Bayesian Statistics for Genetics
Lecture 1: Introduction

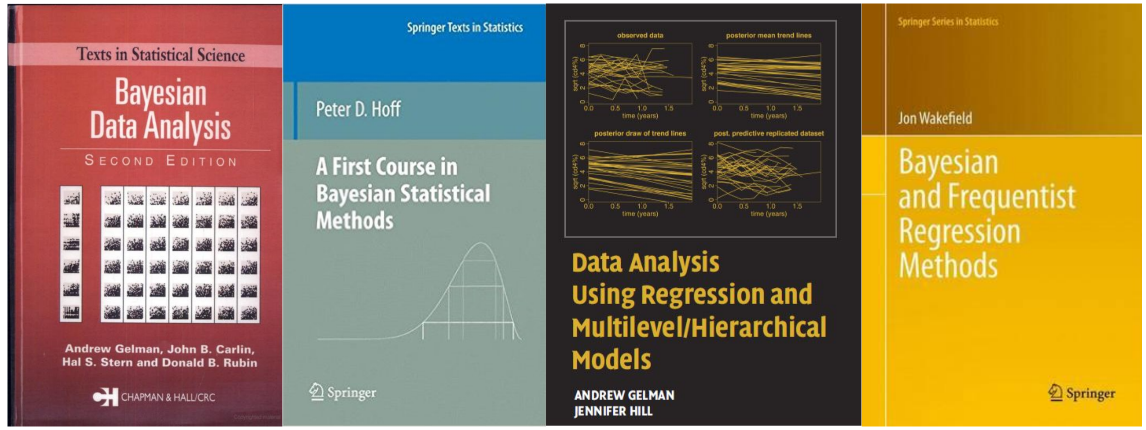
Ken Rice

Swiss Institute in Statistical Genetics

September, 2017

Overview

Just the key points from a large subject...



- What is Bayes' Rule, a.k.a. Bayes' Theorem?
- What is Bayesian inference?
- Where can Bayesian inference be helpful?
- How does it differ from frequentist inference?

Note: *other* literature contains many pro- and anti-Bayesian polemics, many of which are ill-informed and unhelpful. We 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.

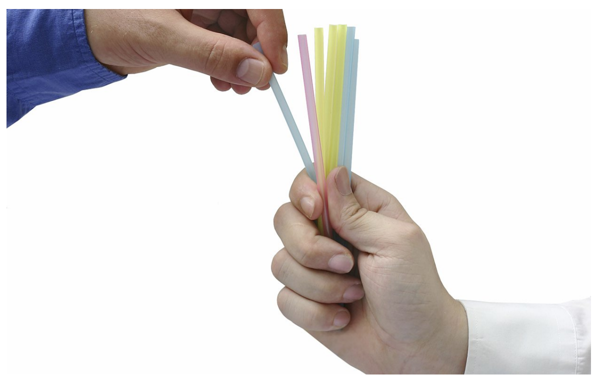
There are different ways of tackling statistical problems, too.

Bayes' Theorem

Before we get to Bayesian statistics, Bayes' *Theorem* is a result from *probability*. Probability is familiar to most people through games of chance;

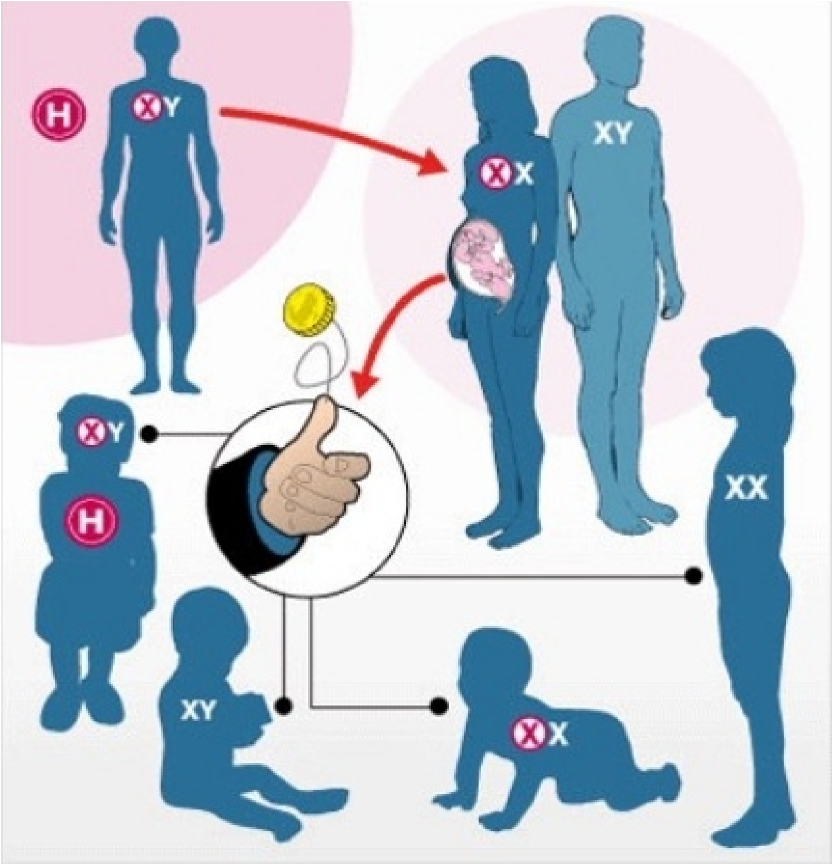







Bayes' Theorem

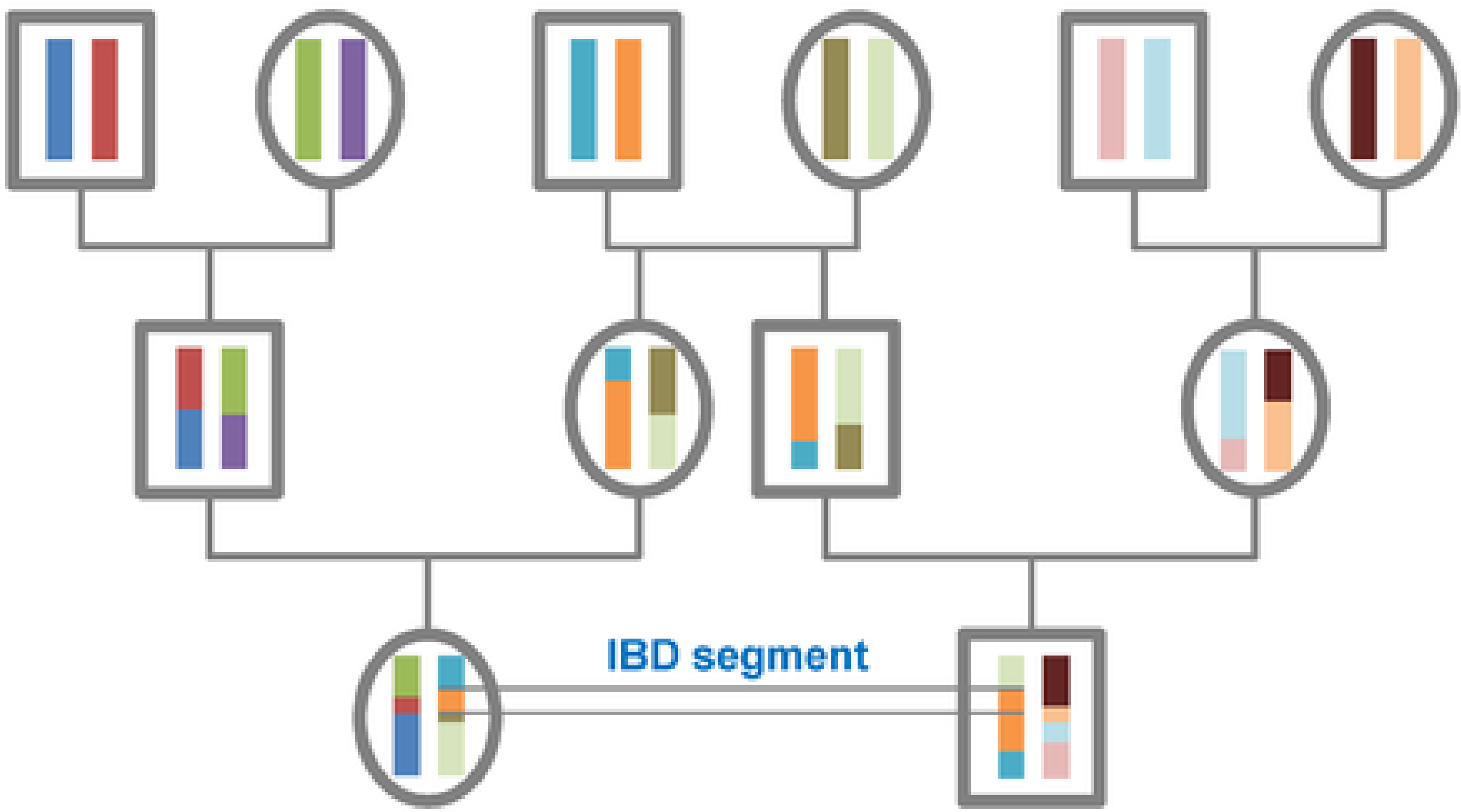
These ideas occur naturally in genetics;



'Mendelian inheritance' means that, at conception, a biological coin toss determines which parental alleles are passed on.

Bayes' Theorem

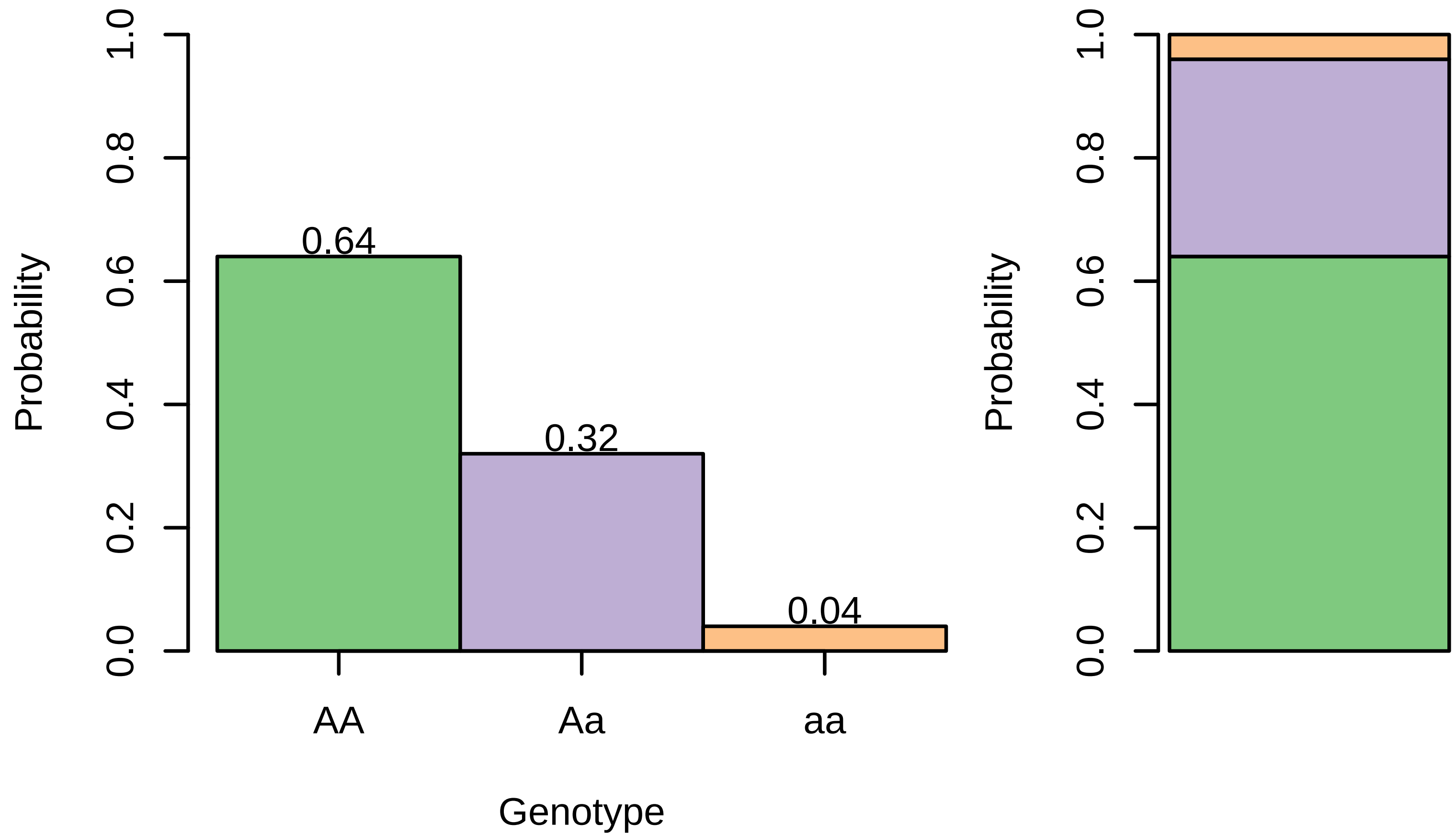
These ideas occur naturally in genetics;



The probability of being 'identical by descent' at any locus depends on the pedigree's genotypes, and structure.

Bayes' Theorem

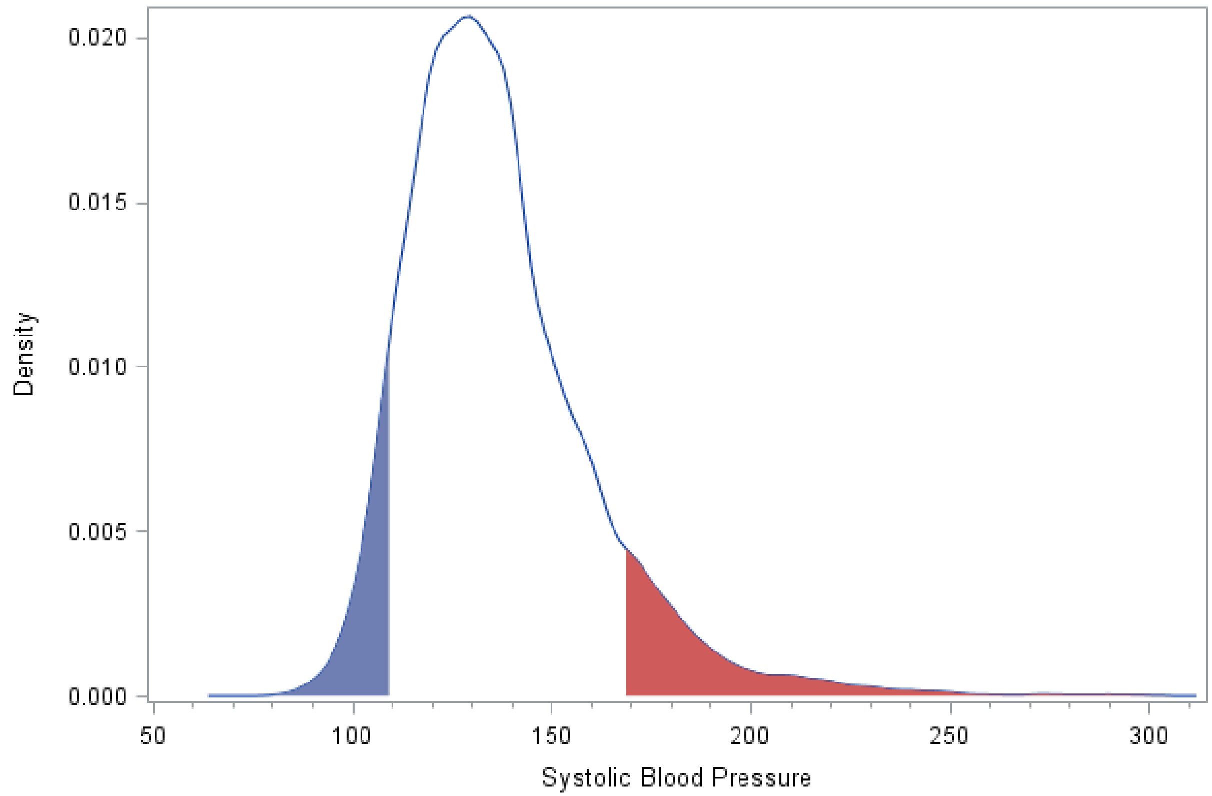
In most studies, "random" means sampling from a population;



Each person randomly-chosen to be genotyped could be AA/Aa/aa, with particular probabilities. Here, having at least one copy of the 'a' allele happens with probability 0.32+0.04=0.36, i.e. 36%.

Bayes' Theorem

Traits can also be random;



In a *density function*, we get the probability of certain sets (e.g. of a randomly-selected adult SBP $>$170mmHg or $<$110mmHg) by evaluating the corresponding **area**.

Bayes' Theorem

There are 'rules' of probability. Denoting the density at outcome y as $p(y)$;

- The total probability of all possible outcomes is 1 - so densities integrate to one;

  $$\int_{\mathcal{Y}} p(y)dy = 1,$$

  where $\mathcal{Y}$ denotes the set of all possible outcomes
- For any $a < b$ in $\mathcal{Y}$,

  $$\mathbb{P}[\,Y \in (a,b)\,] = \int_a^b p(y)dy$$
- For general events;

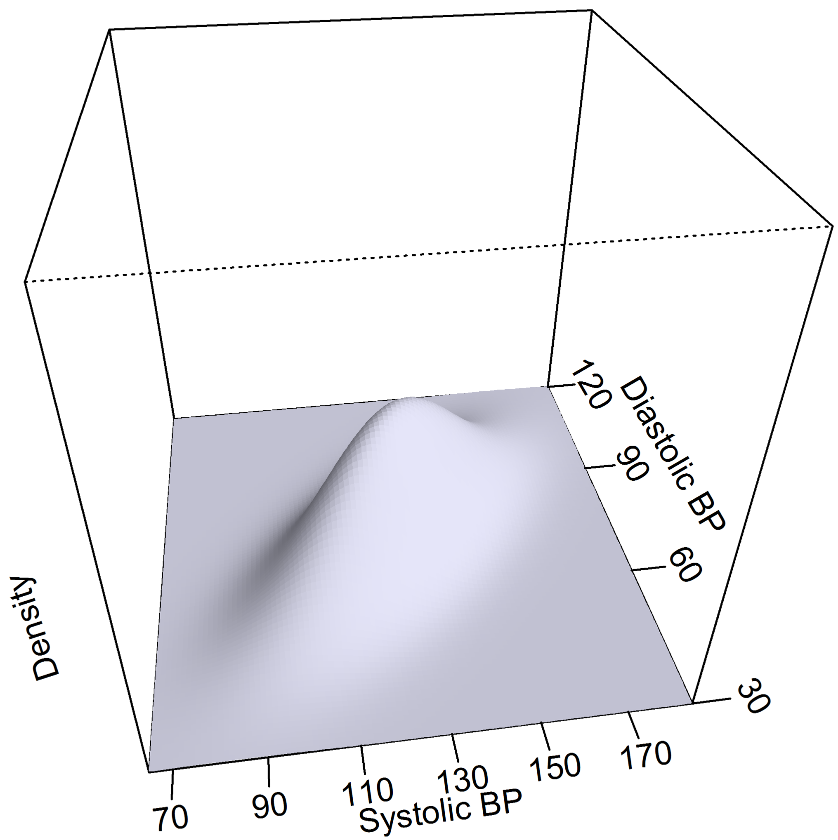
  $$\mathbb{P}[\,Y \in \mathcal{Y}_0\,] = \int_{\mathcal{Y}_0} p(y)dy,$$

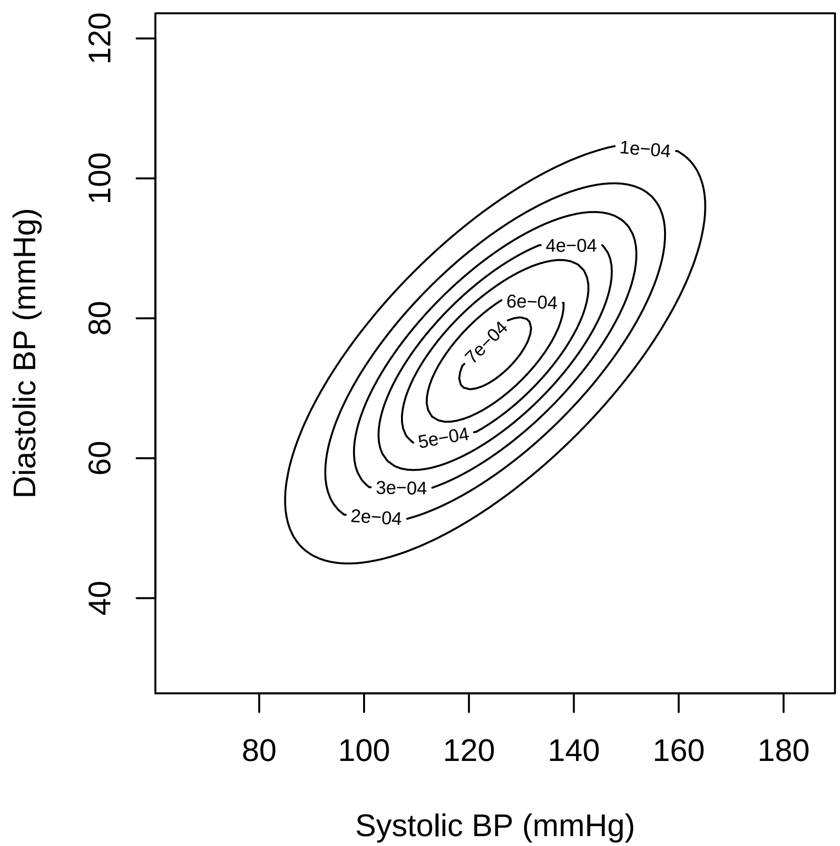
  where $\mathcal{Y}_0$ is some subset of the possible outcomes $\mathcal{Y}$

(For discrete events, replace integration by addition if you prefer)

Bayes' Theorem

For two random variables, the density is a surface;

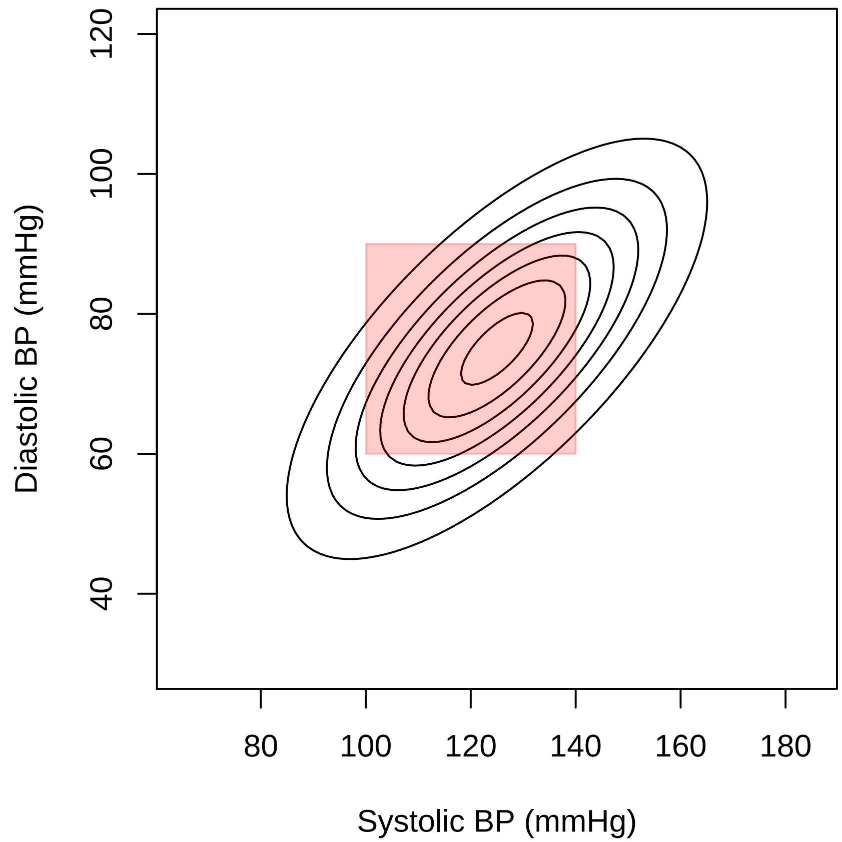


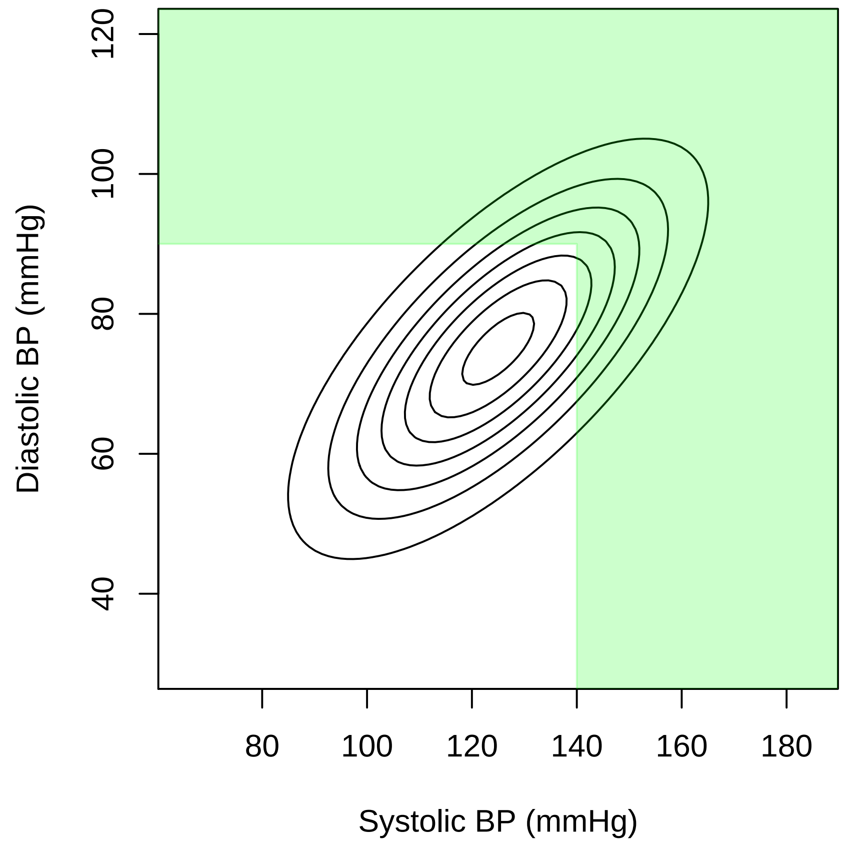


... where the total 'volume' is 1, i.e. $\int_{\mathcal{X},\mathcal{Y}} p(x,y)dxdy = 1$.

Bayes' Theorem

To get the probability of outcomes in a region we again integrate;



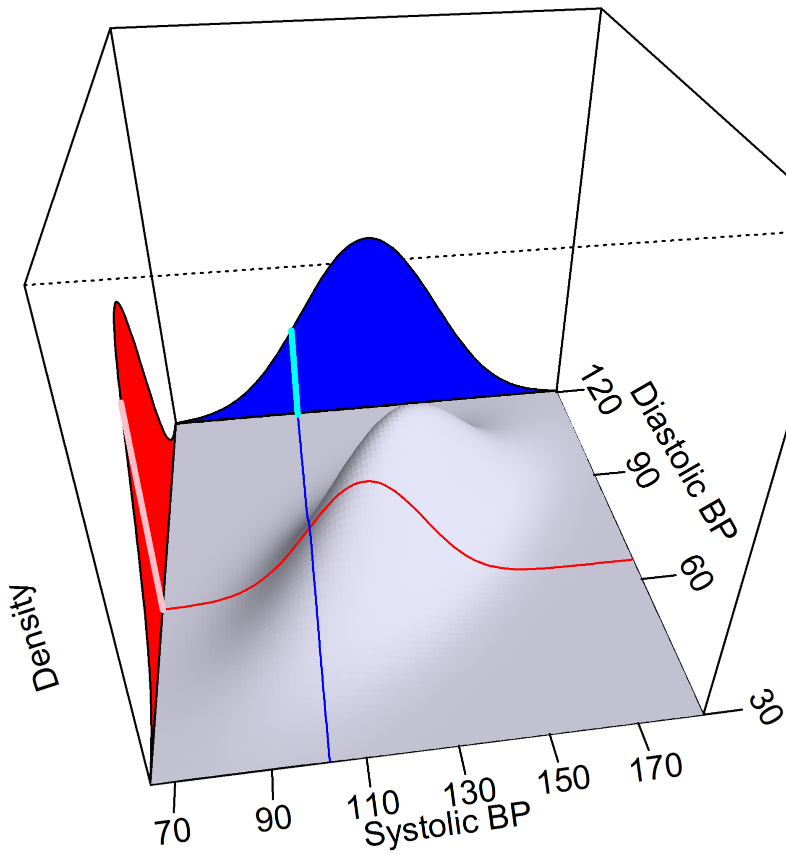


$$\mathbb{P}\left[\begin{array}{c} 100 < \text{SBP} < 140 \\ \& \\ 60 < \text{DBP} < 90 \end{array}\right] \approx 0.52$$

$$\mathbb{P}\left[\begin{array}{c} \text{SBP} > 140 \\ \text{OR} \\ \text{DBP} > 90 \end{array}\right] \approx 0.28$$

Bayes' Theorem

For continuous variables (say systolic and diastolic blood pressure) think of *conditional densities* as 'slices' through the distribution;



Formally,

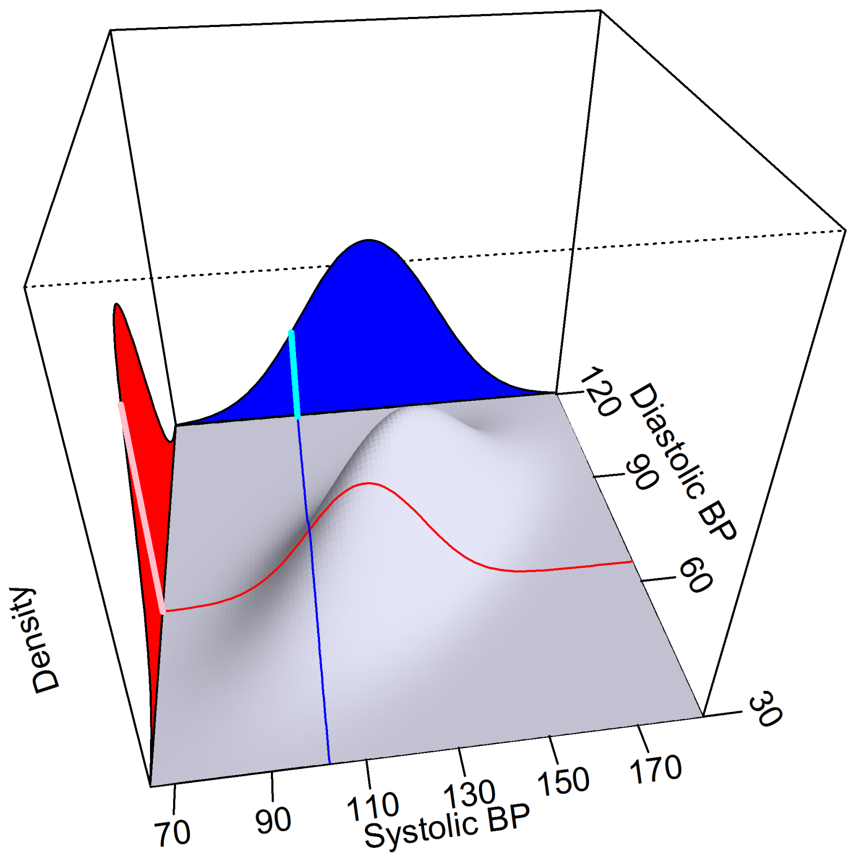
$$
\begin{aligned}
p(x|y=y_0) &= p(x,y_0)\Big/\int_{\mathcal{X}} p(x,y_0)dx \\
p(y|x=x_0) &= p(x_0,y)\Big/\int_{\mathcal{Y}} p(x_0,y)dy,
\end{aligned}
$$

and we often write these as just $p(x|y)$, $p(y|x)$. Also, the *marginal densities* (shaded curves) are given by

$$
\begin{aligned}
p(x) &= \int_{\mathcal{Y}} p(x,y)dy \\
p(y) &= \int_{\mathcal{X}} p(x,y)dx.
\end{aligned}
$$

Bayes' Theorem

Bayes' theorem connects different conditional distributions –



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

$$p(x|y) = p(y|x)\frac{p(x)}{p(y)}.$$

Because we know $p(x|y)$ must integrate to one, we can also write this as

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

Bayes' Theorem states that the conditional density is proportional to the marginal *scaled by* the other conditional density.

Bayesian statistics

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

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

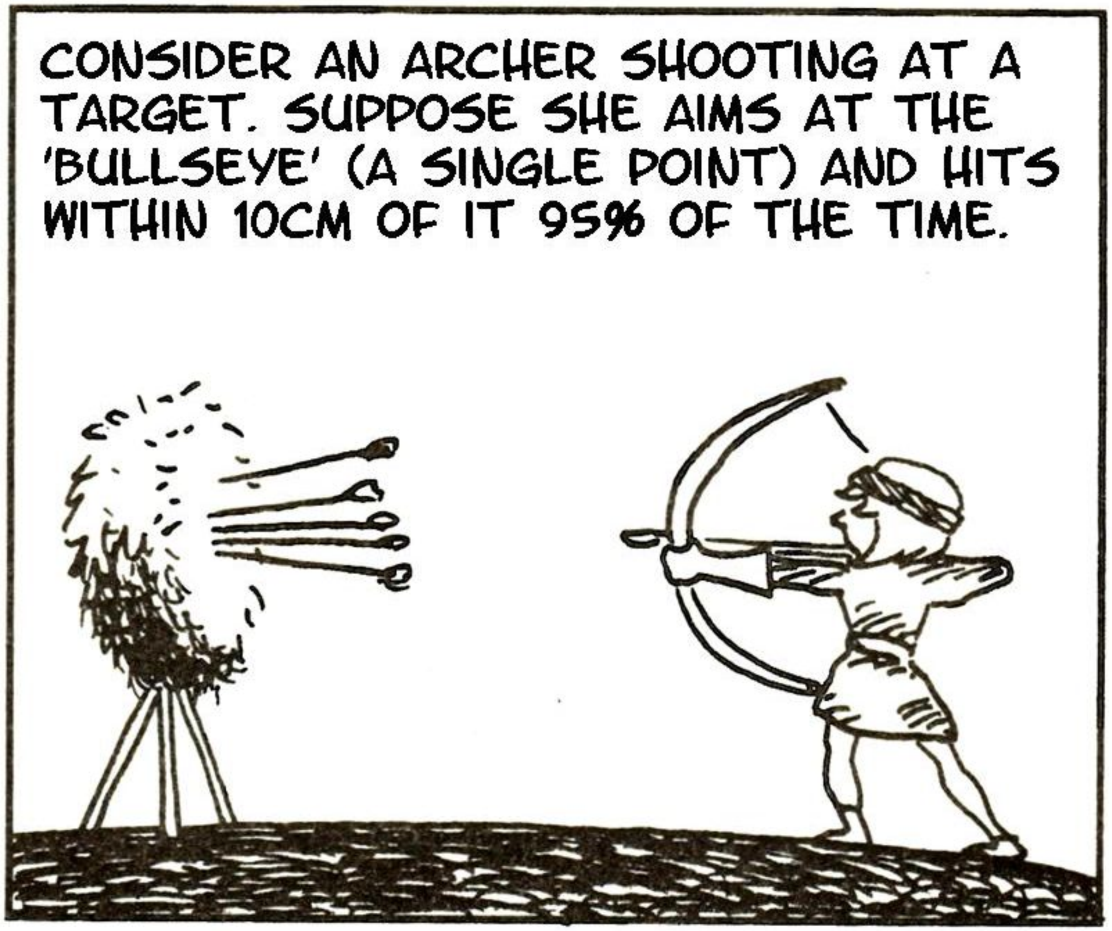
In *addition* to describing random variables, Bayesian statistics uses the 'language' of probability to describe what is known about unknown parameters.

Note: Frequentist statistics , 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.

Bayesian inference

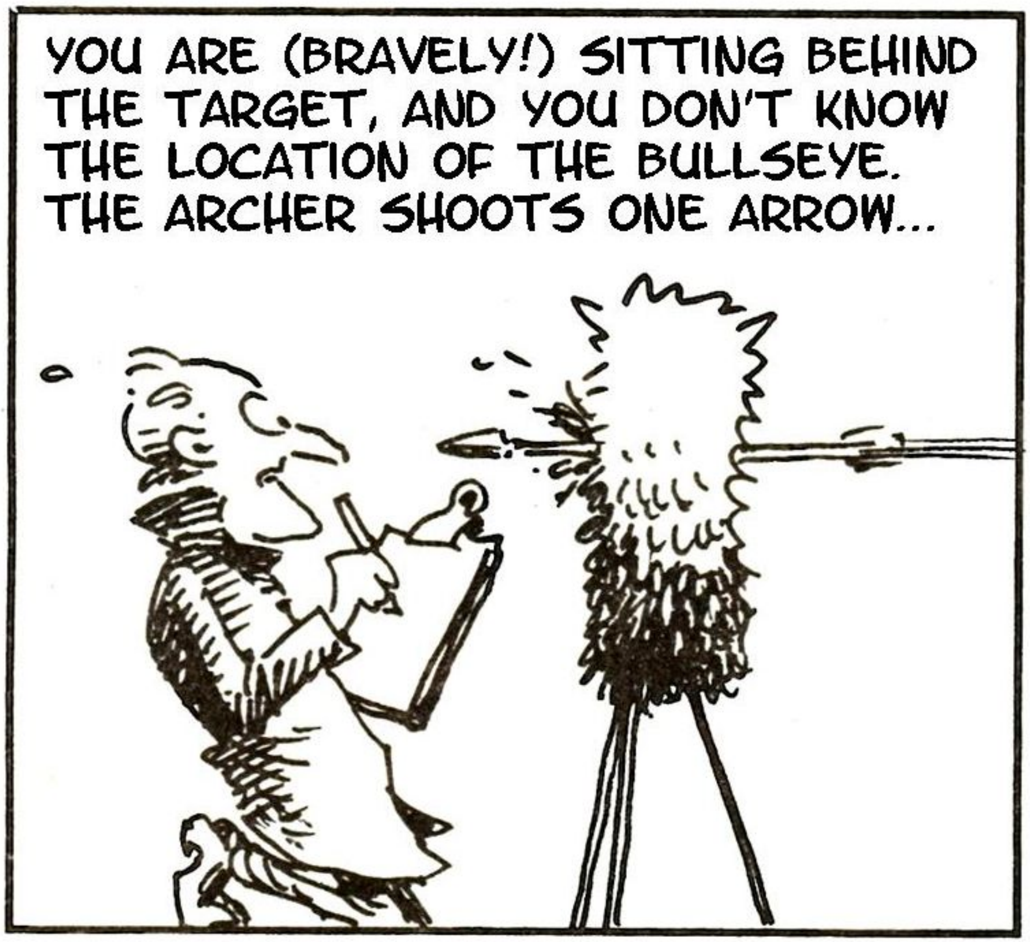
How does it work? Let's take aim...



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

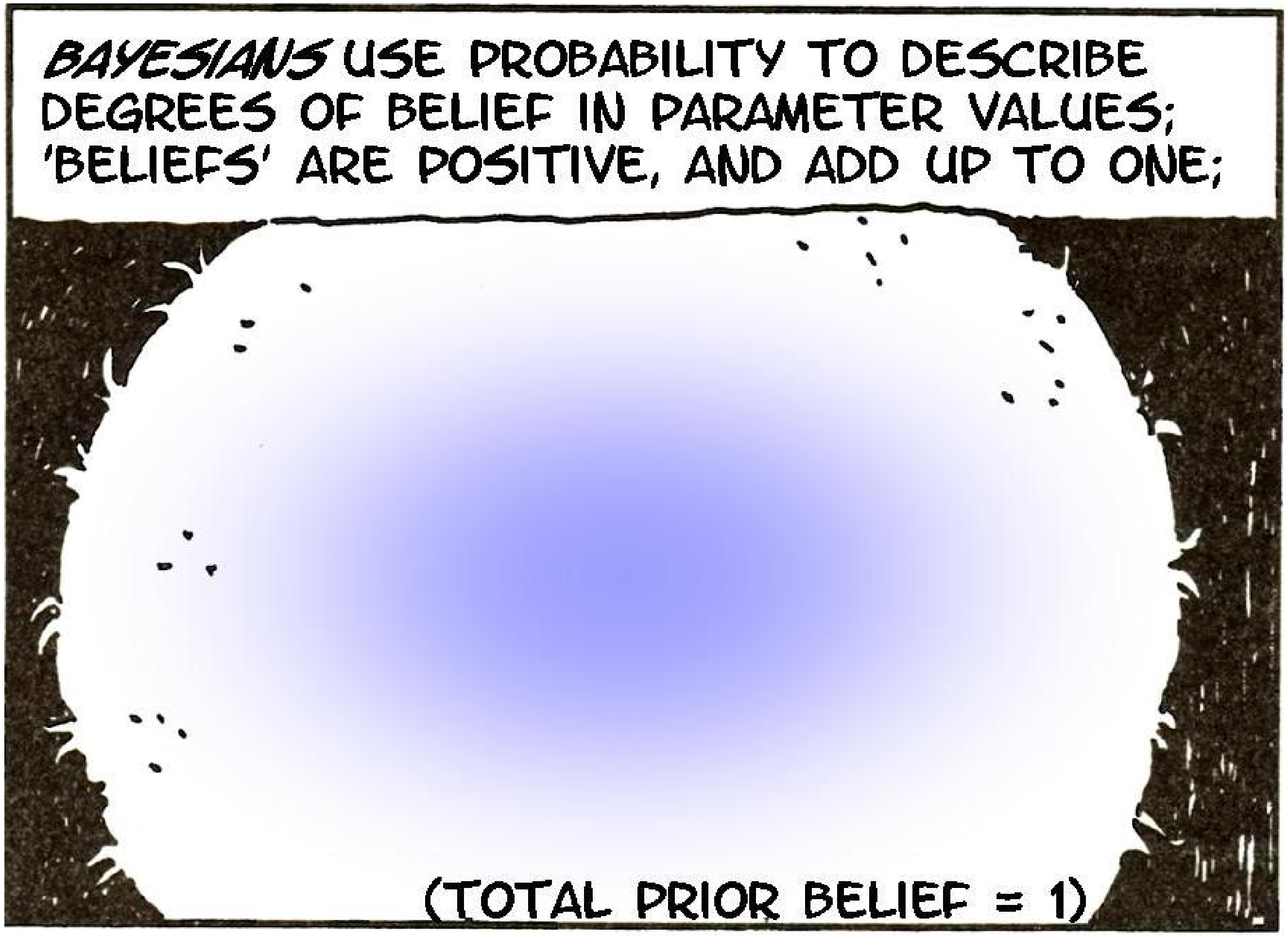
Bayesian inference

How does it work? Let's take aim...



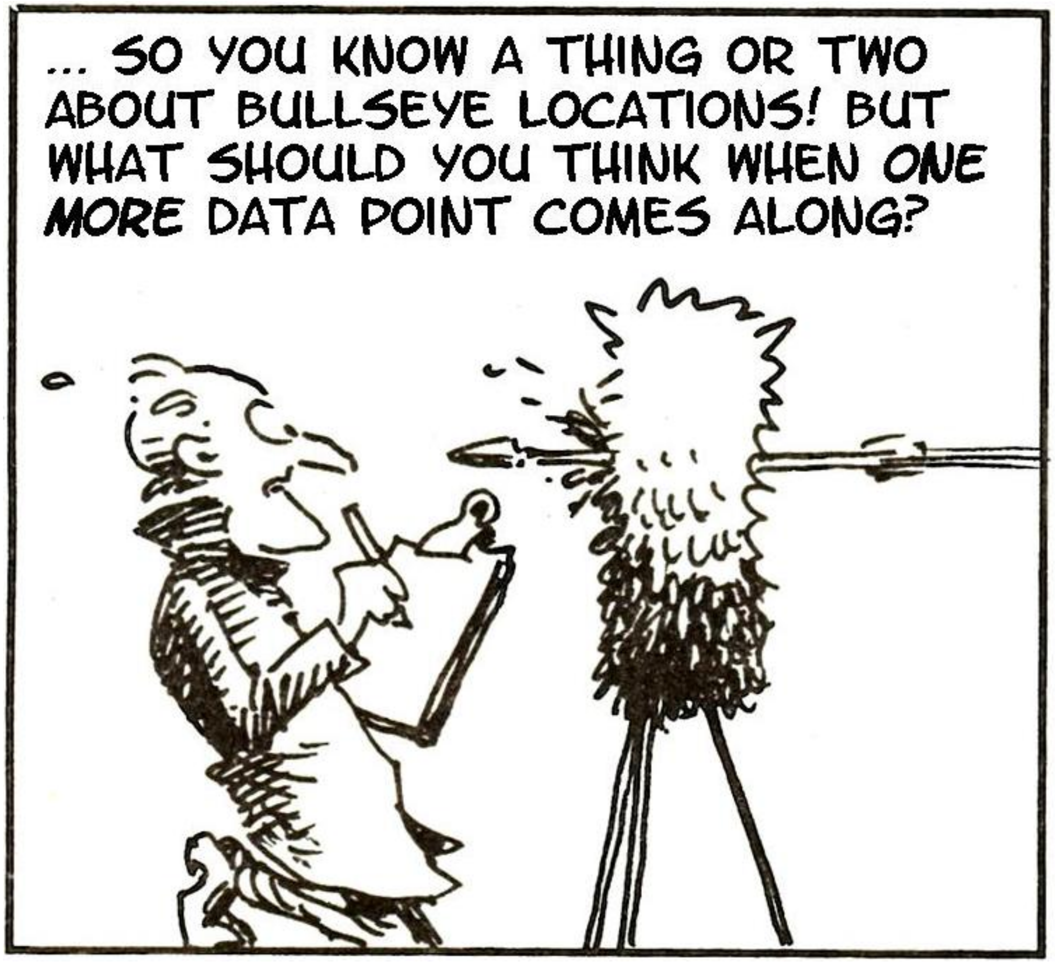
Bayesian inference

You don't know the location **exactly**, but do have some ideas...



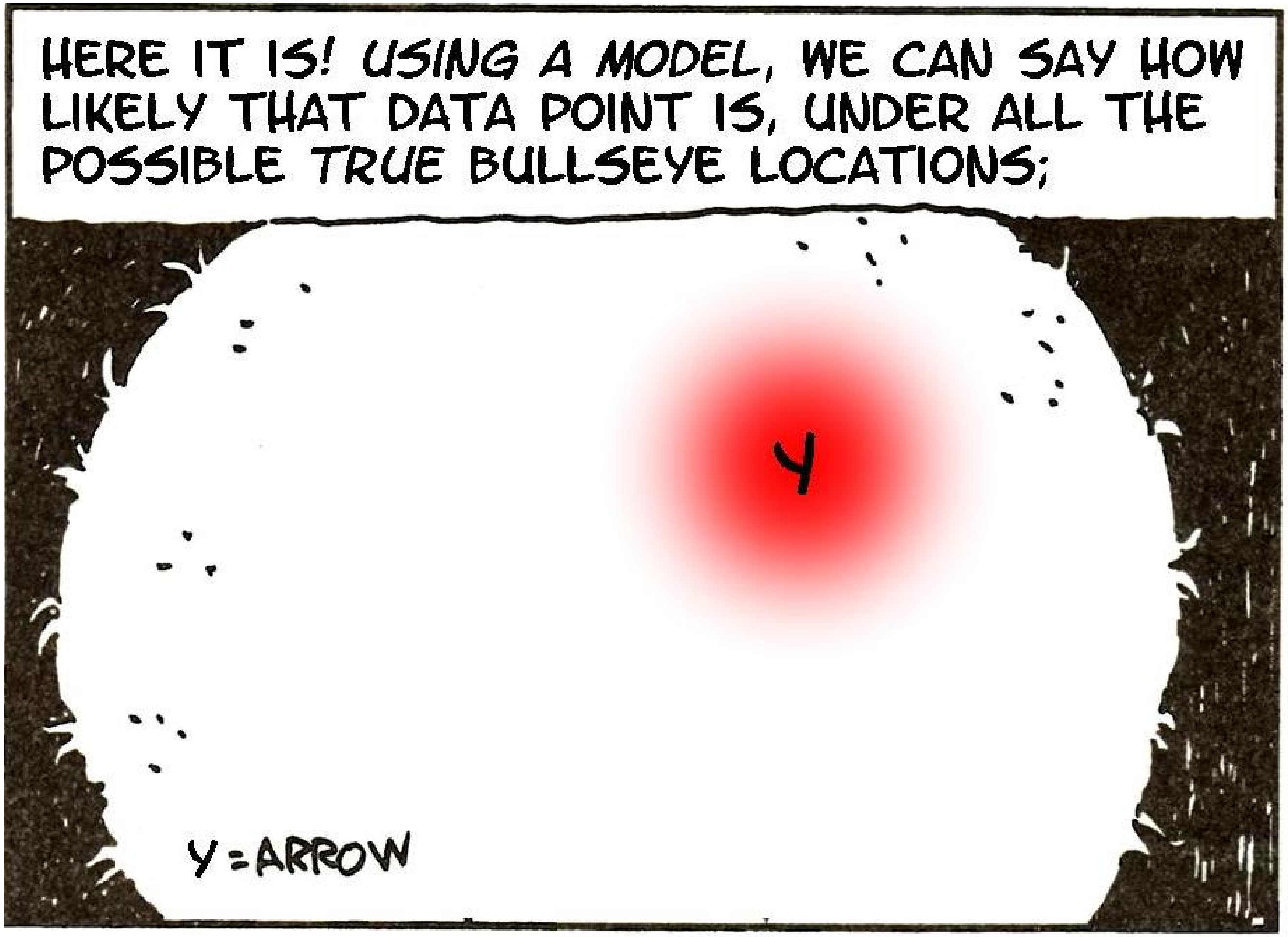
Bayesian inference

You don't know the location **exactly**, but do have some ideas...



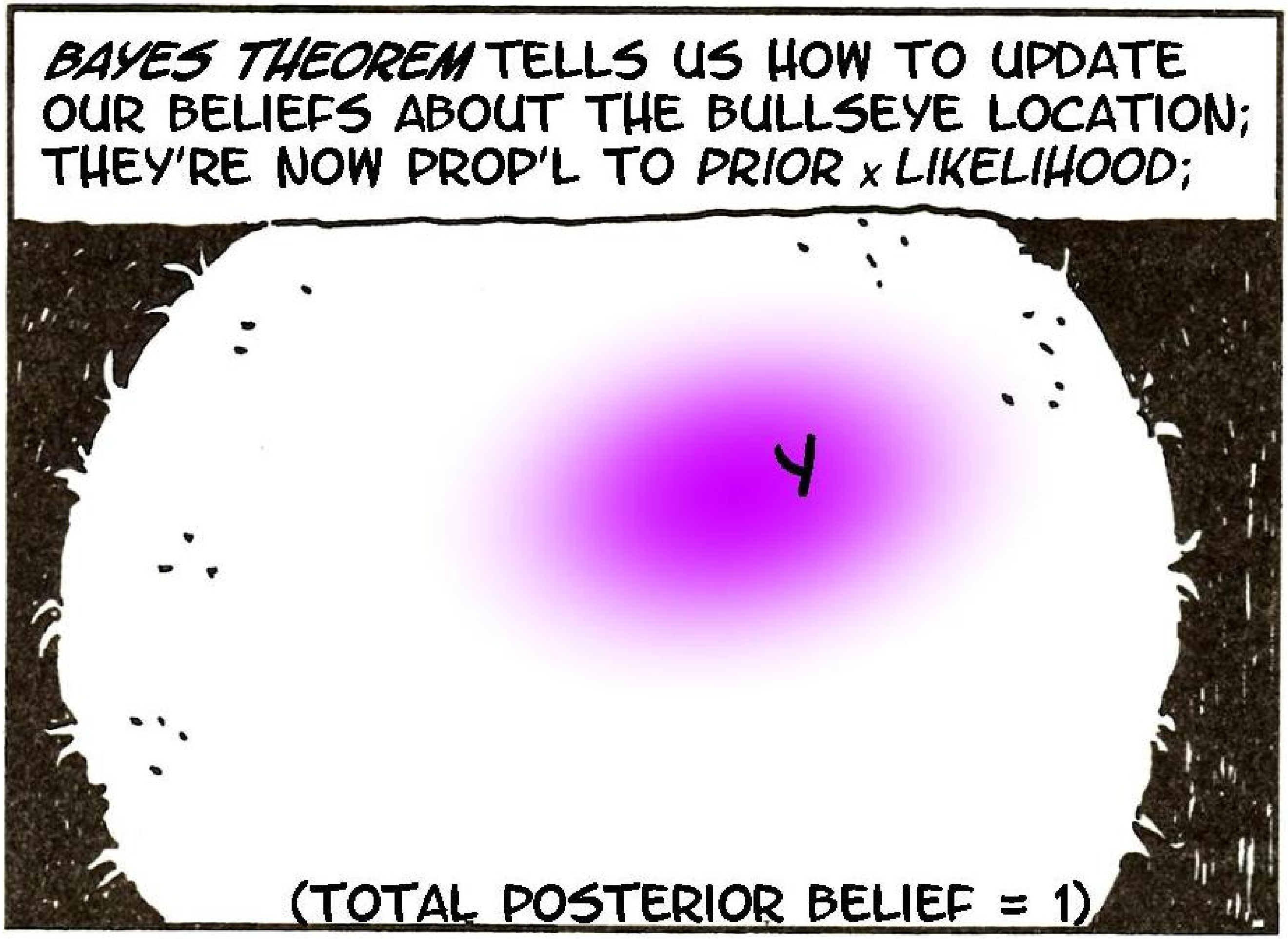
Bayesian inference

What to do when the data comes along?



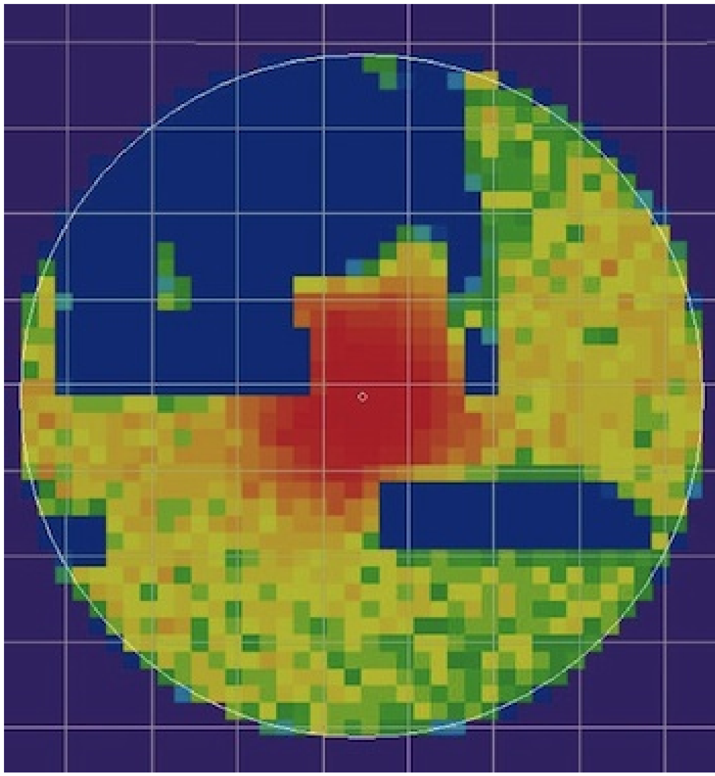
Bayesian inference

What to do when the data comes along?



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 $\boldsymbol{\theta}$, excluding the information in the data – denoted $p(\boldsymbol{\theta})$
- **Likelihood:** based on modeling assumptions, how (relatively) likely the data $\boldsymbol{y}$ are *if* the truth is $\boldsymbol{\theta}$ – denoted $p(\boldsymbol{y}|\boldsymbol{\theta})$

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;

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior}. \end{aligned}$$

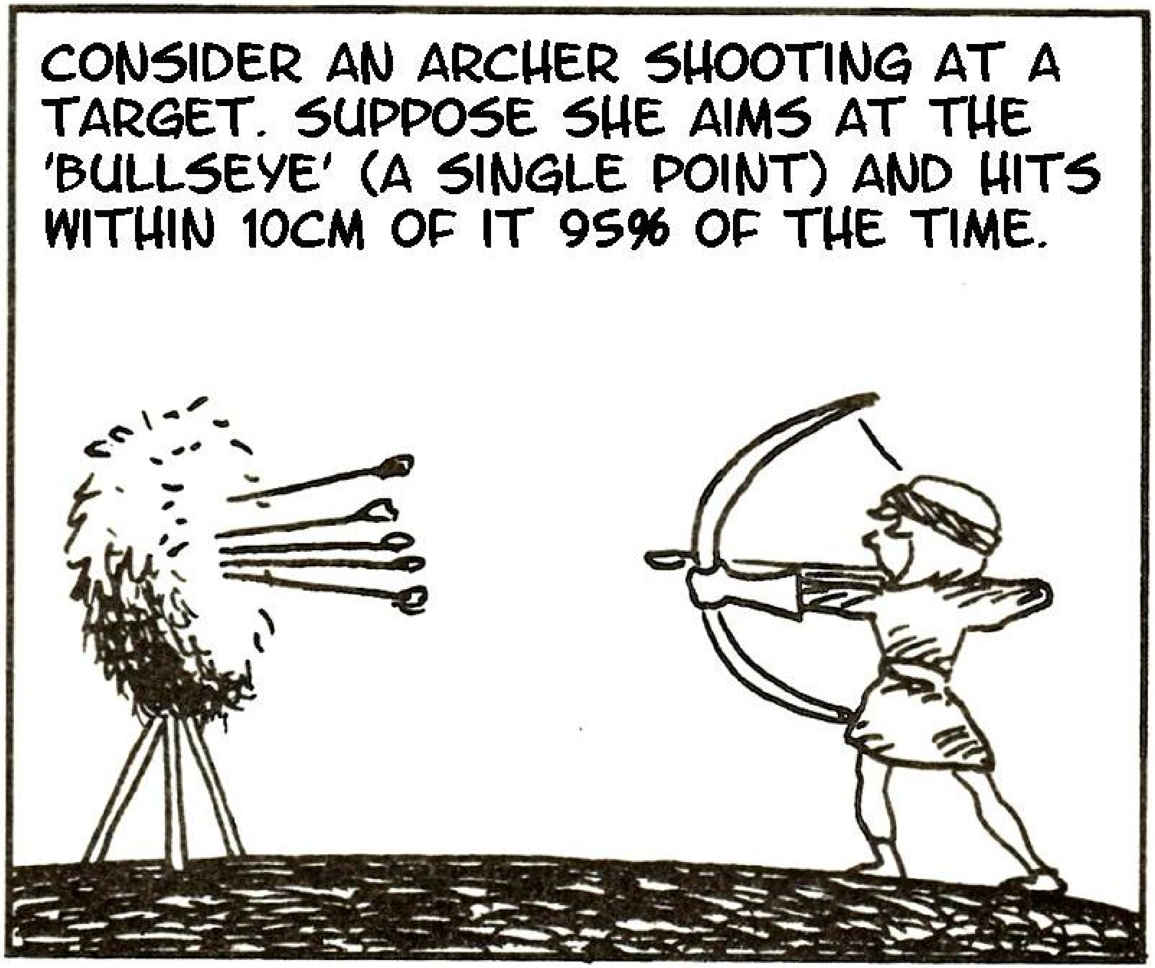
... and that's it! (essentially!)

- Given modeling assumptions & prior, process is automatic
- Keep adding data, and updating knowledge, as data becomes available... knowledge will concentrate around true $\boldsymbol{\theta}$

How does this differ from frequentist inference?

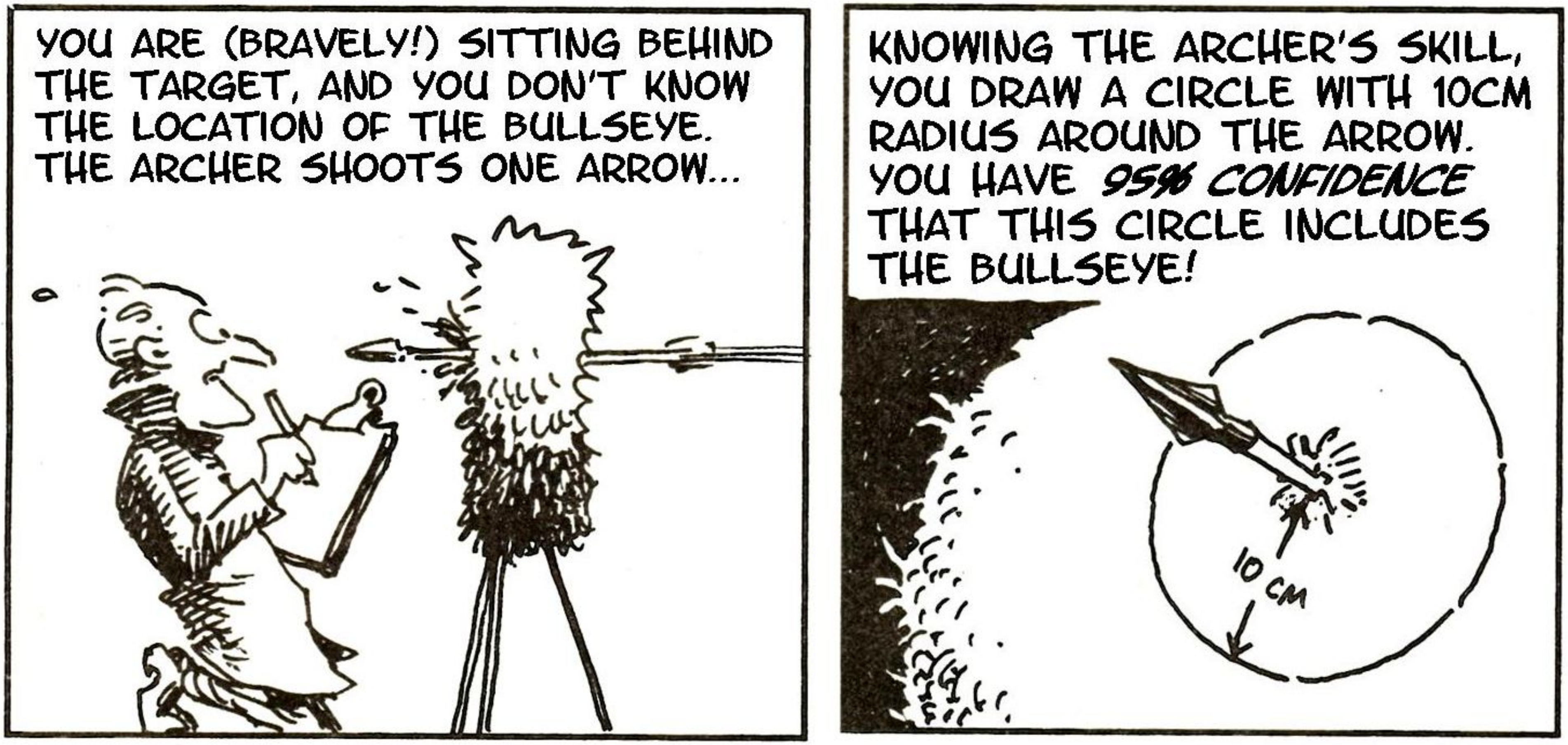
Freq'ist inference (I know, shoot me!)

Frequentist inference, set all a-quiver;



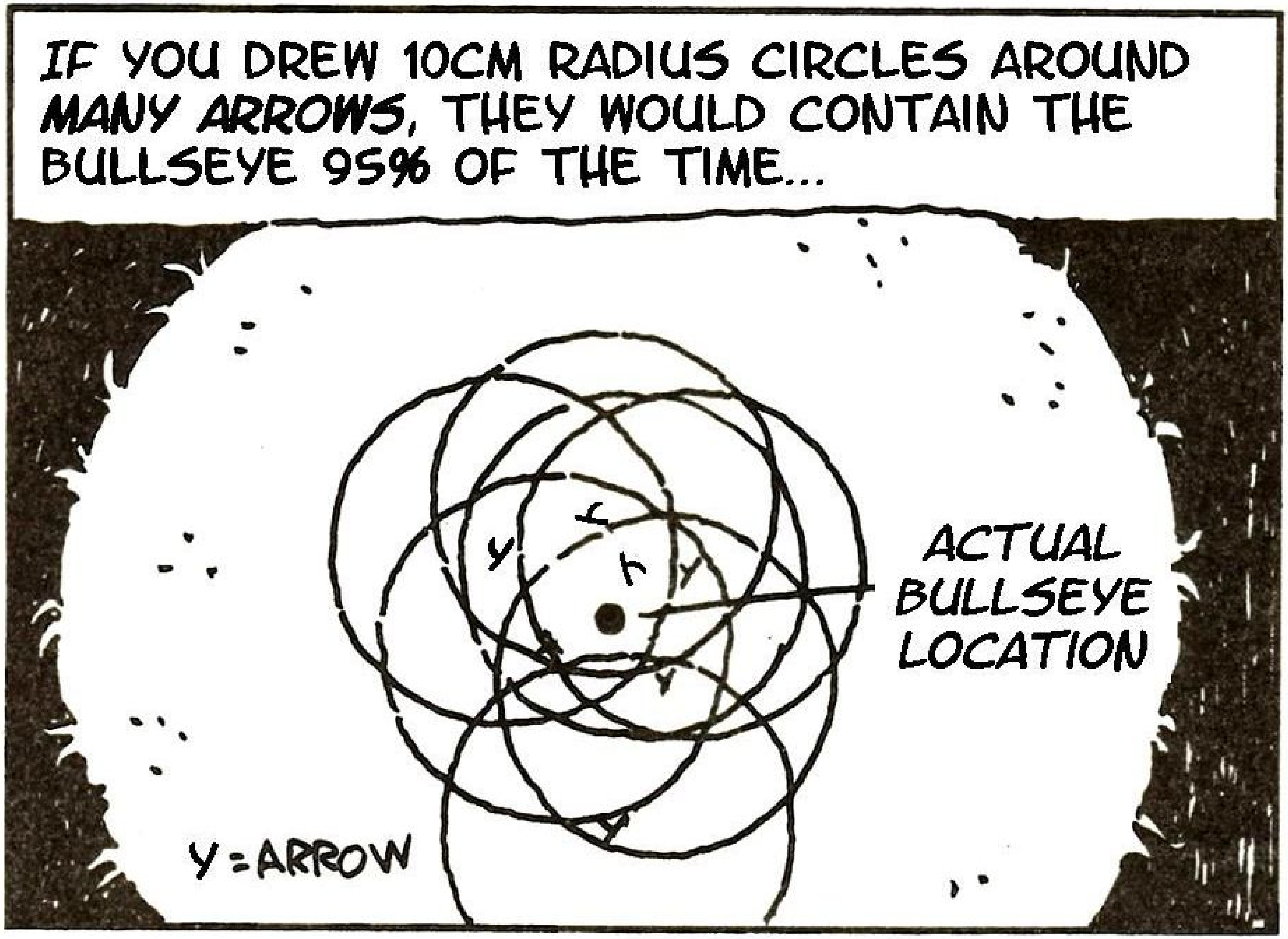
Freq'ist inference (I know, shoot me!)

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.

Parameters and likelihoods

The unknown 'parameter' in this example is the bullseye location. More generally, parameters quantify unknown population characteristics;

- Frequency of a particular SNP variant in that population
- Mean systolic BP in that population
- Mean systolic BP in that population, in those who have a particular SNP variant

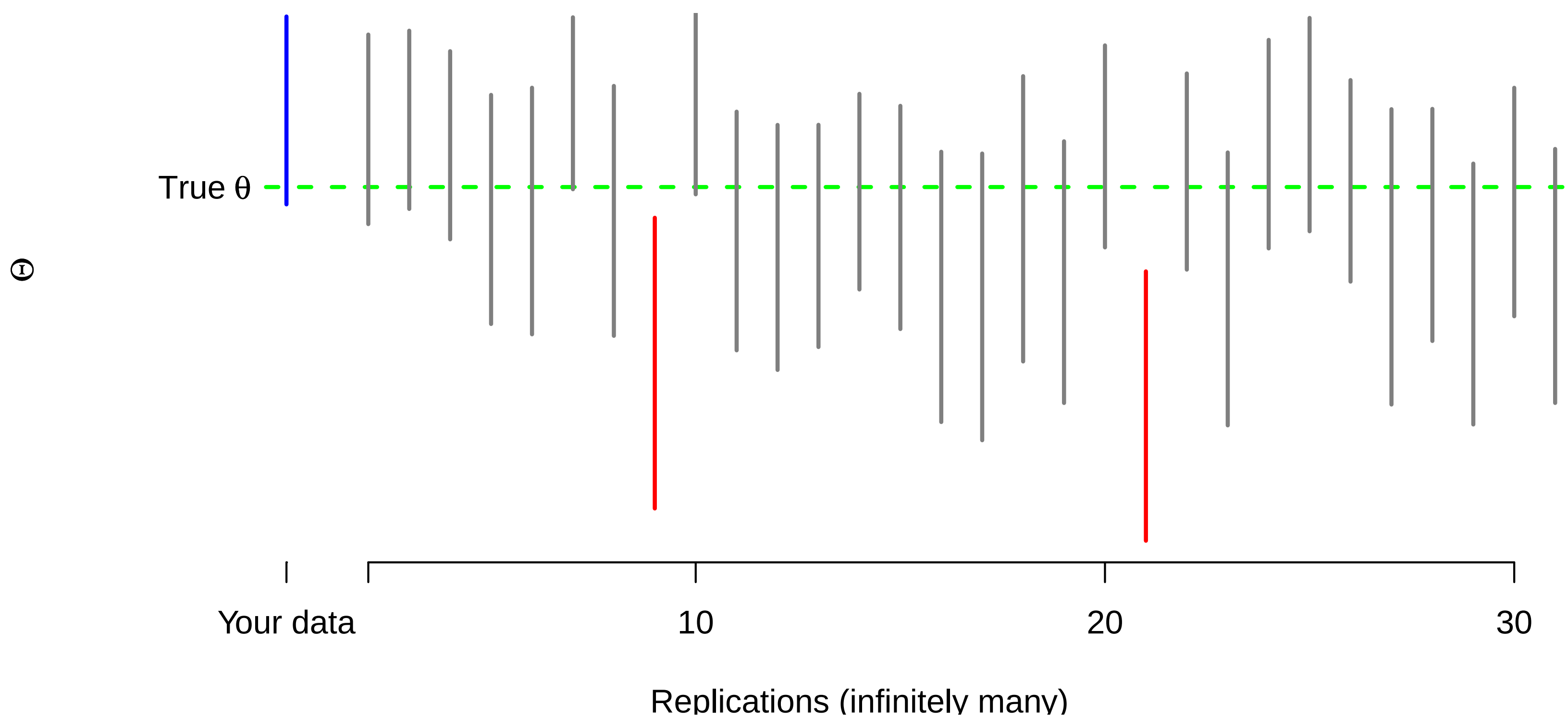
Parameters are traditionally denoted as Greek letters (θ,β ...ξ) and we write $p(y|\theta)$ to define the distribution of Y given a particular value of θ.

- Varying y, $p(y|\theta)$ tells how **relatively** likely different outcomes y are for fixed θ
- Varying θ, $p(y|\theta)$ (known as a *likelihood*) describes how relatively likely a given y is, at different θ

... more detailed examples follow in Session 2.

Frequentist inference: intervals

In almost all frequentist inference, confidence intervals take the form $\hat{\theta} \pm 1.96 \times \widehat{\text{stderr}}$ where the *standard error* quantifies the 'noise' in some estimate $\hat{\theta}$ of parameter θ.



(The 1.96 comes from $\hat{\theta}$ following a Normal distribution, approximately — more later)

Frequentist inference: intervals

Usually, we imagine running the '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

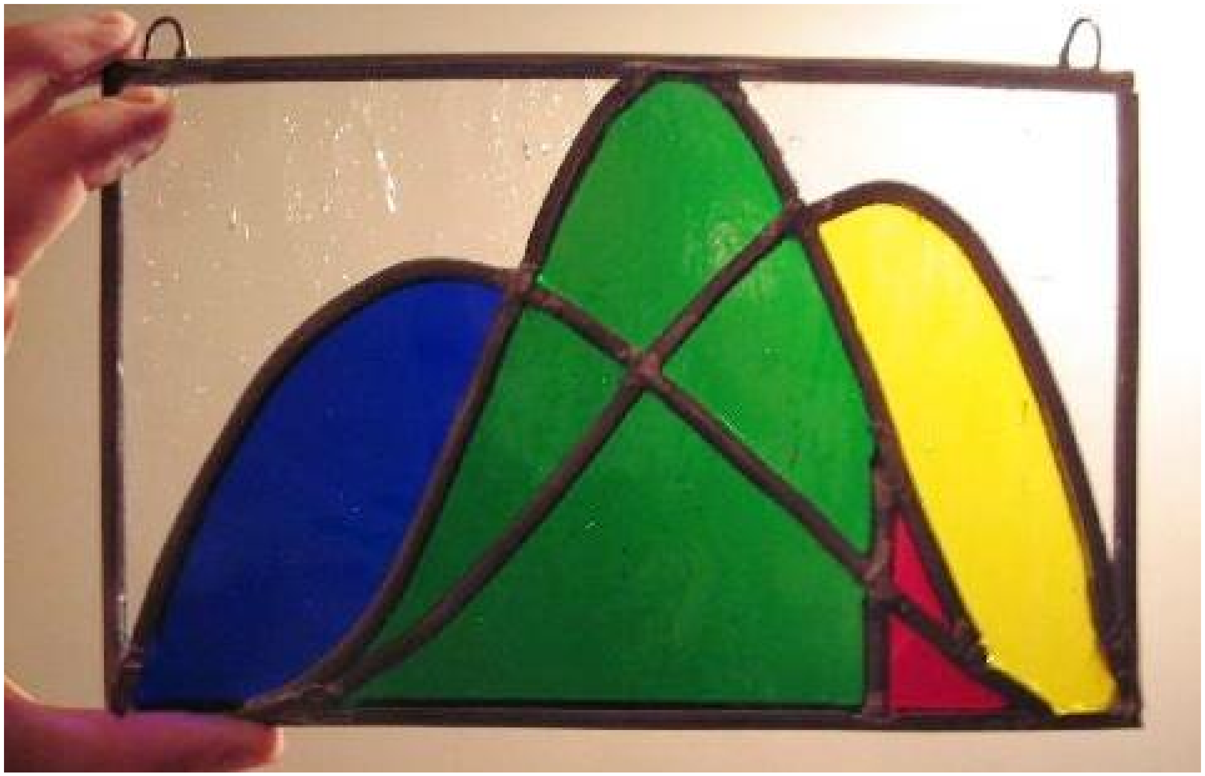
Larry Wasserman, All of Statistics

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. Bayes does not! ...and this is how scientists *tend* to think about data

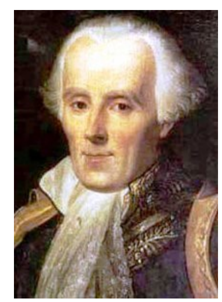
Back to Bayesian simplicity

Bayesian inference can be made, er, transparent;



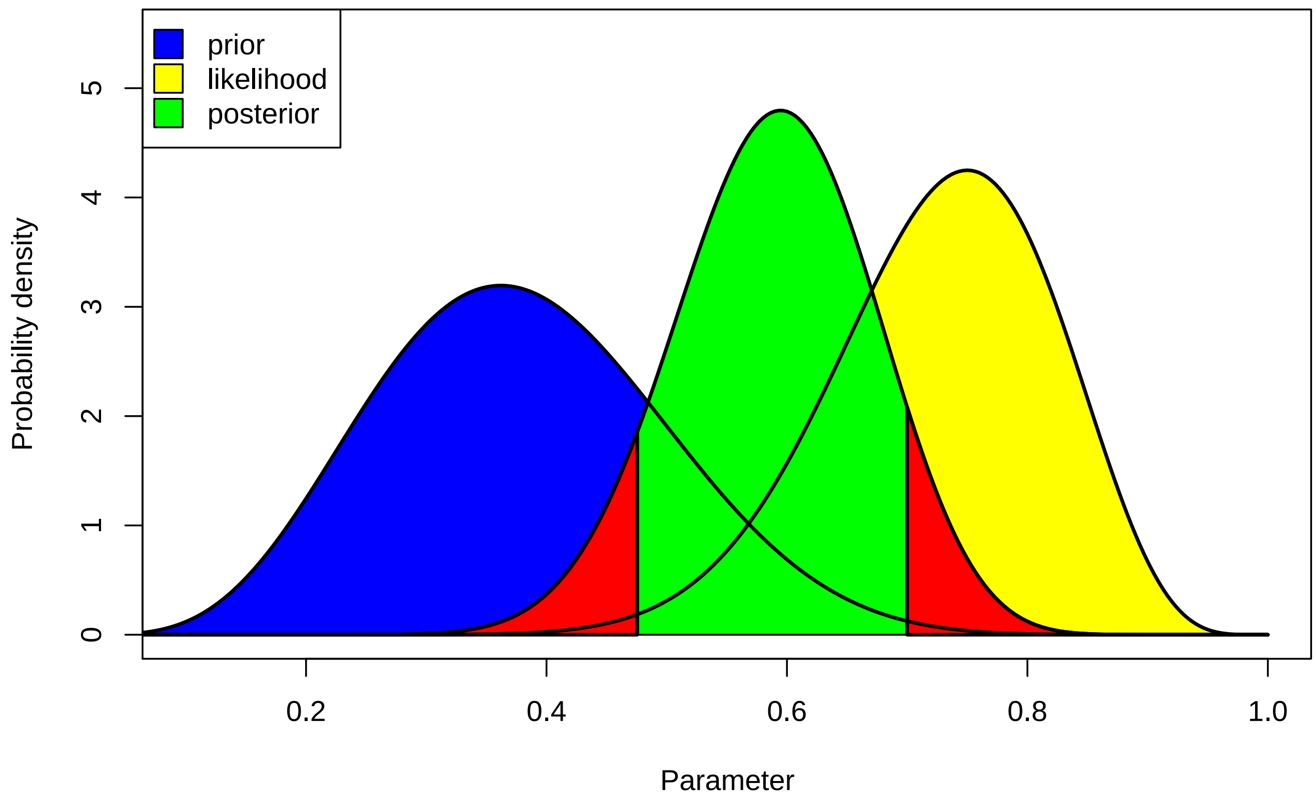
Common sense reduced to computation

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



Back to Bayesian simplicity

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

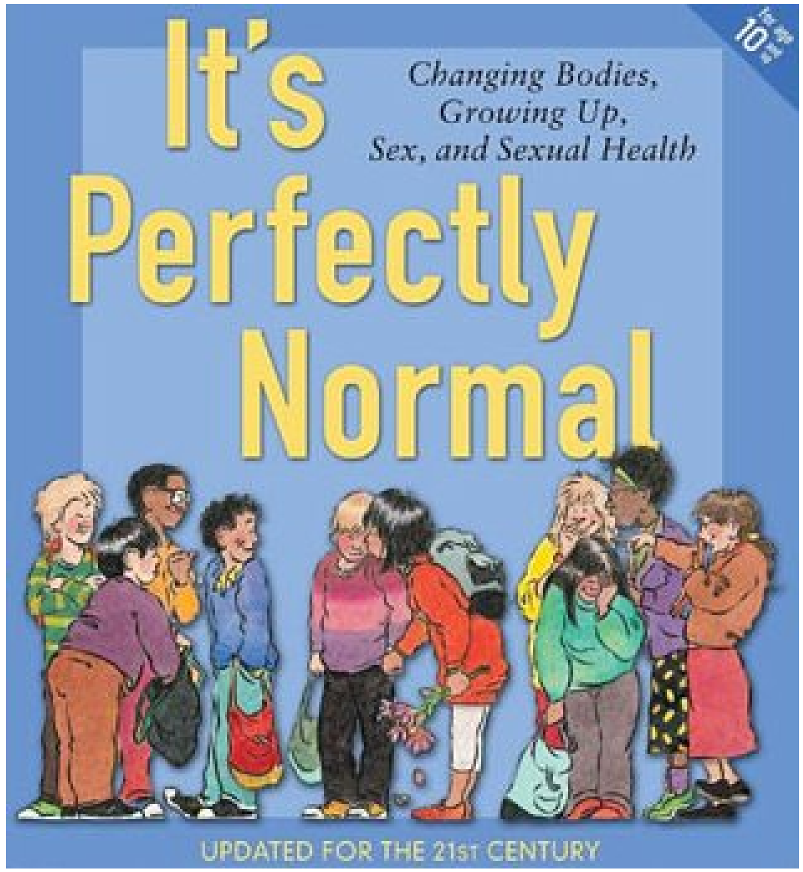


A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule

Stephen Senn, Statistician & Bayesian Skeptic (mostly)

But where do priors come from?

An important day at statistician-school?



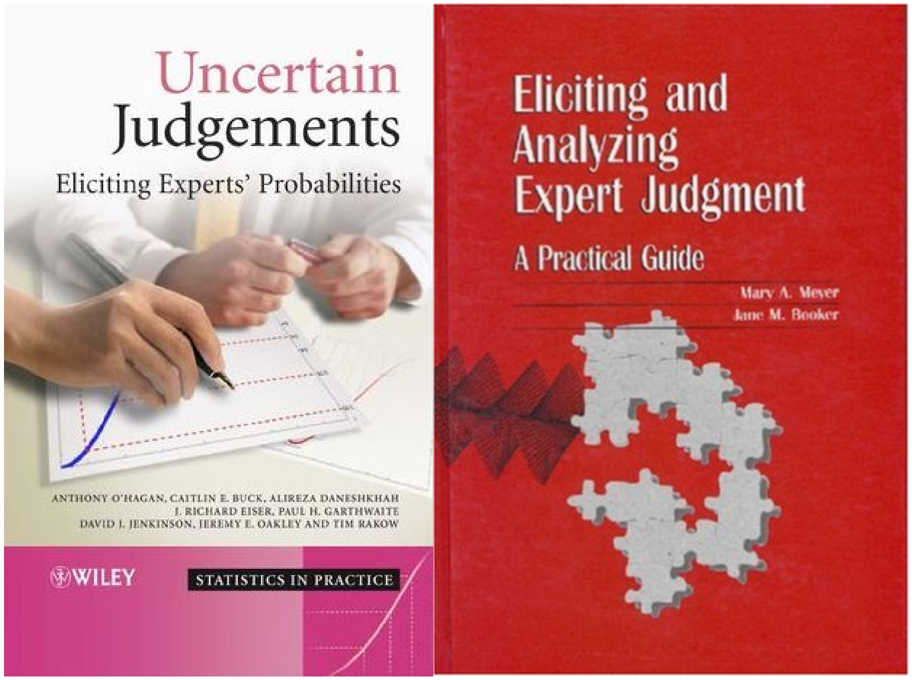
There's nothing wrong, dirty, unnatural or even *unusual* about making assumptions – carefully. Scientists & statisticians all make assumptions... even if they don't like to talk about them.

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;

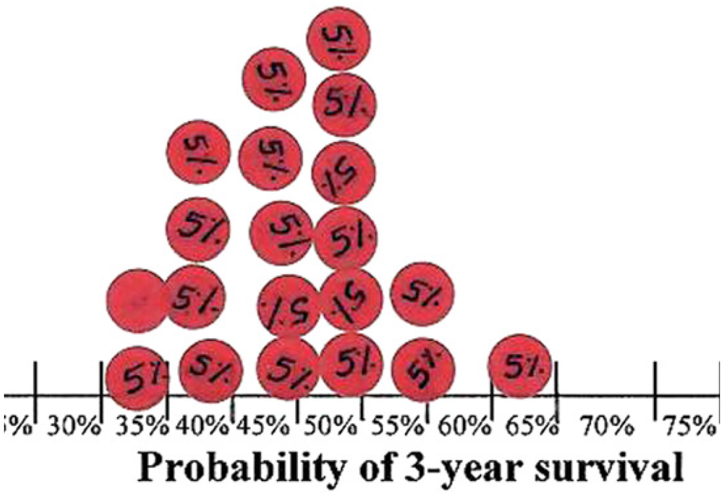


- Discuss parameters experts understand – e.g. code variables in familiar units, make comparisons relative to an easily-understood reference, *not* with age=height=IQ=0
- Avoid leading questions (just as in survey design)
- The 'language' of probability is unfamiliar; help users express their uncertainty

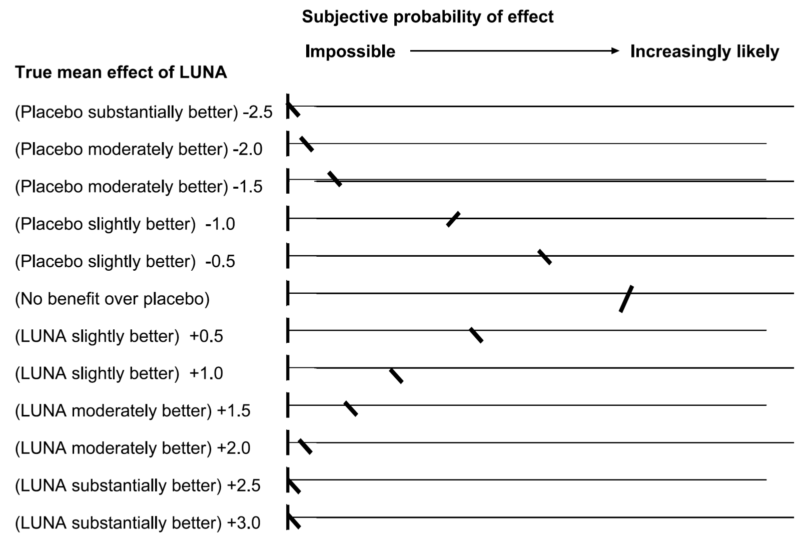
Kynn (2008, JRSSA) is a good review, describing many pitfalls.

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

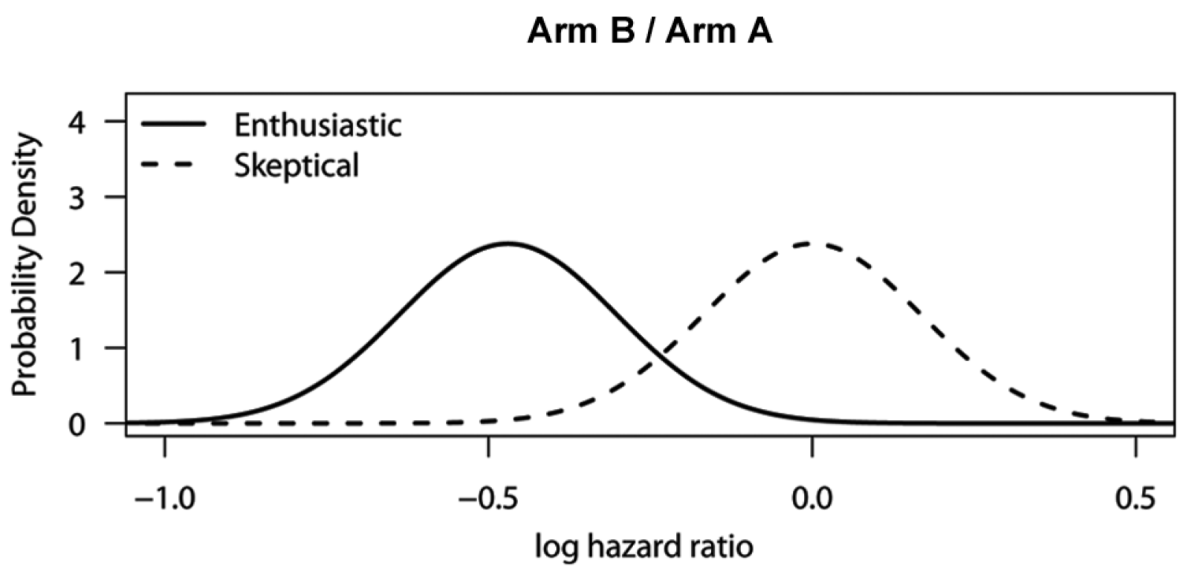


Normalize marks (Latthe *et al* 2005, J Obs Gync) for prior on pain effect of LUNA vs placebo

- 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)



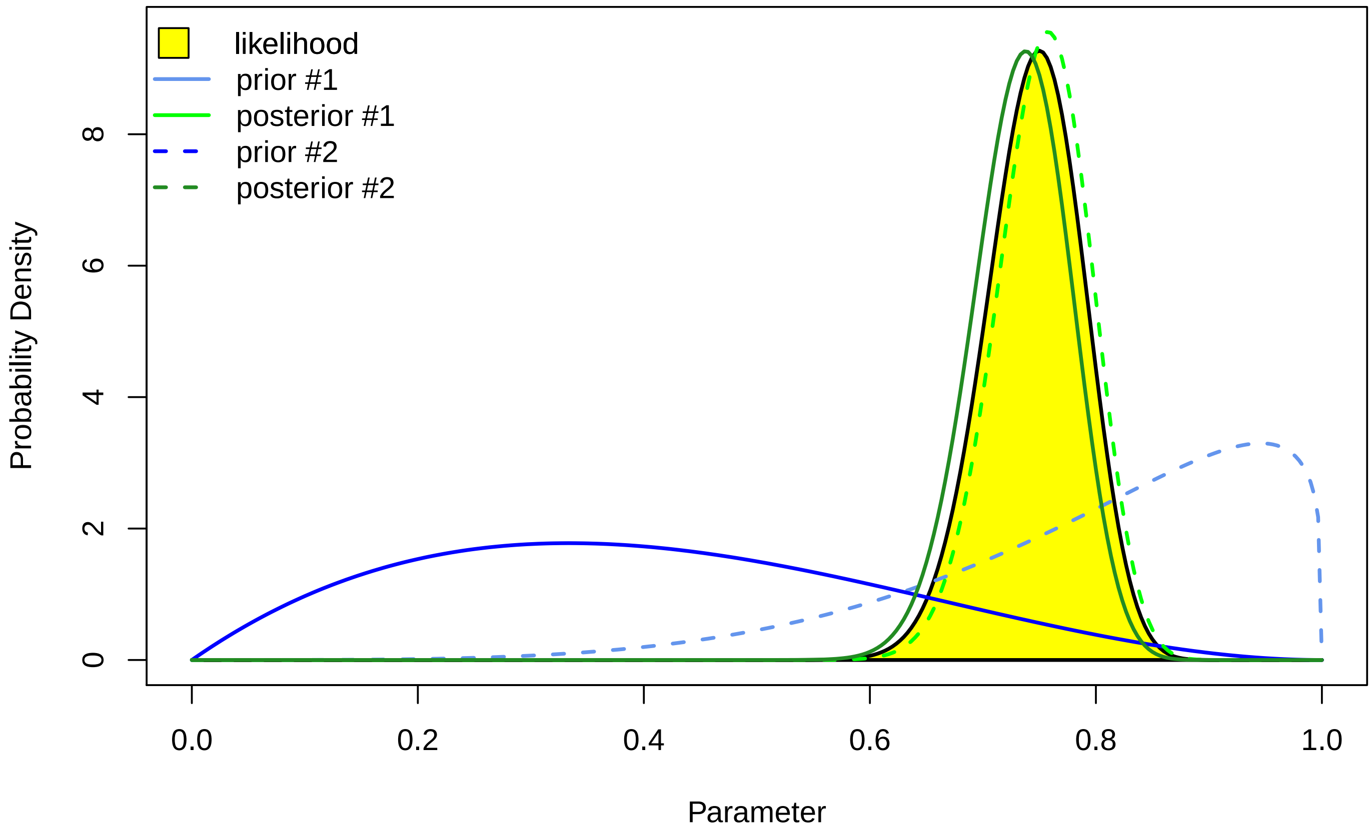
Parmer *et al* (1996, JNCI) popularized the definitions, they are now common in trials work

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)?

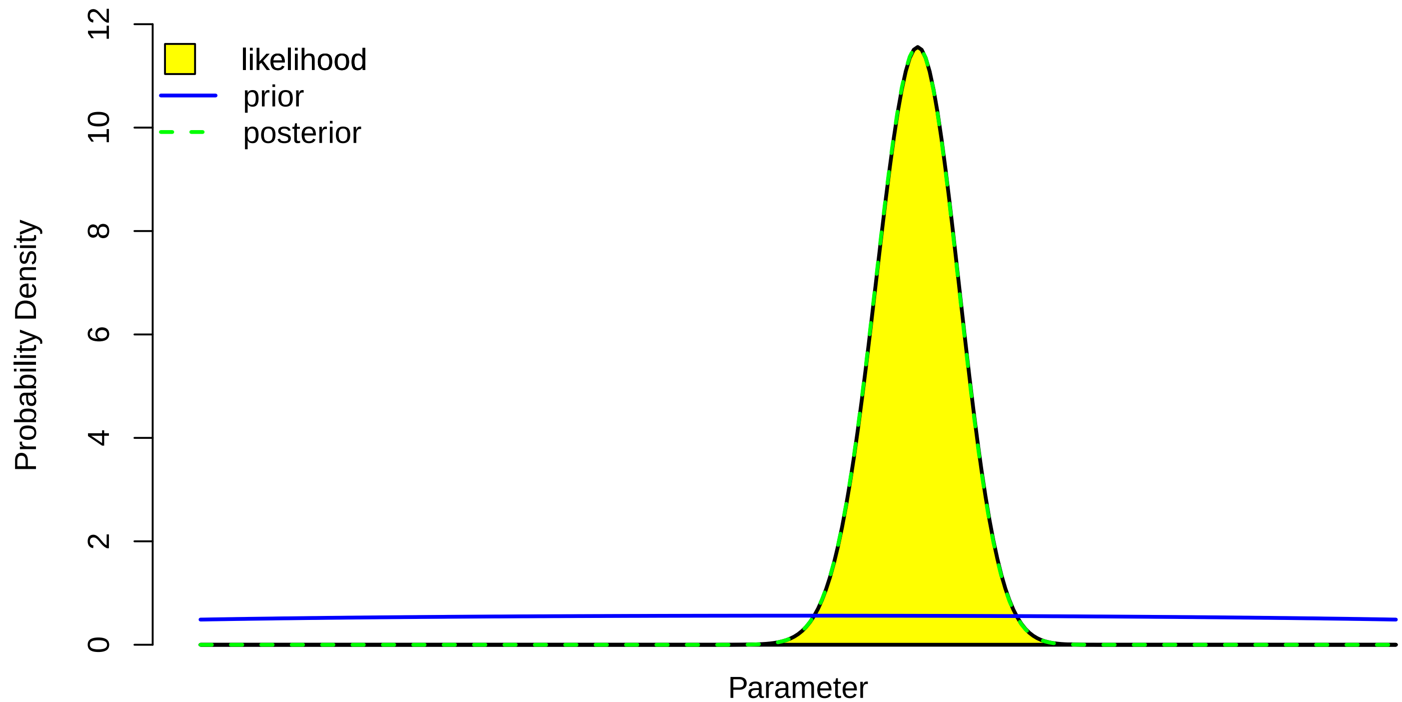
When the data provide a lot more information than the prior, this happens; (recall the stained glass color-scheme)



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)?

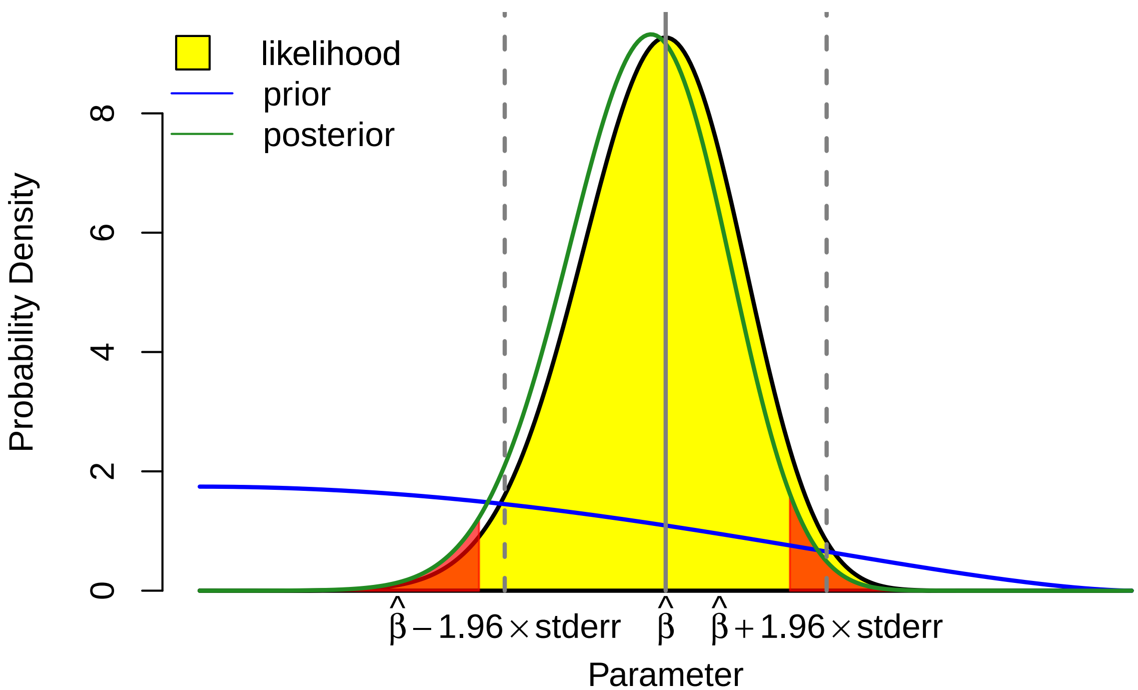
A related idea; try using very flat priors to represent ignorance;



- Flat priors do NOT actually represent ignorance! Most of their support is for *very* extreme parameter values
- For parameters in 'famous' regression models, this idea works okay – it's more generally known as 'Objective Bayes'
- For many other situations, it doesn't, so be careful! (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.

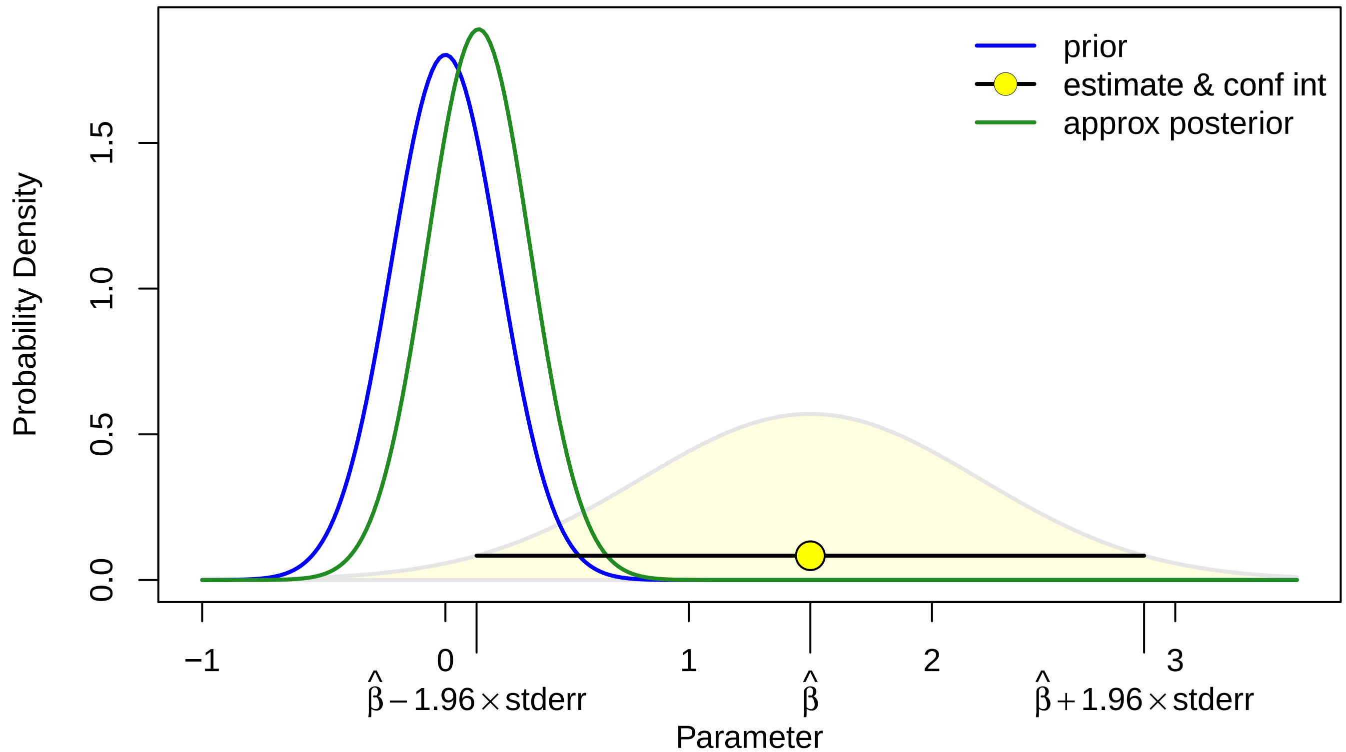
One way to do this;

- Take point-estimate $\hat{\beta}$ and corresponding standard error $stderr$, calculate precision $1/stderr^2$
- Elicit prior mean β_0 and prior standard deviation σ; calculate prior precision $1/\sigma^2$
- 'Posterior' precision $= 1/stderr^2 + 1/\sigma^2$ (which gives overall uncertainty
- 'Posterior' mean $=$ *precision-weighted mean* of $\hat{\beta}$ and β_0

Note: This is a (very) quick-and-dirty approach; we'll see much more precise approaches in later sessions.

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

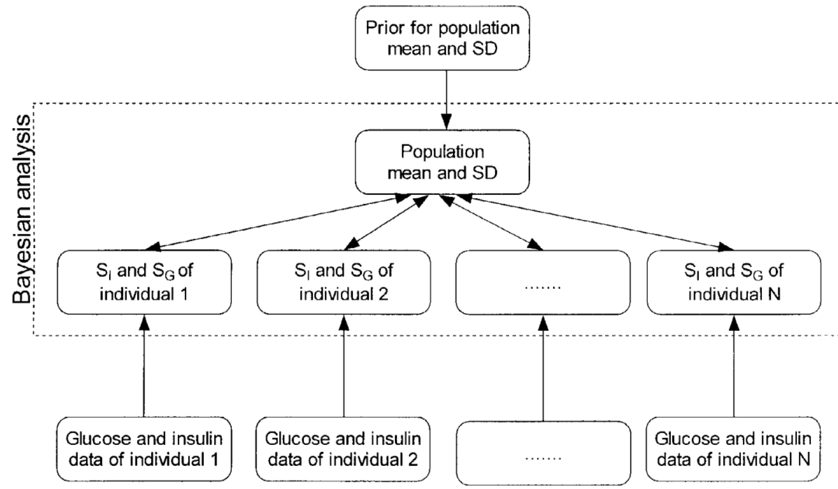
Many results that are rigorously proved and accepted start shrinking in later studies.

- 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

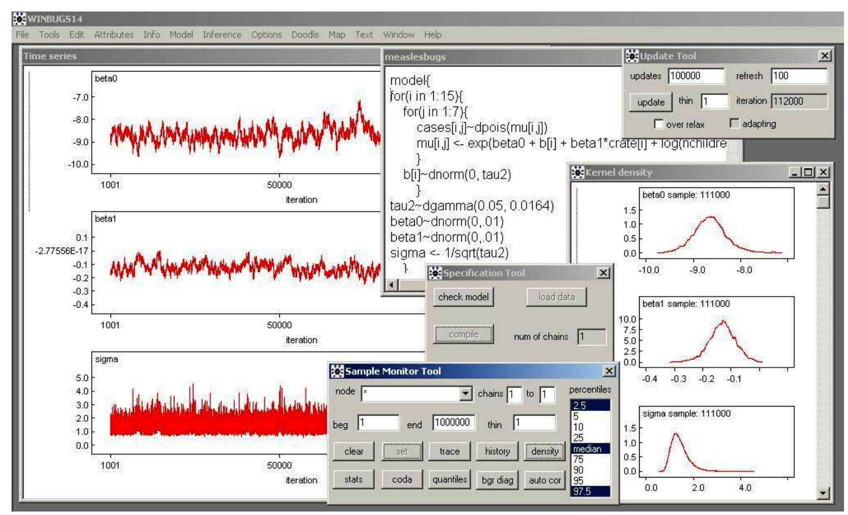
Where is Bayes commonly used?

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 popular regression methods, with large n, Bayesian and frequentist ideas often don't disagree much. This is (provably!) 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)

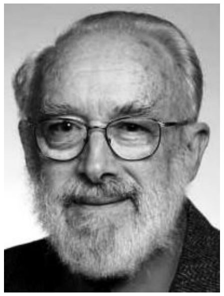
And while e.g. hierarchical modeling & multiple imputation are easier to justify in Bayesian terms, they aren't *un*frequentist.

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

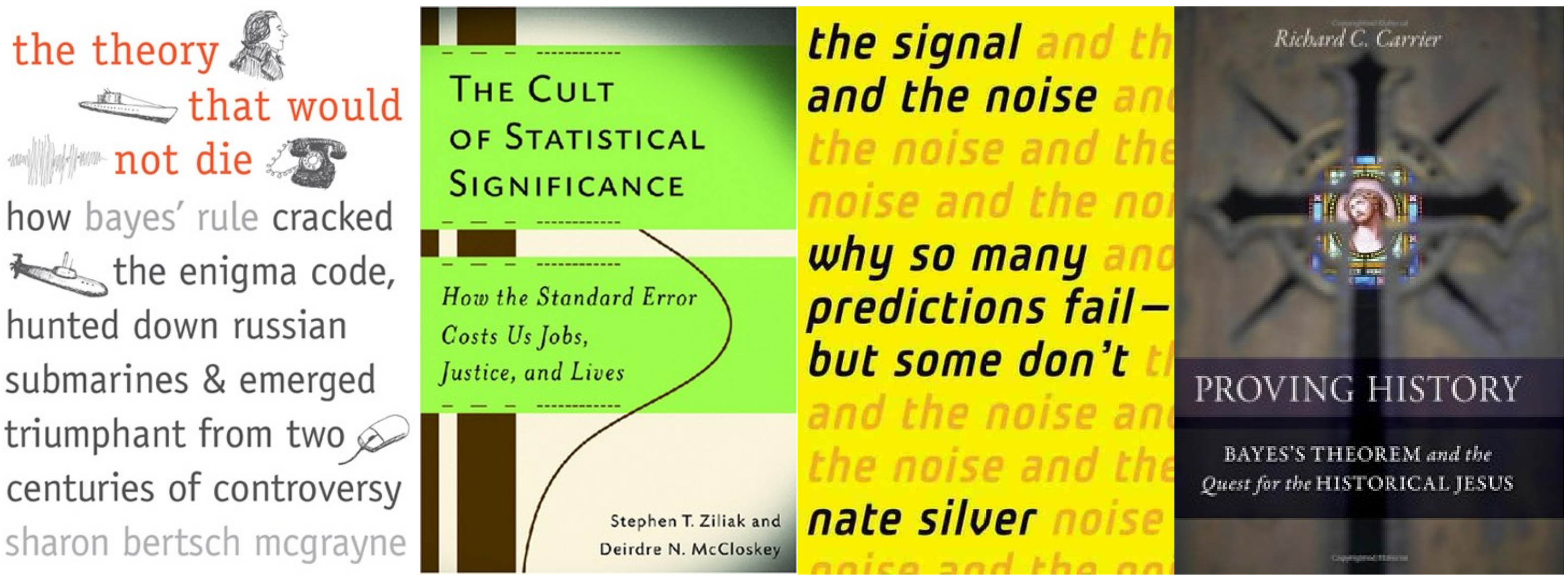
Dennis Lindley (1923–2013)
in The Future of Statistics: A Bayesian 21st Century (1975)

- 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!

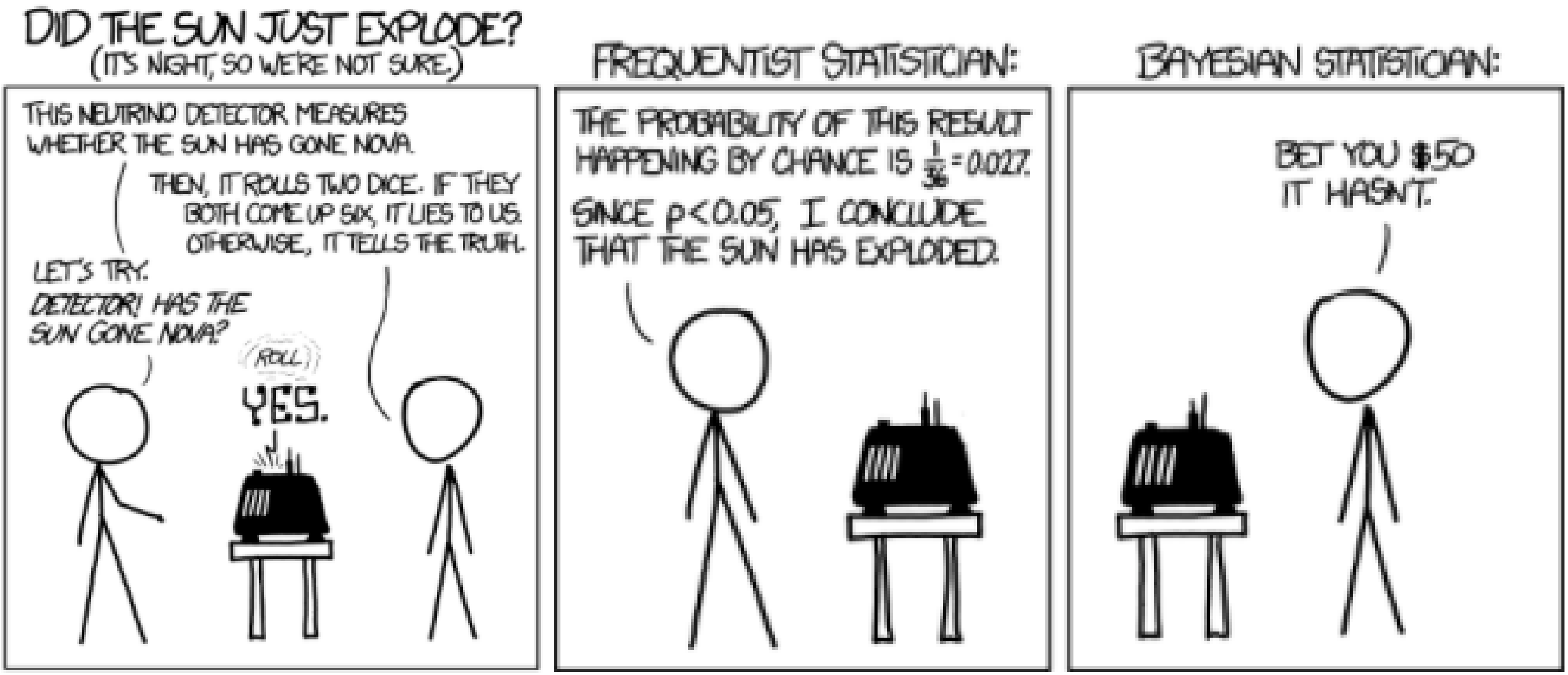
But writers of dramatic/romantic stories about Bayesian "heresy" [NYT] tend (I think) to over-egg the actual differences;



- 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 yet again, on Frequentists vs Bayesians;

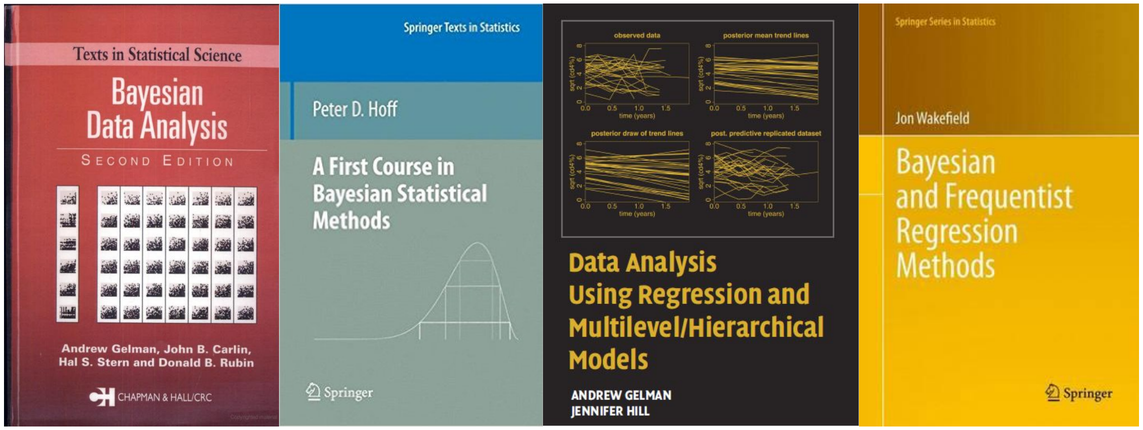


Here, the fun relies on setting up a straw-man; p-values are not the only tools used in a *skillful* frequentist analysis.

Note: 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 – the course site will feature more resources. We will focus on Bayesian approaches to ;

- Regression-based modeling
- Testing
- Learning about multiple parameters (testing)
- Combining data sources (imputation, meta-analysis)

– but the general principles apply very broadly.

Summary

Bayesian statistics:

- Is useful in many settings, and intuitive
- 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. Practical uses abound.

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
- Ken's website, or Jon's website