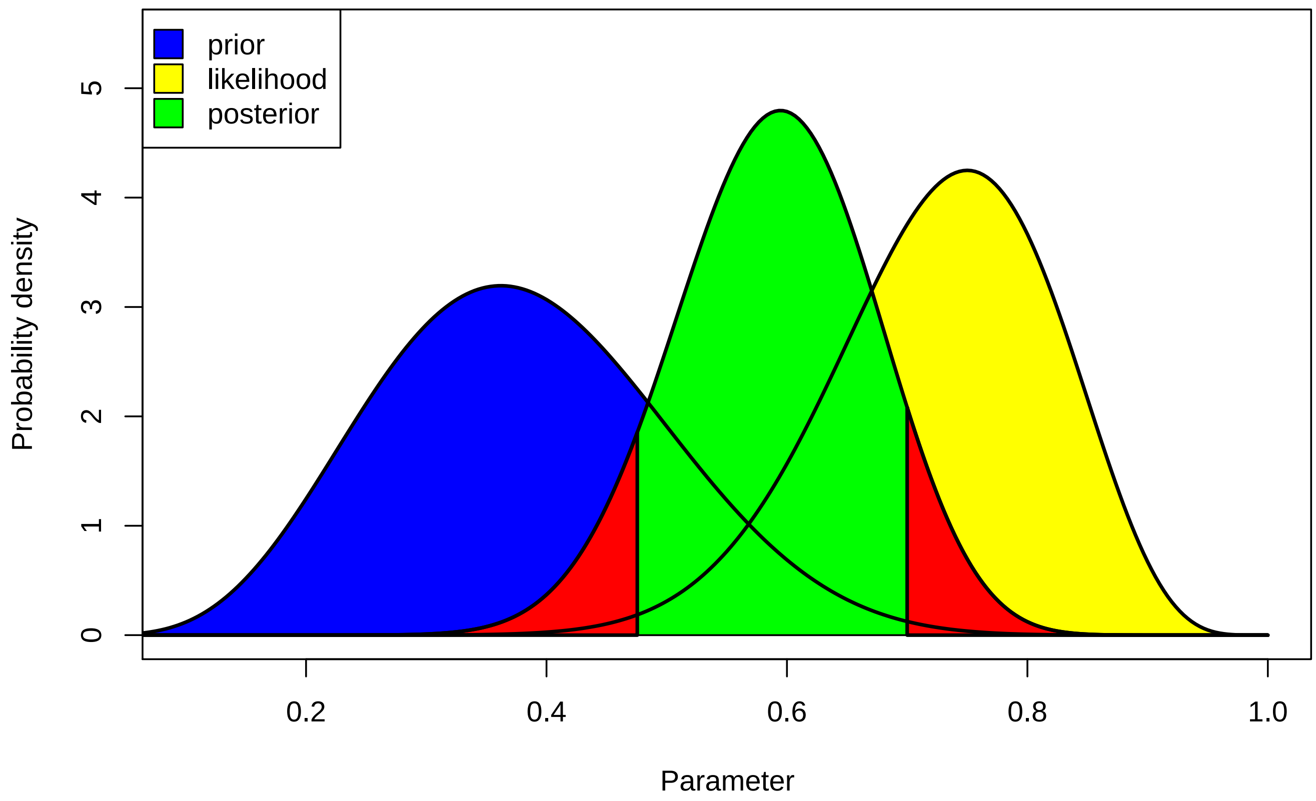
Common sense reduced to computation

Pierre-Simon, marquis de Laplace (1749–1827)
Inventor of Bayesian inference



Bayesian inference

The same example; recall posterior $\propto$ prior $\times$ likelihood;

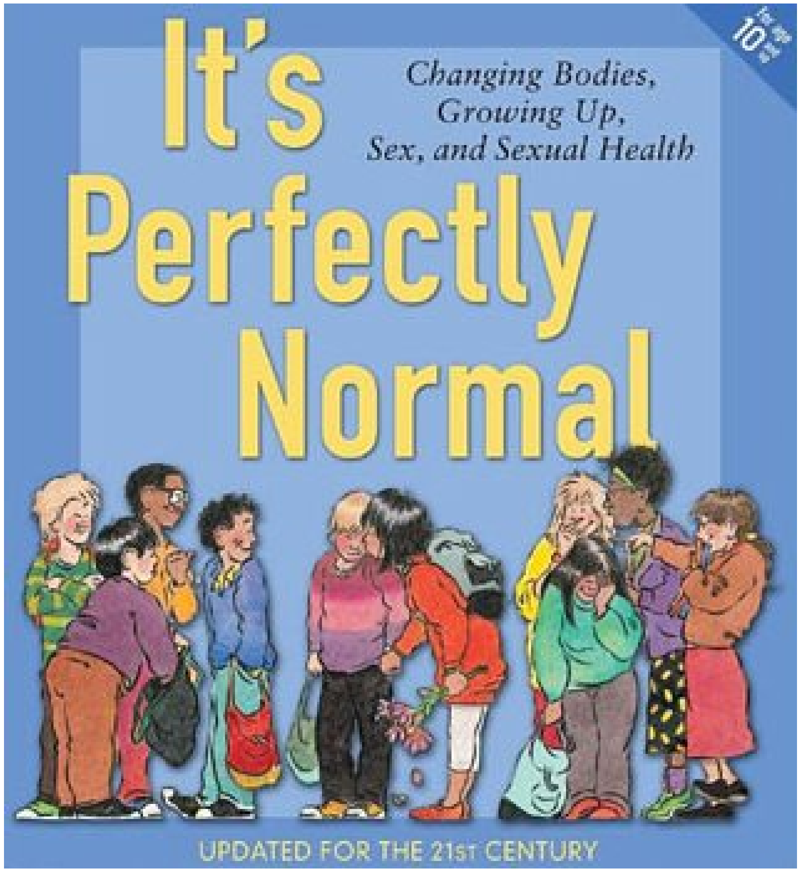


A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule

Stephen Senn, Statistician & Bayesian Skeptic (mostly)

But where do priors come from?

An important day at statistician-school?



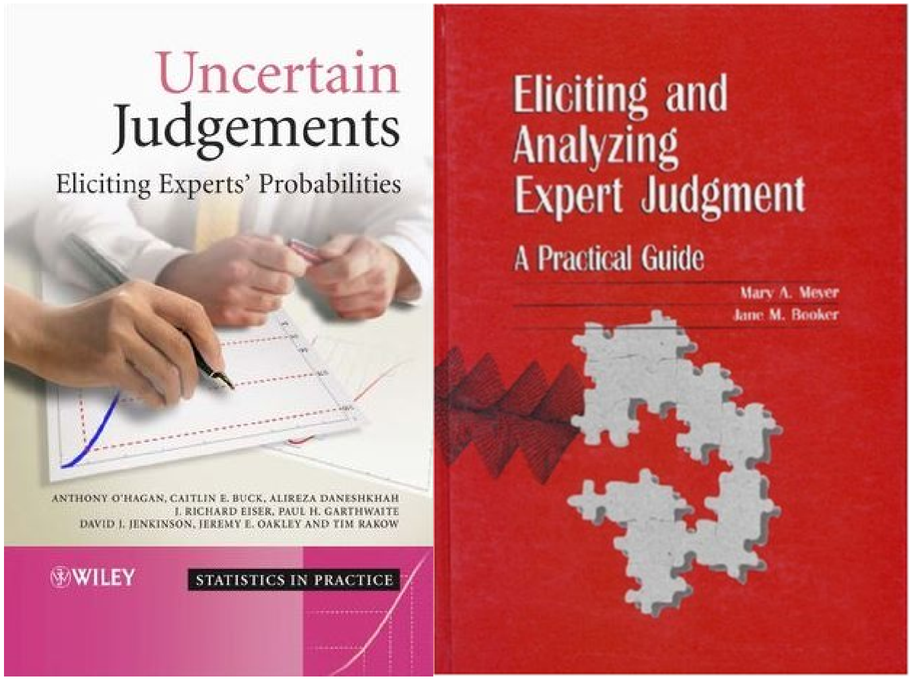
There's nothing wrong, dirty, unnatural or even *unusual* about making assumptions – carefully. Scientists & statisticians all make assumptions... even if they don't like to talk about them.

But where do priors come from?

Priors come from all data *external* to the current study, i.e. everything else.

'Boiling down' what subject-matter experts know/think is known as *eliciting* a prior.

It's not easy (see right) but here are some simple tips;

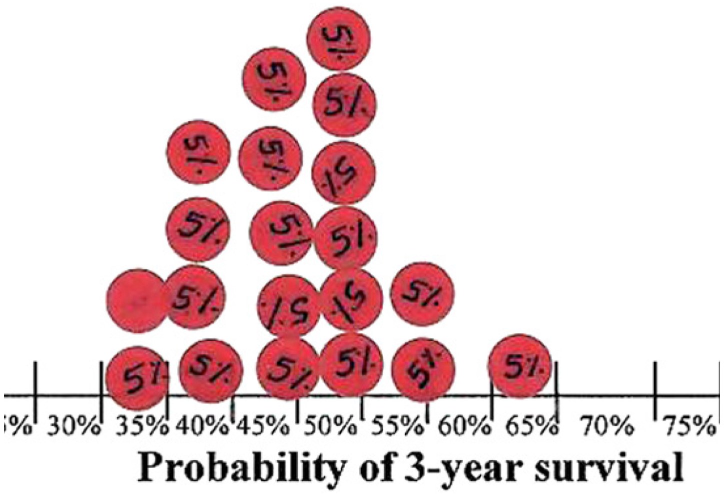


- Discuss parameters experts understand – e.g. code variables so intercept is mean outcome in people with average covariates, *not* with age=height=IQ=0
- Avoid leading questions (just as in survey design)
- The 'language' of probability is unfamiliar; help users express their uncertainty

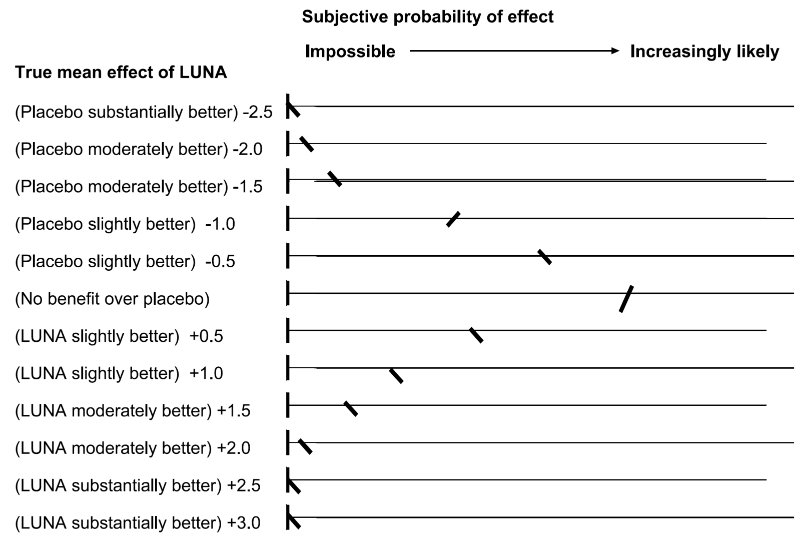
Kynn (2008, JRSSA) is a good review, describing many pitfalls.

But where do priors come from?

Ideas to help experts 'translate' to the language of probability;



Use 20×5% stickers (Johnson *et al* 2010, J Clin Epi) for prior on survival when taking warfarin

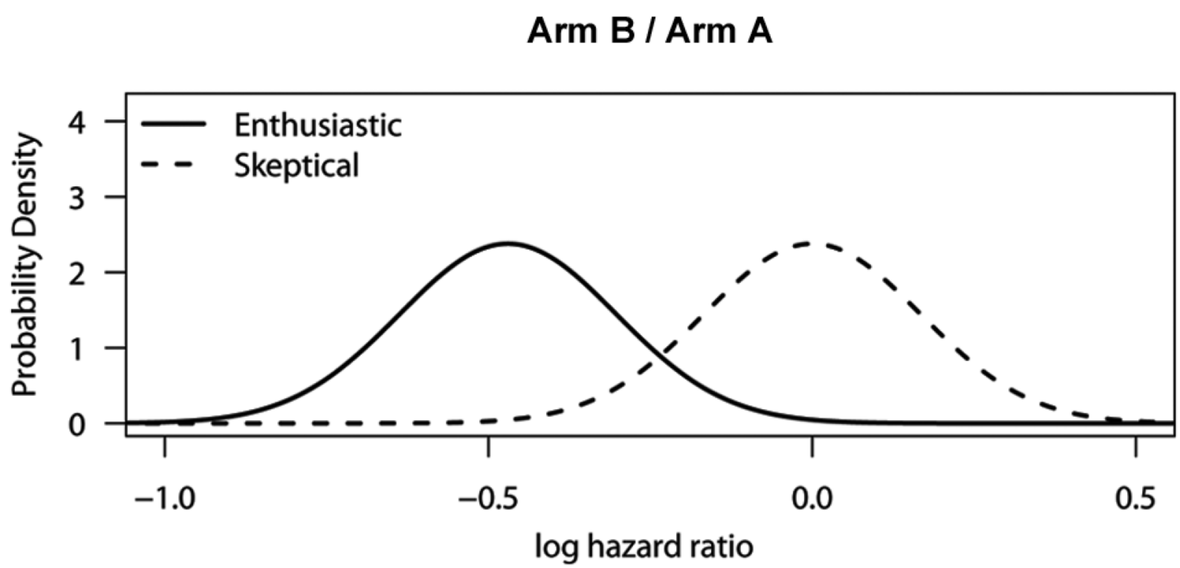


Normalize marks (Latthe *et al* 2005, J Obs Gync) for prior on pain effect of LUNA vs placebo

- Typically these 'coarse' priors are smoothed. Providing the basic shape remains, exactly how much you smooth is unlikely to be critical in practice.
- Elicitation is also *very* useful for non-Bayesian analyses – it's similar to study design & analysis planning

But where do priors come from?

If the experts disagree? Try it both ways; (Moatti, Clin Trl 2013)



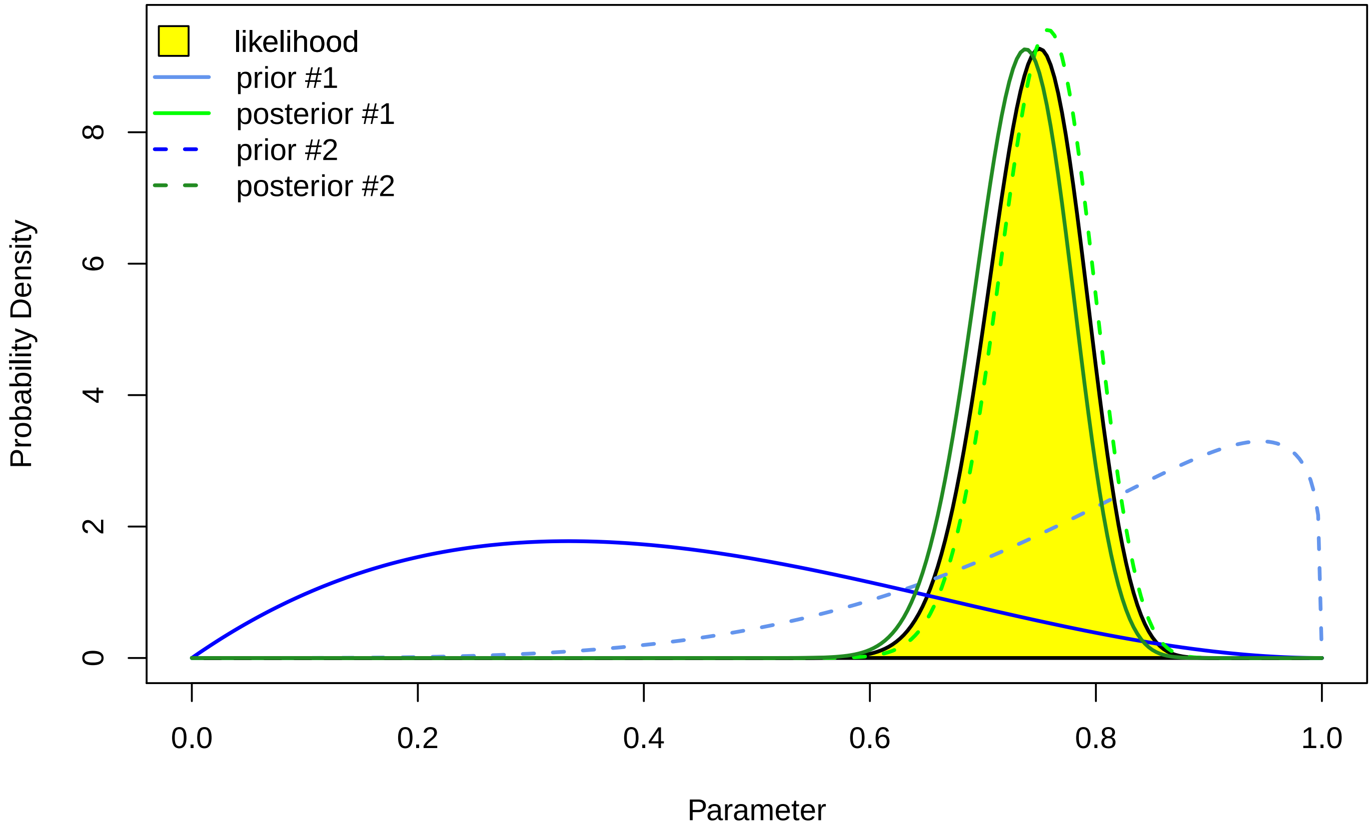
Parmer *et al* (1996, JNCI) popularized the definitions, they are now common in trials work

Known as 'Subjunctive Bayes'; if one had *this* prior and the data, *this* is the posterior one would have. If one had *that* prior... etc.

If the posteriors differ, what You believe based on the data depends, importantly, on Your prior knowledge. To convince *other* people expect to have to convince skeptics – and note that convincing [rational] skeptics is what science *is all about*.

When don't priors matter (much)?

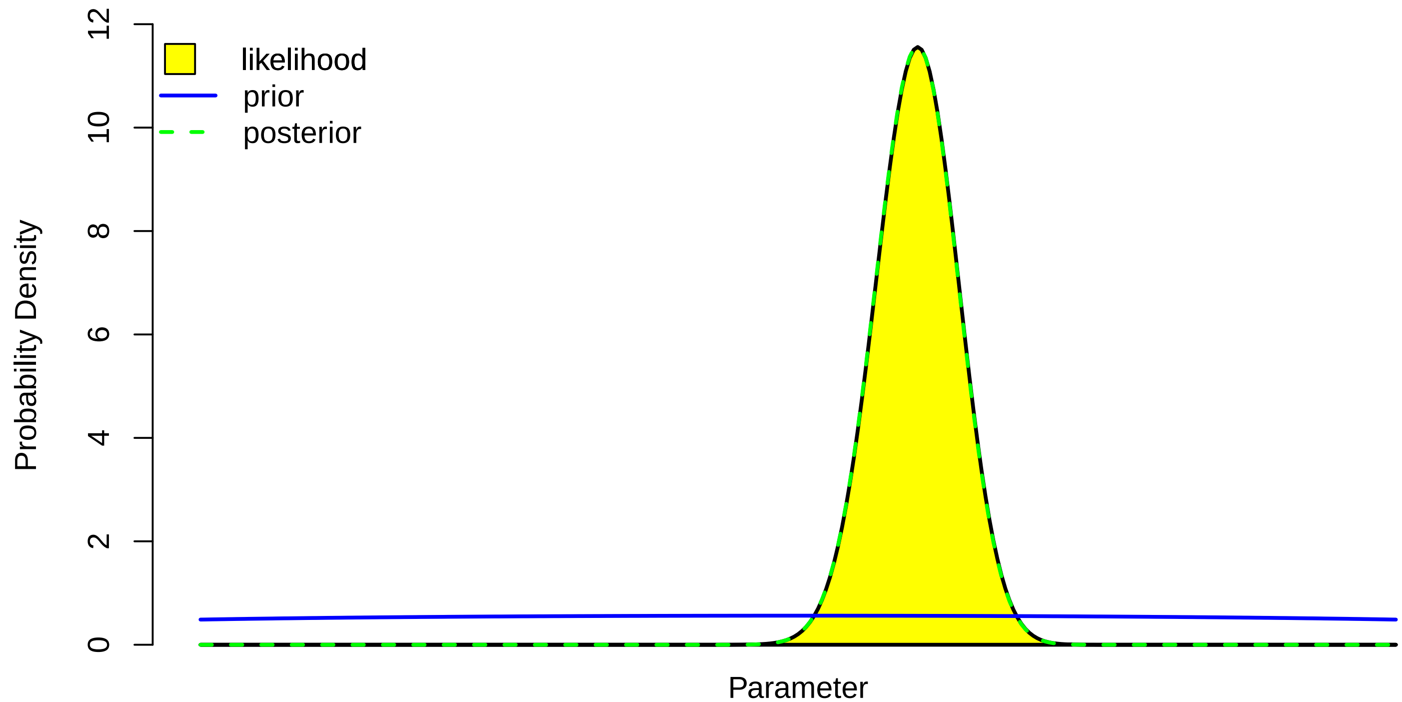
When the data provide a lot more information than the prior, this happens; (recall the stained glass color-scheme)



These priors (& many more) are *dominated* by the likelihood, and they give very similar posteriors – i.e. everyone agrees. (Phew!)

When don't priors matter (much)?

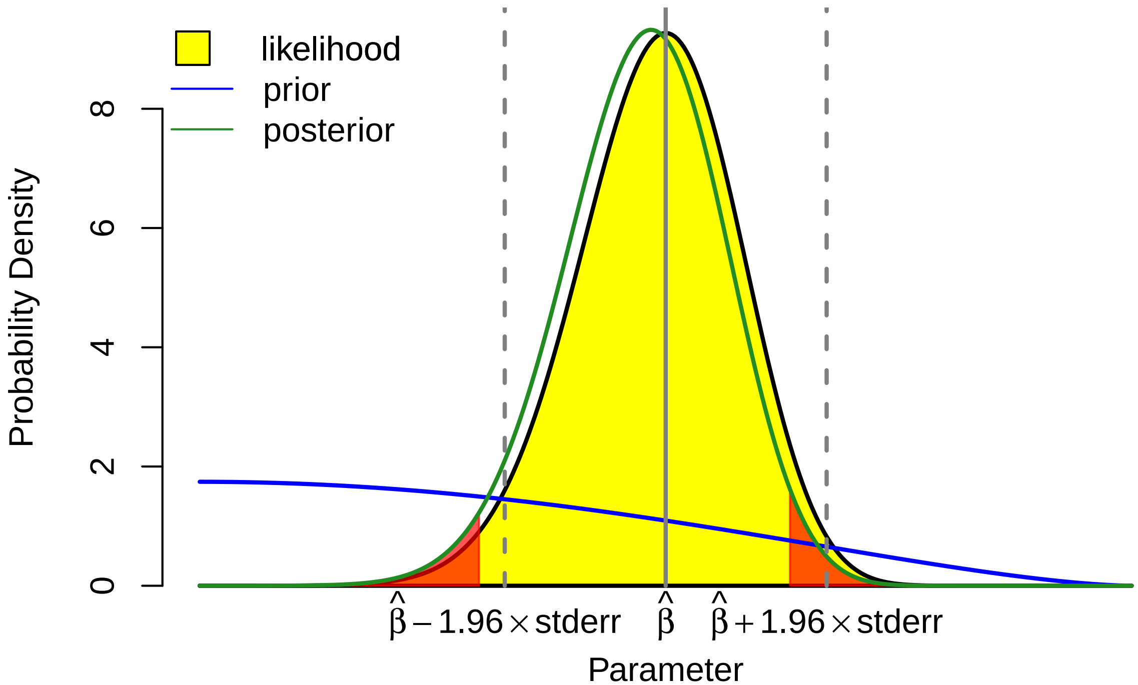
A related idea; try using very flat priors to represent ignorance;



- Flat priors do NOT actually represent ignorance! Most of their support is for *very* extreme parameter values
- For β parameters in '1st year' regression models, this idea works okay – it's more generally known as 'Objective Bayes'
- For many other situations, it doesn't, so use it carefully. (And also recall that prior elicitation is a useful exercise)

When don't priors matter (much)?

Back to having very informative data – now zoomed in;



The likelihood *alone* (yellow) gives the classic 95% confidence interval. But, to a good approximation, it goes from 2.5% to 97.5% points of Bayesian posterior (red) – a 95% *credible* interval.

- With large samples*, sane frequentist confidence intervals and sane Bayesian credible intervals are essentially identical
- With large samples*, it's actually *okay* to give Bayesian interpretations to 95% CIs, i.e. to say we have ≈95% posterior belief that the true β lies within that range

* *and some regularity conditions*

When don't priors matter (much)?

We can exploit this idea to be 'semi-Bayesian'; multiply what the likelihood-based interval says by Your prior. For Normal priors*;

$$
\begin{aligned}
\text{Prior: } \beta &\sim N(\mu_0, \sigma_0^2) \\
\text{Likelihood: } &\text{approx } N(\hat\beta, \widehat{\text{StdErr}}[\hat\beta]^2) \\
\text{Posterior: } \beta &\sim N\left(\mu_0 w + \hat\beta(1-w), \frac{1}{1/\sigma_0^2 + 1/\widehat{\text{StdErr}}[\hat\beta]^2}\right), \\
\text{where } \quad & w = \frac{1/\sigma_0^2}{1/\sigma_0^2 + 1/\widehat{\text{StdErr}}[\hat\beta]^2}
\end{aligned}
$$

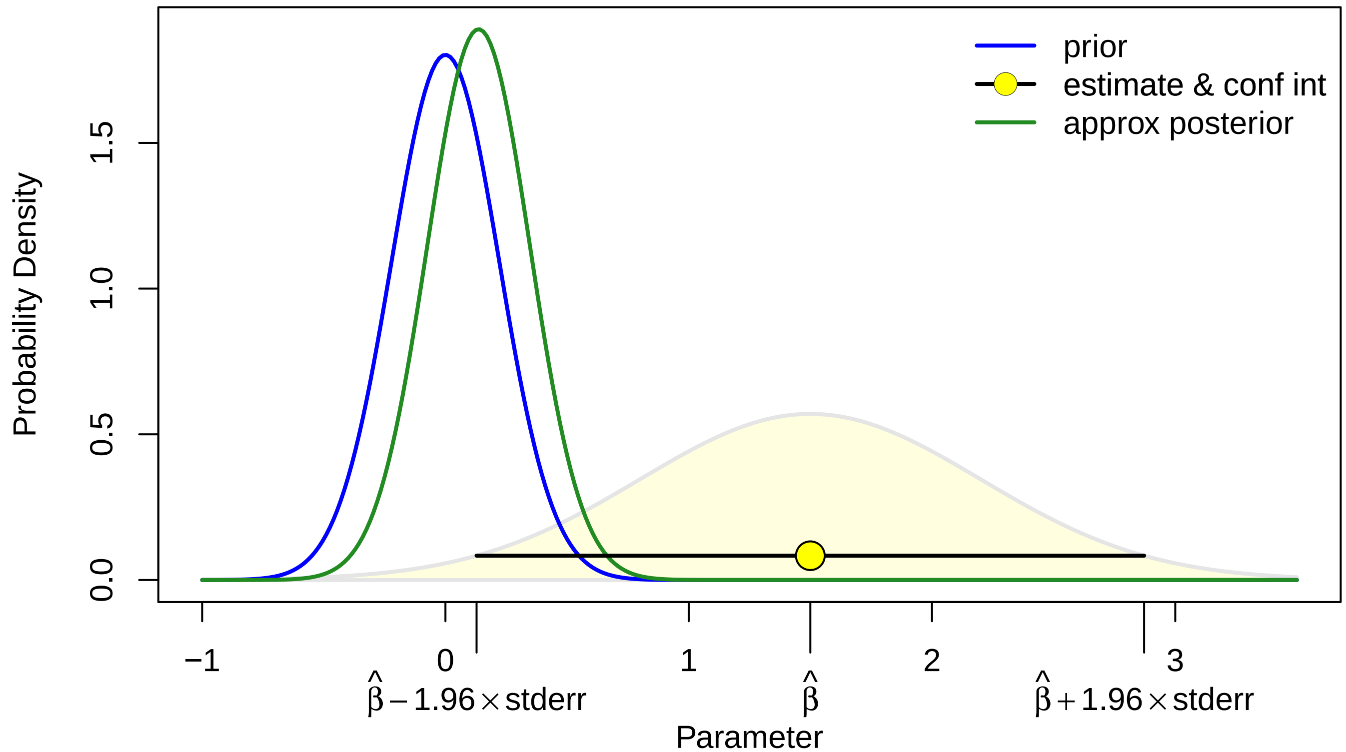
- Posterior's mean weights the prior mean (μ_0) and the classic estimate ($\hat\beta$)
- Weights ($w, 1-w$) for each reflect their precision (1/variance)
- Overall precision = sum of each source's precision

Note: these are *exactly* the same calculations as fixed-effects meta-analysis – which also computes just a sensible average.

* for non-Normal priors you'll want a computer, but it's still quick to do

When don't priors matter (much)?

Let's try it, for a prior strongly supporting small effects, and with data from an imprecise study;



- 'Textbook' classical analysis says 'reject' ($p < 0.05$, woohoo!)
- Compared to the CI, the posterior is 'shrunk' toward zero; posterior says we're sure true β is very small (& so hard to replicate) & we're unsure of its sign. So, hold the front page

When don't priors matter (much)?

Hold the front page... does that sound familiar?

Problems with the 'aggressive dissemination of noise' are a current hot topic...

THE NEW YORKER

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

BY JONAH LEHRER

DECEMBER 13, 2010

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on

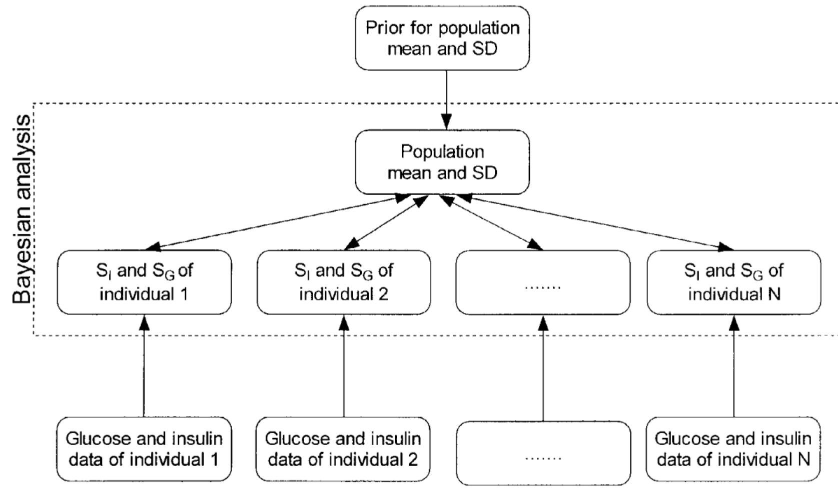
Many results that are rigorously proved and accepted start shrinking in later studies.

- In previous example, approximate Bayes helps stop over-hyping – 'full Bayes' is better still, when you can do it
- *Better* classical analysis also helps – it *can* note e.g. that study tells us little about β that's useful, not just $p < 0.05$
- No statistical approach will stop selective reporting, or fraud. Problems of biased sampling & messy data *can* be fixed (a bit) but only using background knowledge & assumptions

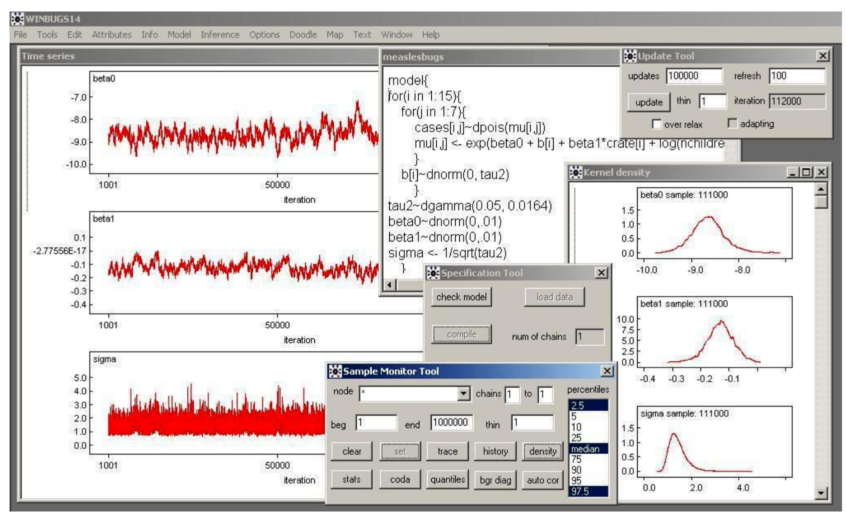
Where is Bayes commonly used?

Allowing approximate Bayes, one answer is ‘almost any analysis’. More-explicitly Bayesian arguments are often seen in;

- Hierarchical modeling
  One expert calls the classic frequentist version a “statistical no-man’s land”



- Compex models – for e.g. messy data, measurement error, multiple sources of data; fitting them is *possible* under Bayesian approaches, but perhaps still not easy



Are all classical methods Bayesian?

We've seen that, for familiar regression problems, with large n, Bayesian and frequentist ideas often don't disagree much. This is true more broadly, though for some situations statisticians haven't yet figured out the details. Some 'fancy' frequentist methods that *can* be viewed as Bayesian are;

- Fisher's exact test – its p-value is the 'tail area' of the posterior under a rather conservative prior (Altham 1969)
- Conditional logistic regression – like Bayesian analysis with particular random effects models (Severini 1999, Rice 2004)
- Robust standard errors – like Bayesian analysis of a 'trend', at least for linear regression (Szpiro *et al* 2010)

And some that can't;

- Many high-dimensional problems (shrinkage, machine-learning)
- Hypothesis testing ('Jeffrey's paradox') ...but NOT significance testing (Rice 2010... available as a talk)

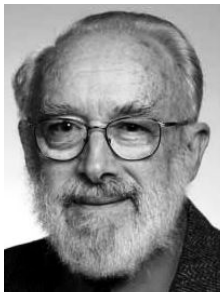
And while e.g. hierarchical modeling & multiple imputation are easier to justify in Bayesian terms, they aren't *un*frequentist.

Fight! Fight! Fight!

Two old-timers slugging out the Bayes *vs* Frequentist battle;

If [Bayesians] would only do as [Bayes] did and publish posthumously we should all be saved a lot of trouble

Maurice Kendall (1907–1983), JRSSA 1968



The only good statistics is Bayesian Statistics

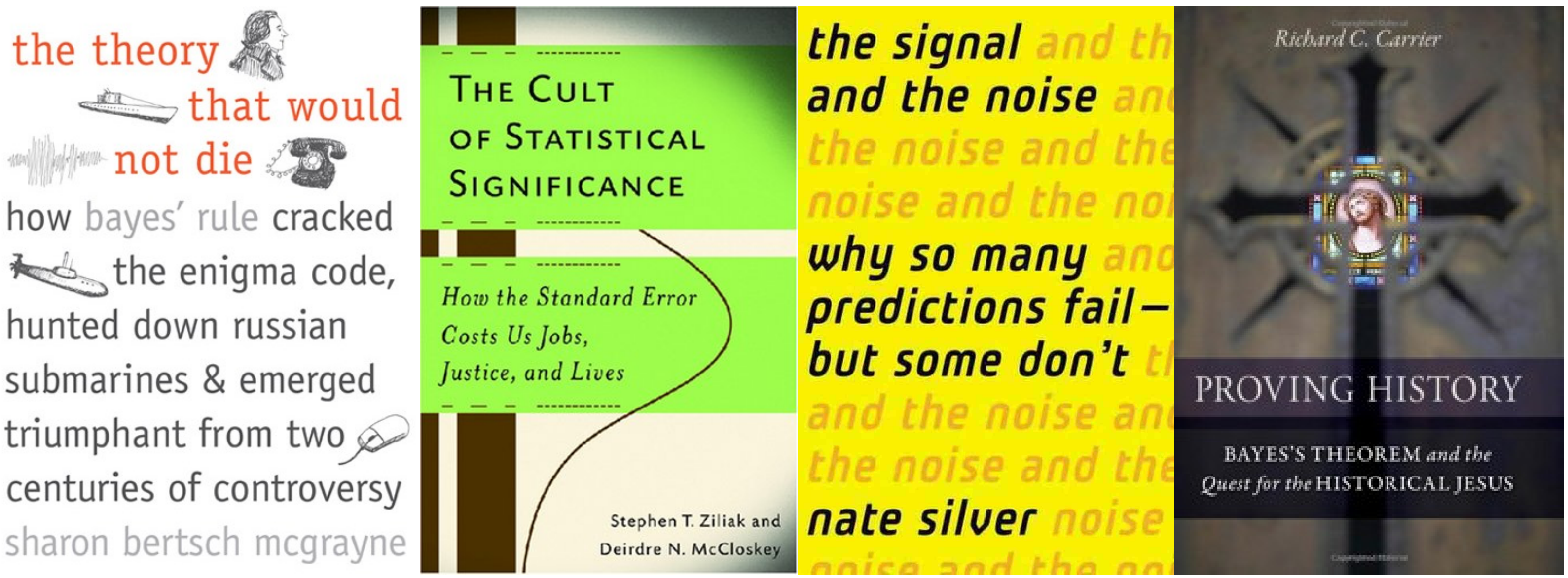
Dennis Lindley (1923–2013)
in The Future of Statistics: A Bayesian 21st Century (1975)

- For many years – until recently – Bayesian ideas in statistics* were widely dismissed, often without much thought
- Advocates of Bayes had to fight hard to be heard, leading to an 'us against the world' mentality – & predictable backlash
- Today, debates *tend* be less acrimonious, and more tolerant

* *and sometimes the statisticians who researched and used them*

Fight! Fight! Fight!

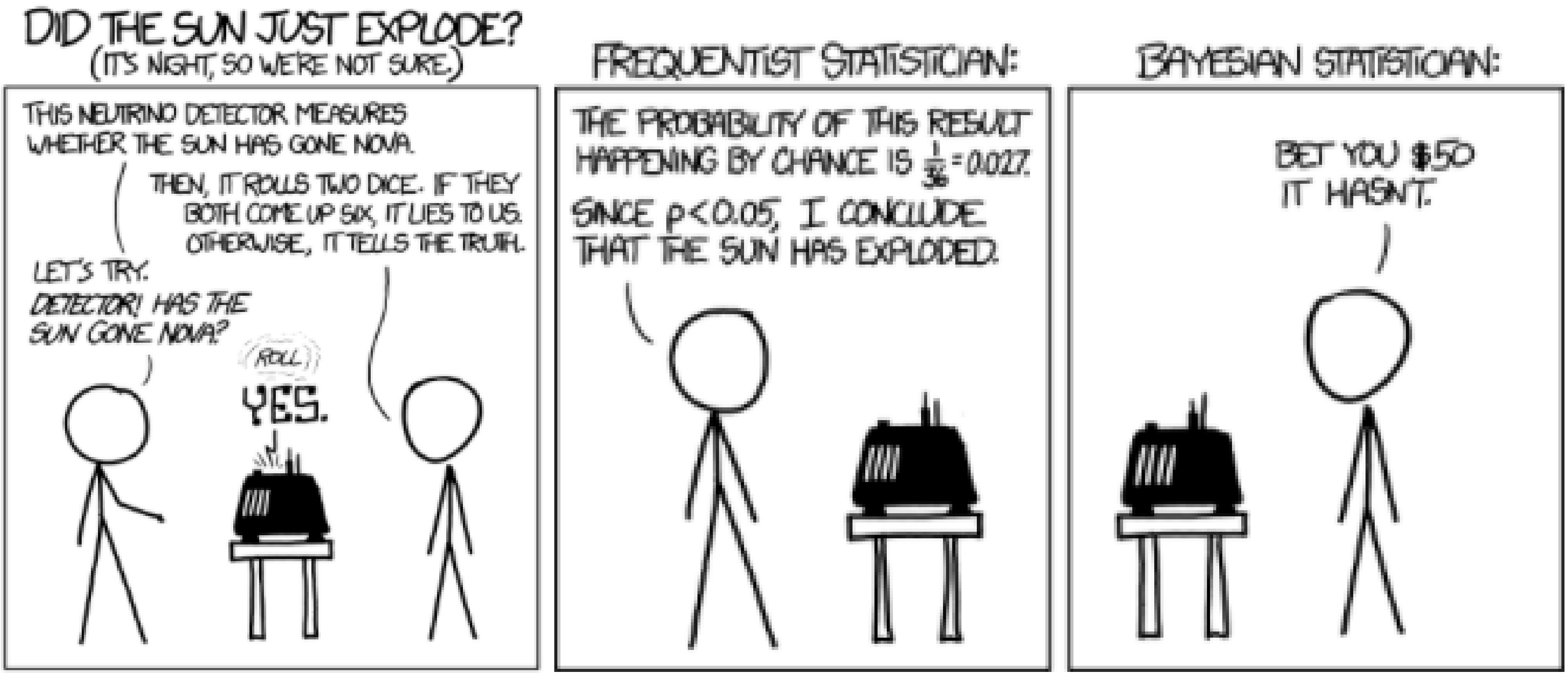
But writers of dramatic/romantic stories about Bayesian "heresy" [NYT] tend (I think) to over-egg the actual differences;



- Among those who actually understand both, it's hard to find people who totally dismiss either one
- Keen people: Vic Barnett's Comparative Statistical Inference provides the most even-handed exposition I know

Fight! Fight! Fight!

XKCD again, on Frequentists vs Bayesians;



Here, the fun relies on setting up a straw-man; p-values are not the only tools used in a *skillful* frequentist analysis.

Note: As you know, statistics can be *hard* – so it's not difficult to find examples where it's done badly, under any system.

What did you miss out?

Recall, there's a *lot* more to Bayesian statistics than I've talked about...

These books are all recommended – and/or get hold of the materials from PhD Stat/Biostat classes. You could look at;

- Model-checking, robustness to different assumptions
- Learning about multiple similar parameters (exchangeability)
- Prediction
- Missing data/causal inference
- Making decisions

– there are good Bayesian approaches to all of these, and good non-Bayesian ones too.

Summary

Bayesian statistics:

- Is useful in many settings, and you should know about it
- Is *often* not very different *in practice* from frequentist statistics; it is often helpful to think about analyses from both Bayesian and non-Bayesian points of view
- Is not reserved for hard-core mathematicians, or computer scientists, or philosophers. If you find it helpful, use it.

Wikipedia's Bayes pages aren't great. Instead, start with the linked texts, or these;

- Scholarpedia entry on Bayesian statistics
- Peter Hoff's book on Bayesian methods
- The Handbook of Probability's chapter on Bayesian statistics
- My website, for these slides
- Biost 526/Epi 540/Pharm 526, with Lurdes Inoue