Bayesian Statistics for Genetics

Lecture 1: Introduction

July, 2020

Overview

We'll cover only the key points from a very 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.

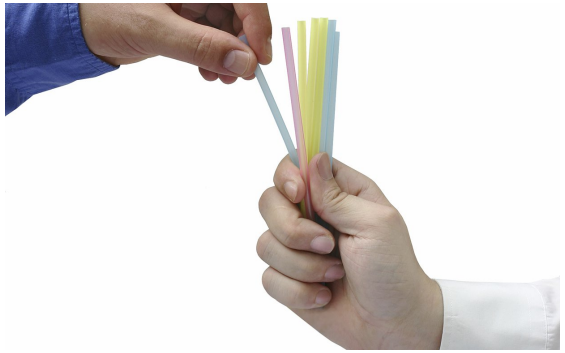
There are different ways of tackling statistical problems, too.

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;







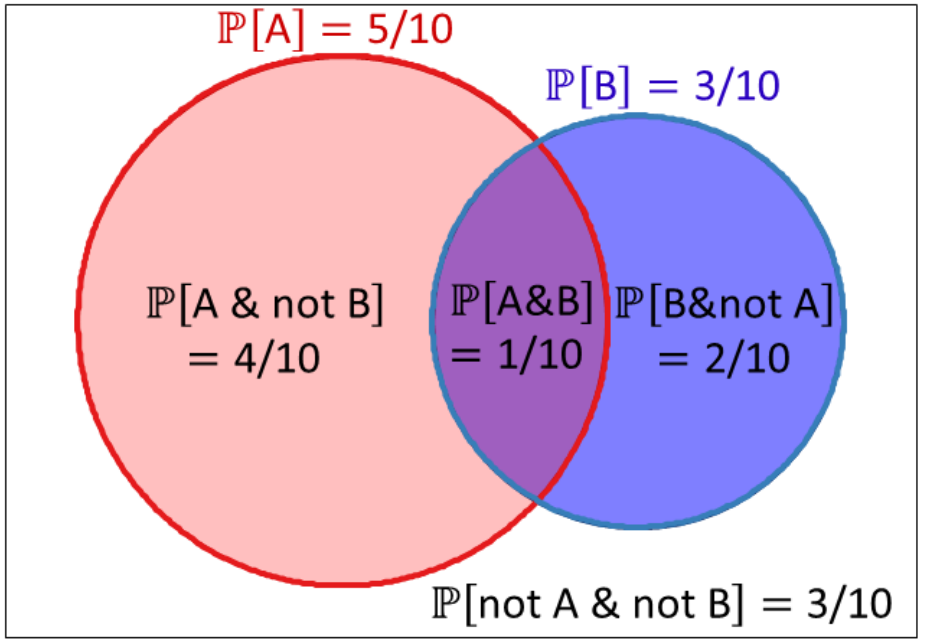

* Sorry! Necessary math ahead!

Bayes' Theorem

Bayes' *Theorem* describes *conditional* probabilities: for events A and B, $\mathbb{P}[\,A|B\,]$ denotes the probability that A happens **given** that B happens. In this example;

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

Bayes' Theorem states how $\mathbb{P}[\,A|B\,]$ and $\mathbb{P}[\,B|A\,]$ are related:



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

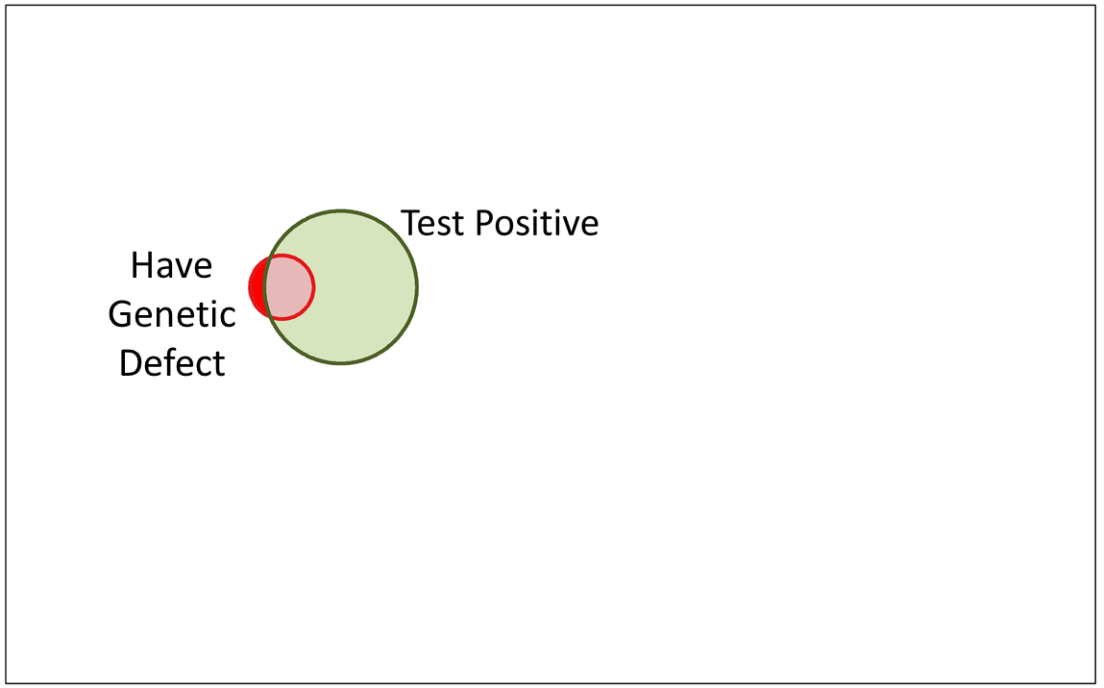
...so here, $1/3 = 1/5 \times \frac{5/10}{3/10}$ (✓)

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 a genetic defect, for a screening test with 80% sensitivity and 95% specificity;

$$\begin{aligned}
\mathbb{P}[\,\text{Test -ve}\,|\,\text{no defect}\,] &= 95\% \\
\mathbb{P}[\,\text{Test +ve}\,|\,\text{defect}\,] &= 80\% \\
\frac{\mathbb{P}[\,\text{Test +ve}]}{\mathbb{P}[\,\text{defect}\,]} &= 5.75 \\
\mathbb{P}[\,\text{defect}\,|\,\text{Test +ve}\,] &\approx 14\%
\end{aligned}$$



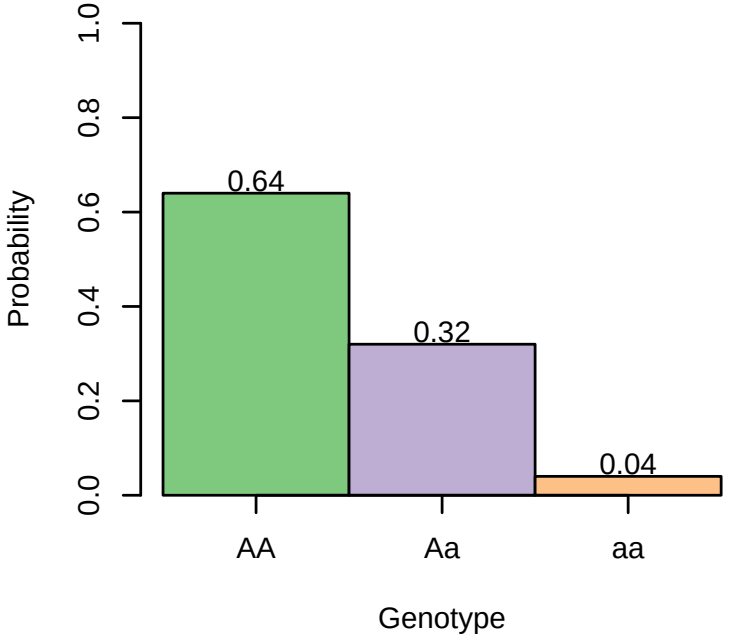
... i.e. most positive results are actually false alarms.

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

Bayes' Theorem

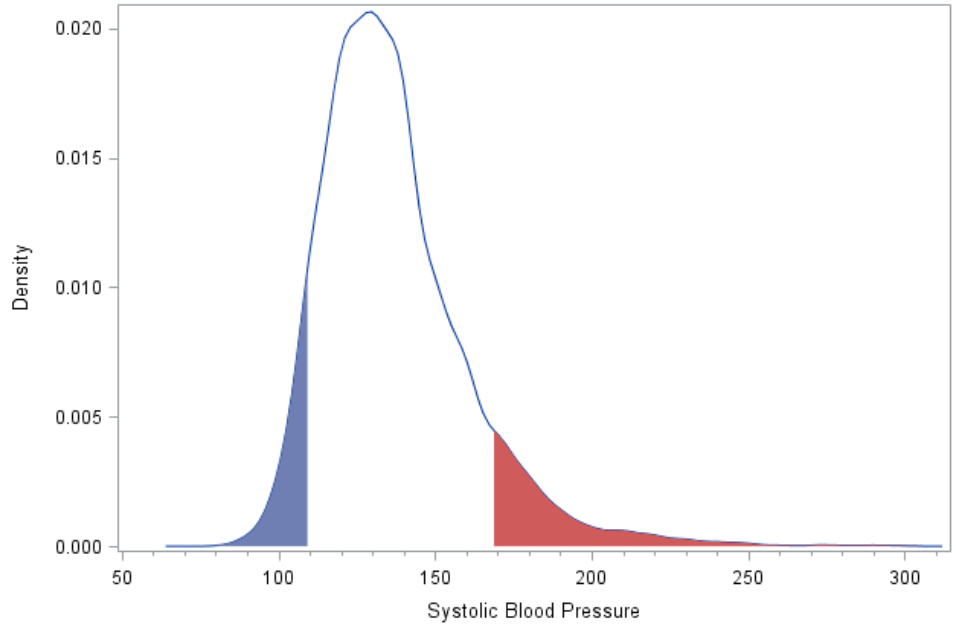
The 'language' of probability is much richer than just Yes/No events;

Categorical (probabilities)



Probability of having at least one copy of the 'a' allele is 0.32+0.04=0.36, i.e. 36%.

Continuous (density function)



Probability of sets (e.g. a randomly-selected adult SBP>170 or <110mmHg) is given by 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;

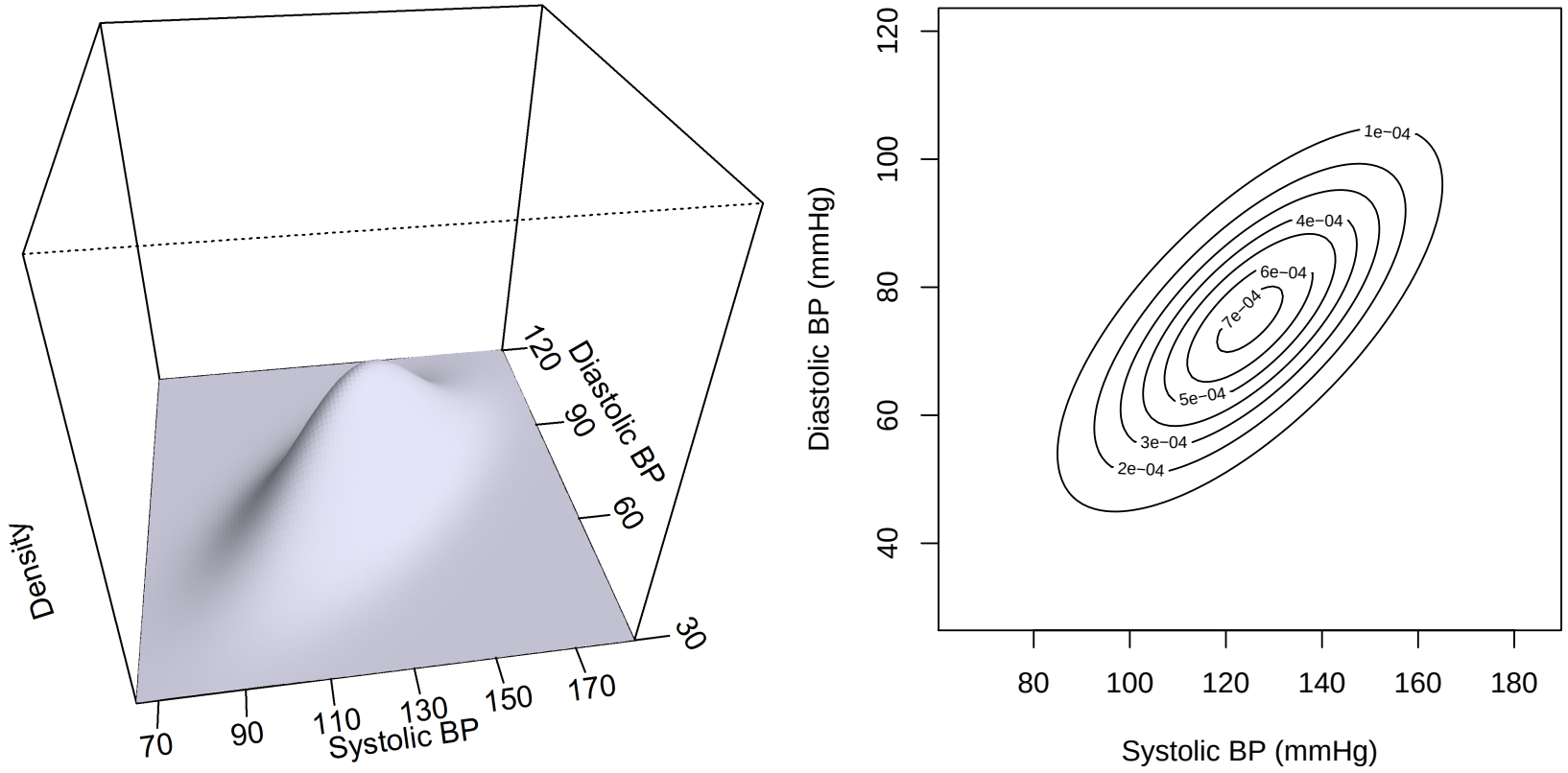
$$\mathbb{P}[\, Y \in \mathcal{Y}_0\,] = \int_{\mathcal{Y}_0} p(y)dy,$$

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

For discrete events, replace integration by addition over possible outcomes.

Bayes' Theorem

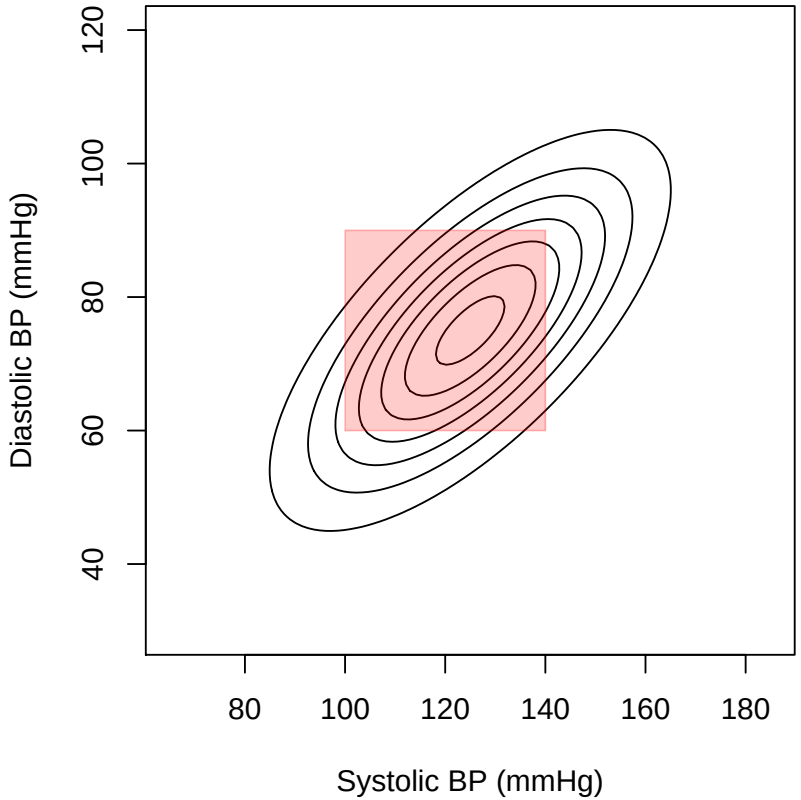
The same ideas for two random variables, where the density is a surface;



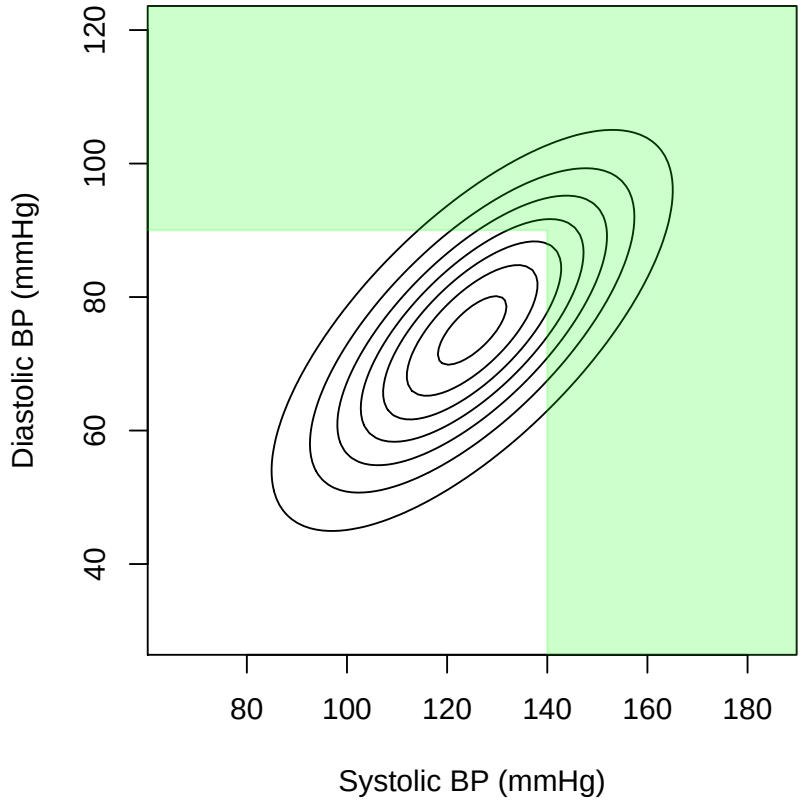
... where the total 'volume' is 1, i.e. $\int_{\mathcal{X},\mathcal{Y}} p(x,y)dxdy = 1$.

Bayes' Theorem

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



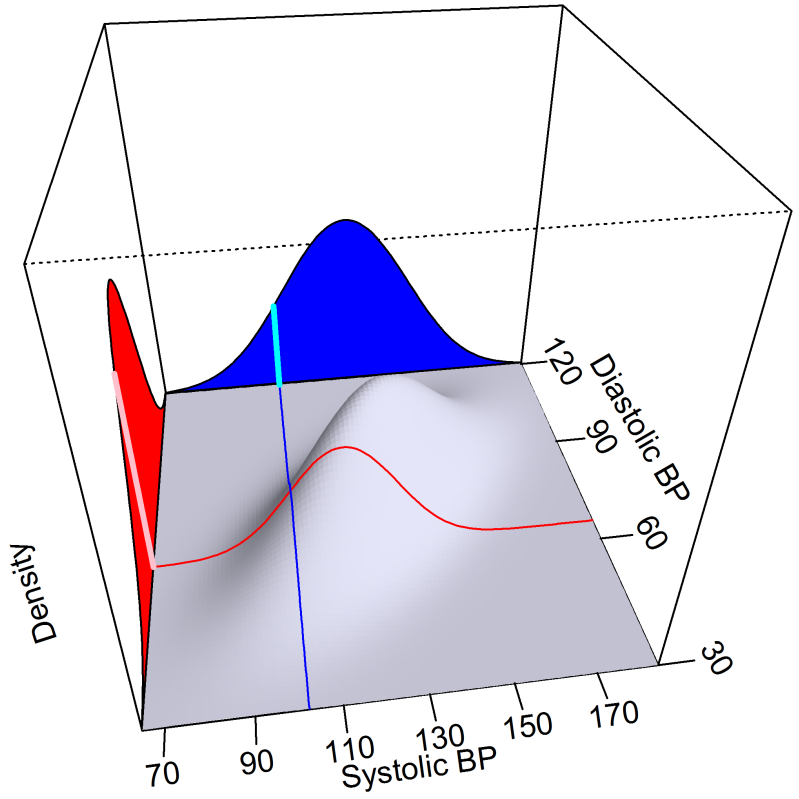
$$\mathbb{P}\left[\begin{array}{c} 100 < \text{SBP} < 140 \\ \& \\ 60 < \text{DBP} < 90 \end{array}\right] \approx 0.52$$



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

Bayes' Theorem

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



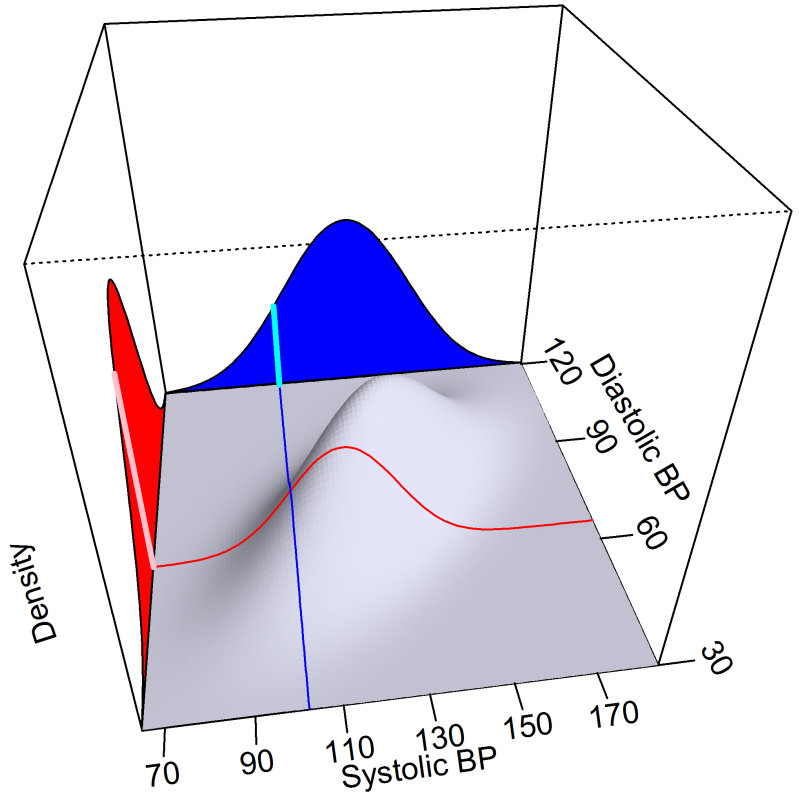
$$p(x|y = y_0) = p(x, y_0) / \int_{\mathcal{X}} p(x, y_0) dx$$
$$p(y|x = x_0) = p(x_0, y) / \int_{\mathcal{Y}} p(x_0, y) dy,$$

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

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

Bayes' Theorem

Bayes' theorem connects different conditional distributions –



Bayes' Theorem says the relationship between conditional densities is;

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

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 math result 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

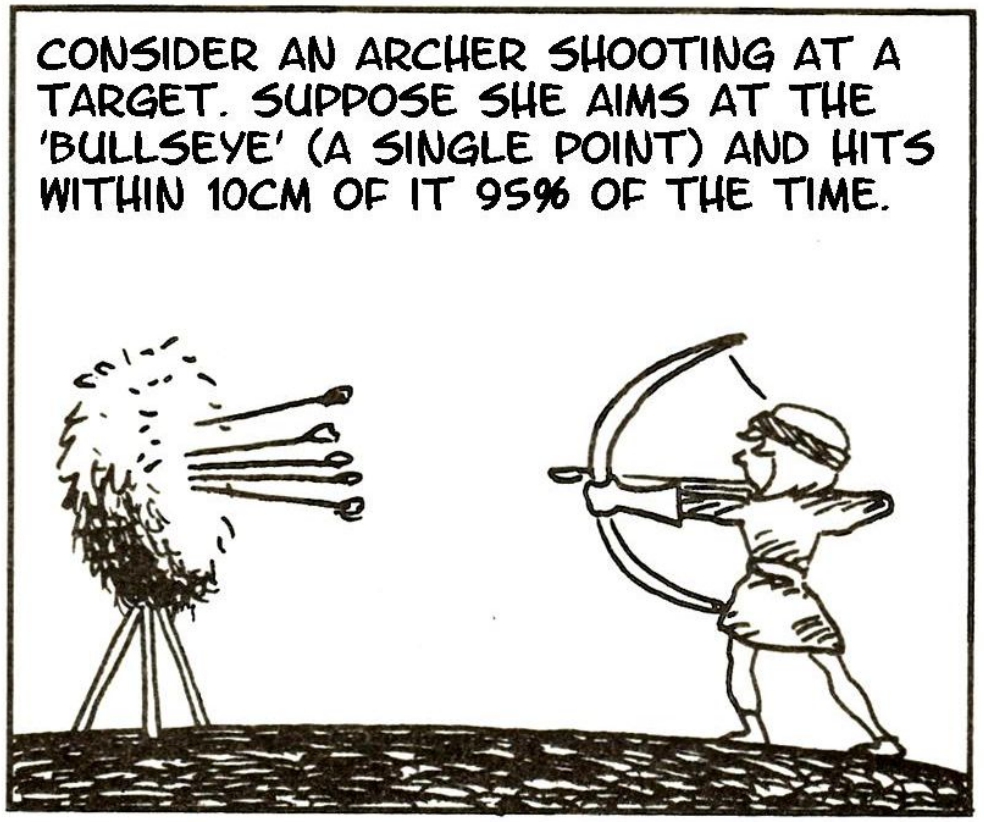
> In *addition* to describing random variables, Bayesian statistics uses the 'language' of probability to describe what is known about unknown parameters.

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

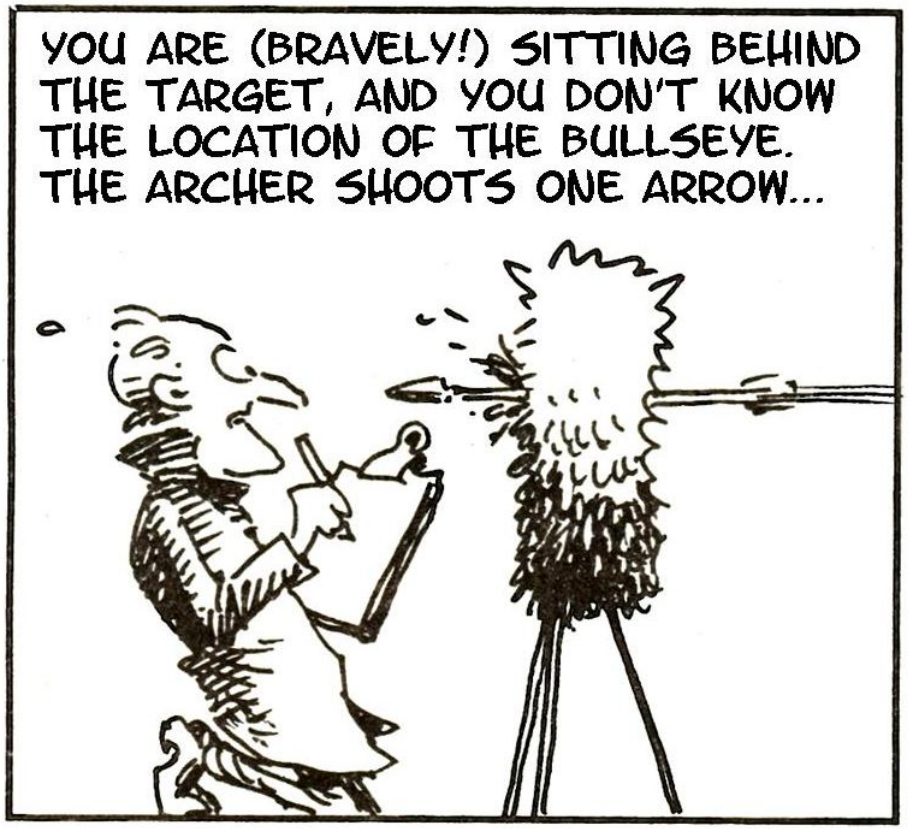
How does it work? Let's take aim...



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

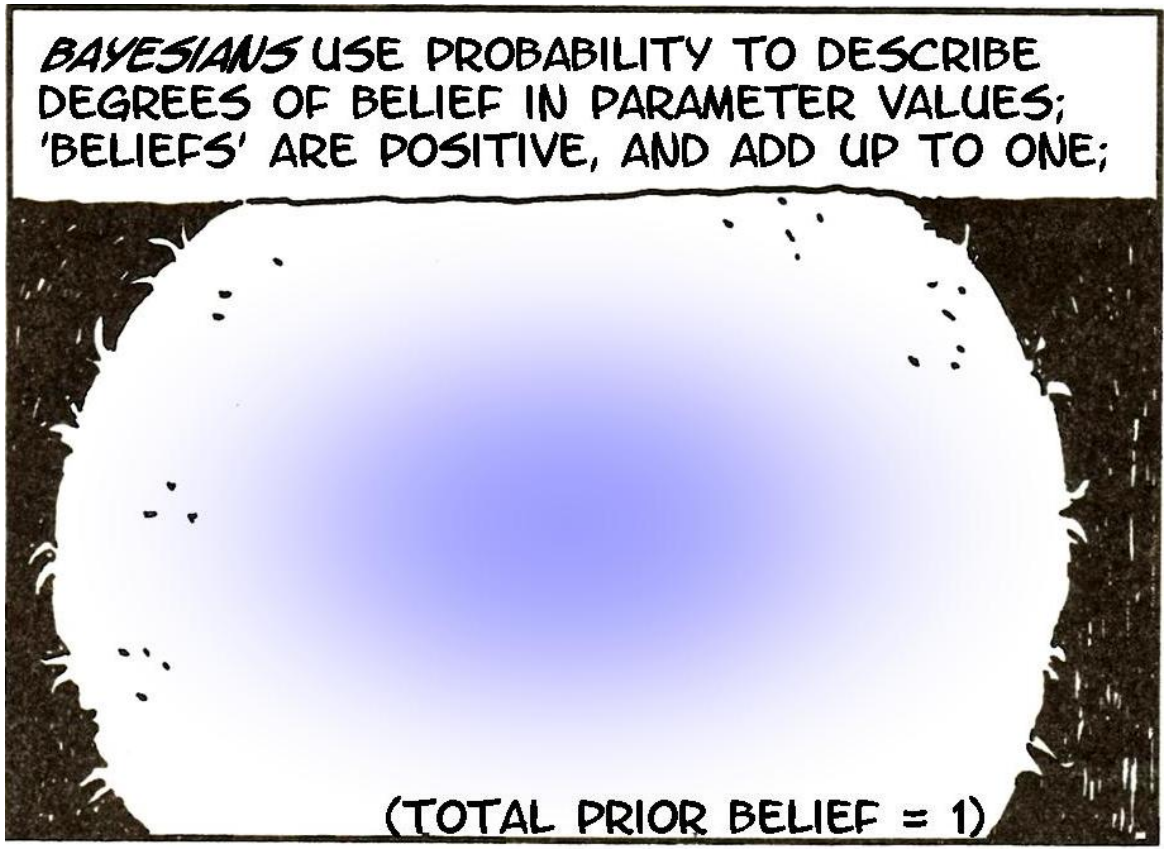
Bayesian inference

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



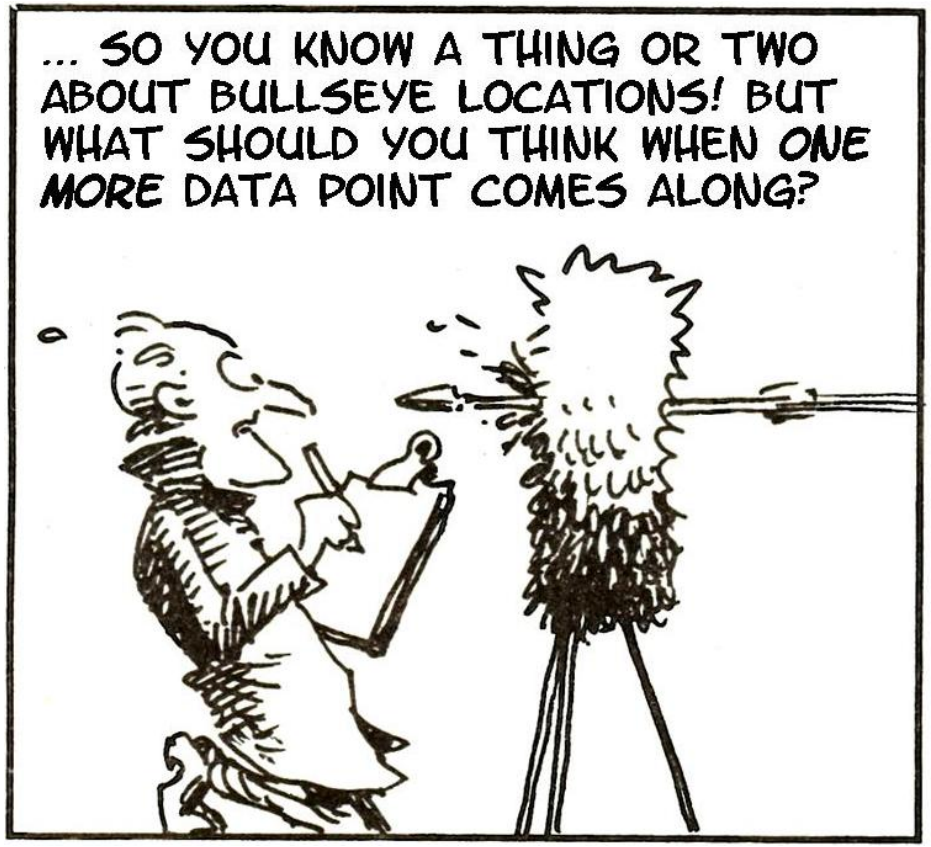
Bayesian inference

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



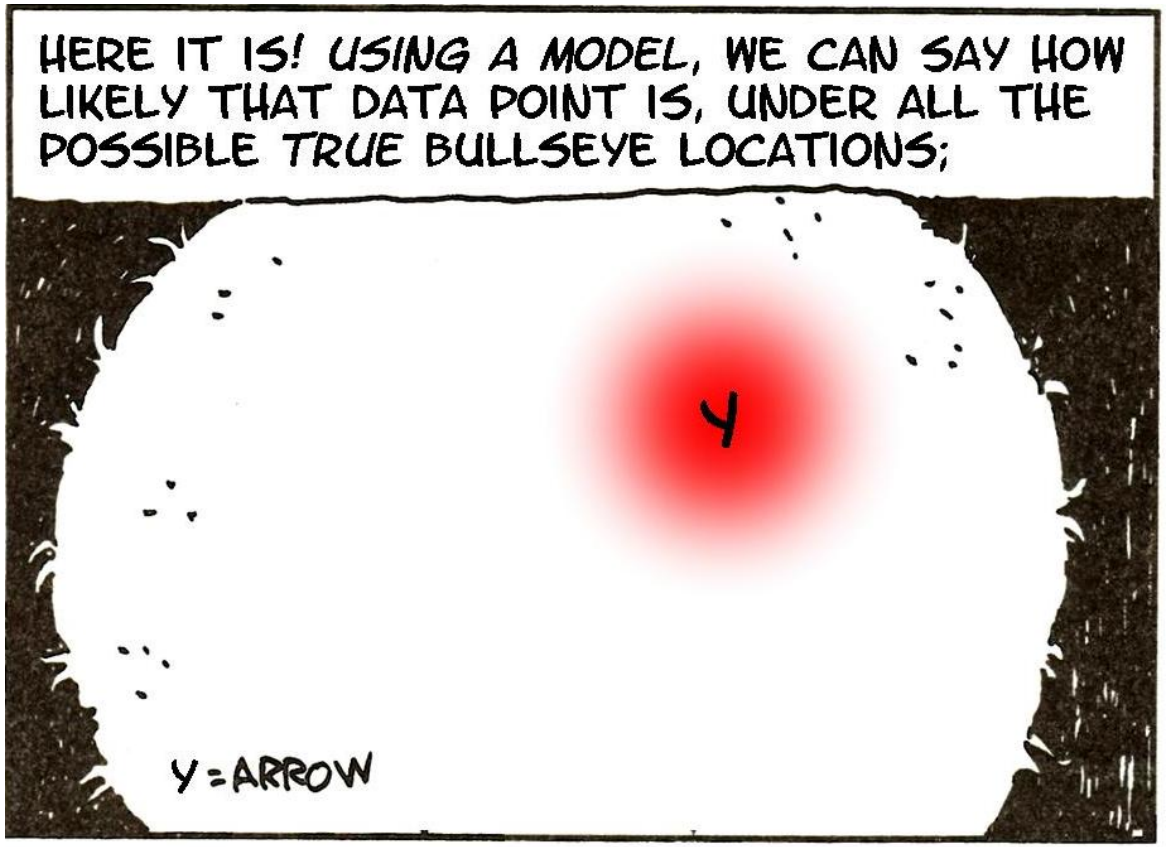
Bayesian inference

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



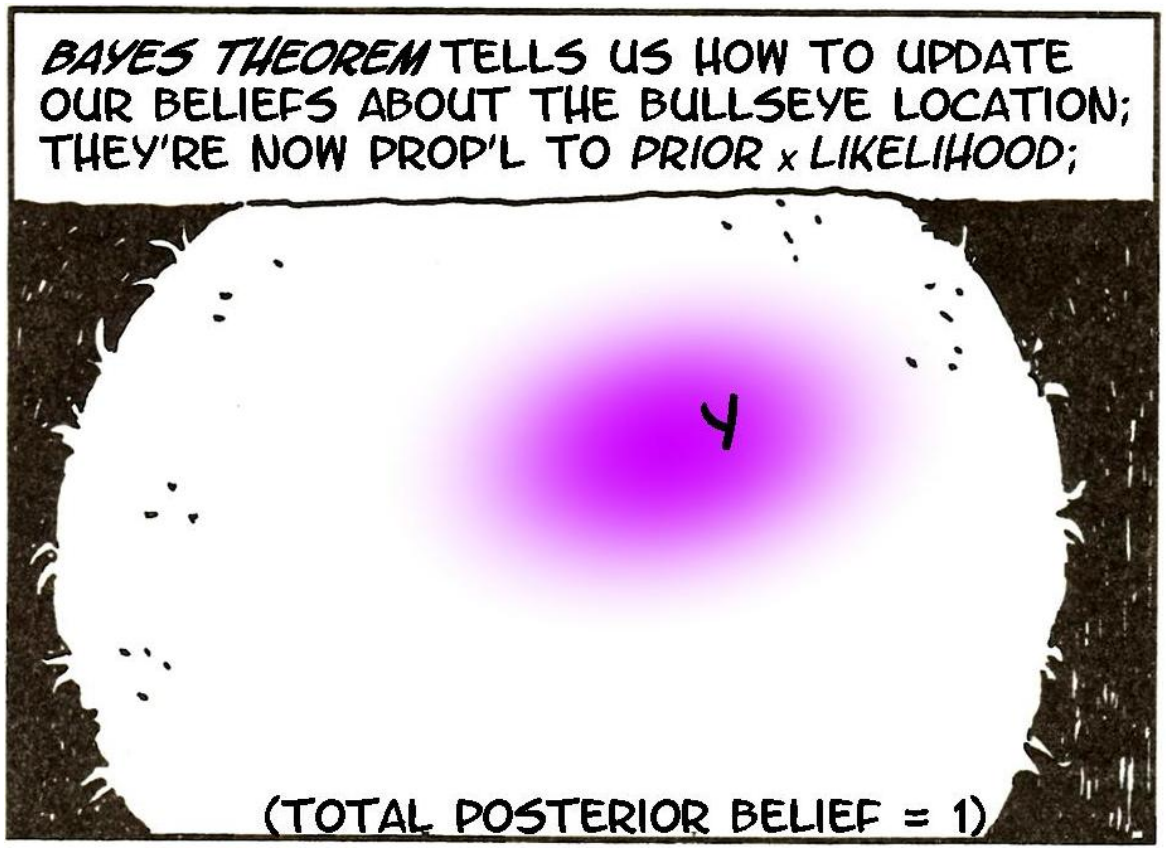
Bayesian inference

What to do when the data comes along?



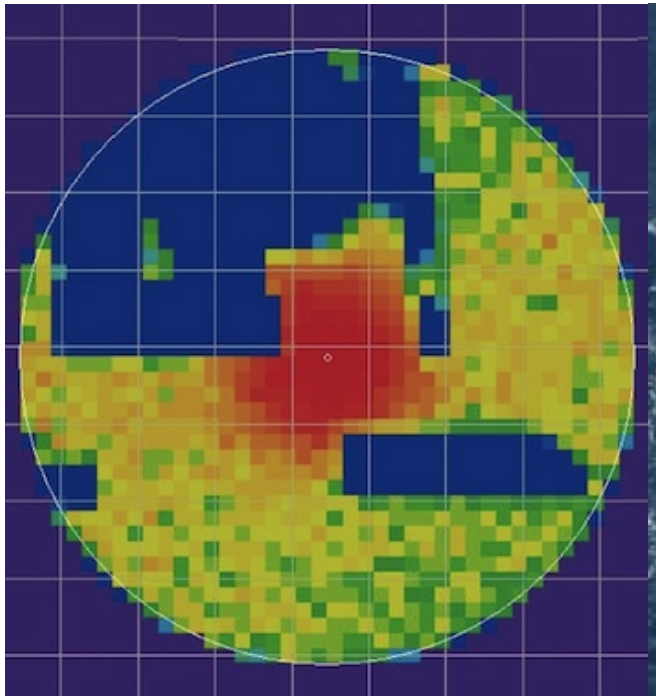
Bayesian inference

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 sampling & 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 You know about $\boldsymbol{\theta}$, combining the prior with the data – denoted $p(\boldsymbol{\theta}|\mathbf{Y})$? Bayes Theorem *used for inference* tells us to multiply;

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

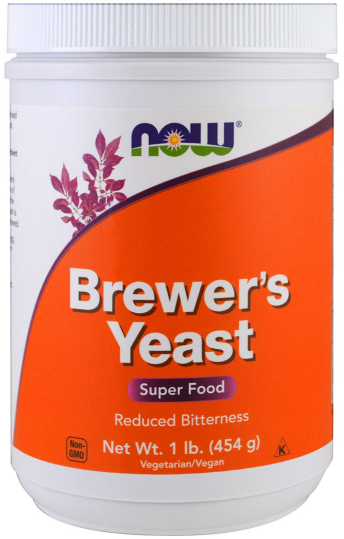
Bayesian inference

... 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 $\boldsymbol{\theta}$
- 'You' denotes any rational person who happens to hold the specified prior beliefs; given the observed data such a person *should* update these to the stated posterior – and it's irrational to believe anything else

Bayesian inference: ASE example

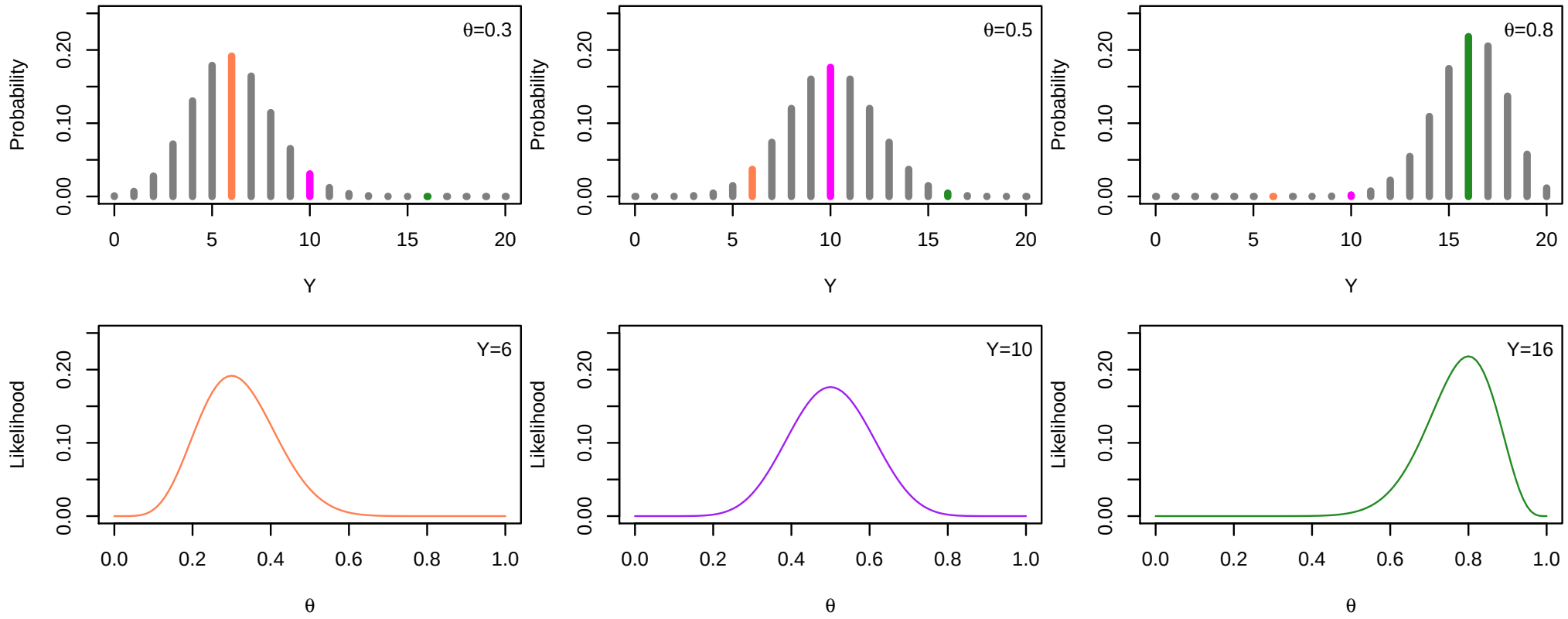
In an allele specific expression (ASE) experiment, 2 strains (BY and RM) are hybridized.



- N denotes the total number of expression reads at a particular location in the genome, Y denotes the number from BY
- We define θ as the probability a read come from BY (not RM)
- How far θ is from 0.5 determines how much allele specific expression there is

Bayesian inference: ASE example

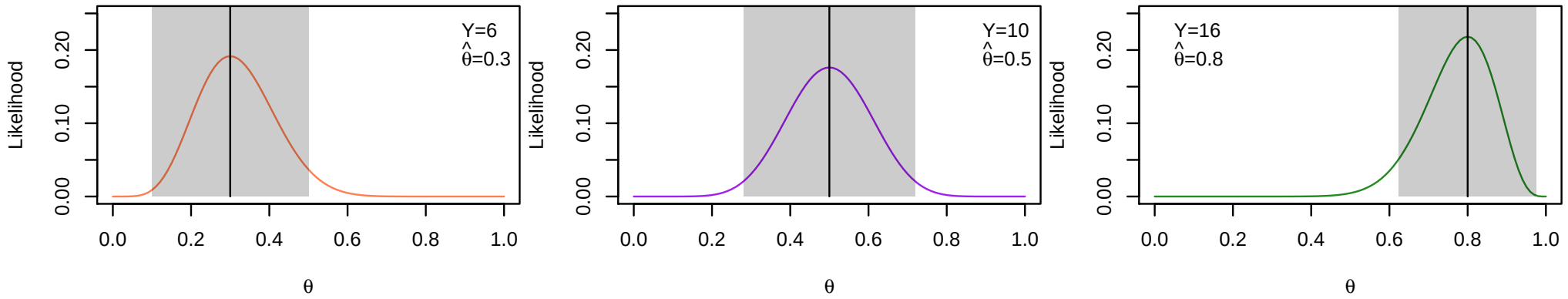
Sampling distribution, for several θ, and likelihood for several observations Y:



These are two ways of looking at $p(y|\theta)$ – varying y and varying θ.

Bayesian inference: ASE example

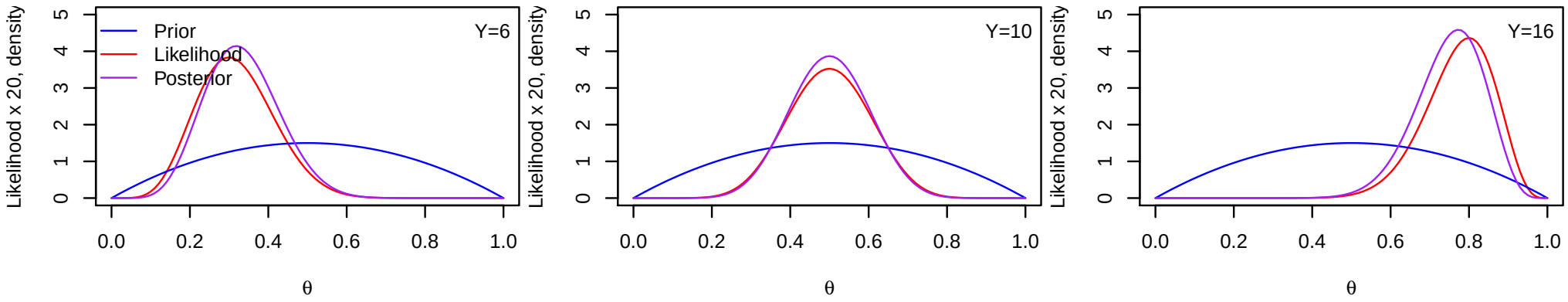
What does classical analysis do here?



- The point estimate (vertical line) is $\hat{\theta} = \bar{Y} = Y/N$, and an estimate of its standard error is given by $\sqrt{\hat{\theta}(1-\hat{\theta})/N}$.
- An approximate 95% confidence interval ("CI", shaded region) is given by $\hat{\theta} \pm 1.96 \times$standard error. This is an interval which, over many experiments, covers the true θ in 95% of them
- The analysis doesn't (& can't) tell us if any given experiment's CI is in the 95% or the 5%

Bayesian inference: ASE example

Here's one Bayesian analysis:



- This prior gives most support near $\theta = 0.5$ (mild allele-specific expression) decreasing to 0 at $\theta = 0, 1$ (expression impossible/guaranteed in BY)
- The prior's influence is to make results *slightly* more conservative than using likelihood alone
- Formally, this is statistical *induction*: reasoning from specific data to general population characteristics.
- Keen people: only relative size of likelihood & prior matters

Bayesian inference: how to summarize a posterior?

Reporting a full posterior $p(\theta|y)$ is too complex for most work. One helpful summary is a point estimate – our 'best guess' at θ, based on the posterior.

There are several definitions of 'best':

Posterior mean	Posterior median	Posterior mode
Center of mass of posterior	Halfway-point of posterior	High point of posterior
$\mathbb{E}[\,\theta\|Y = y\,] = \int \theta p(\theta\|y)$	$\theta' : \int_{-\infty}^{\theta'} p(\theta\|y) = 1/2$	$\text{argmax}_\theta\, p(\theta\|y)$

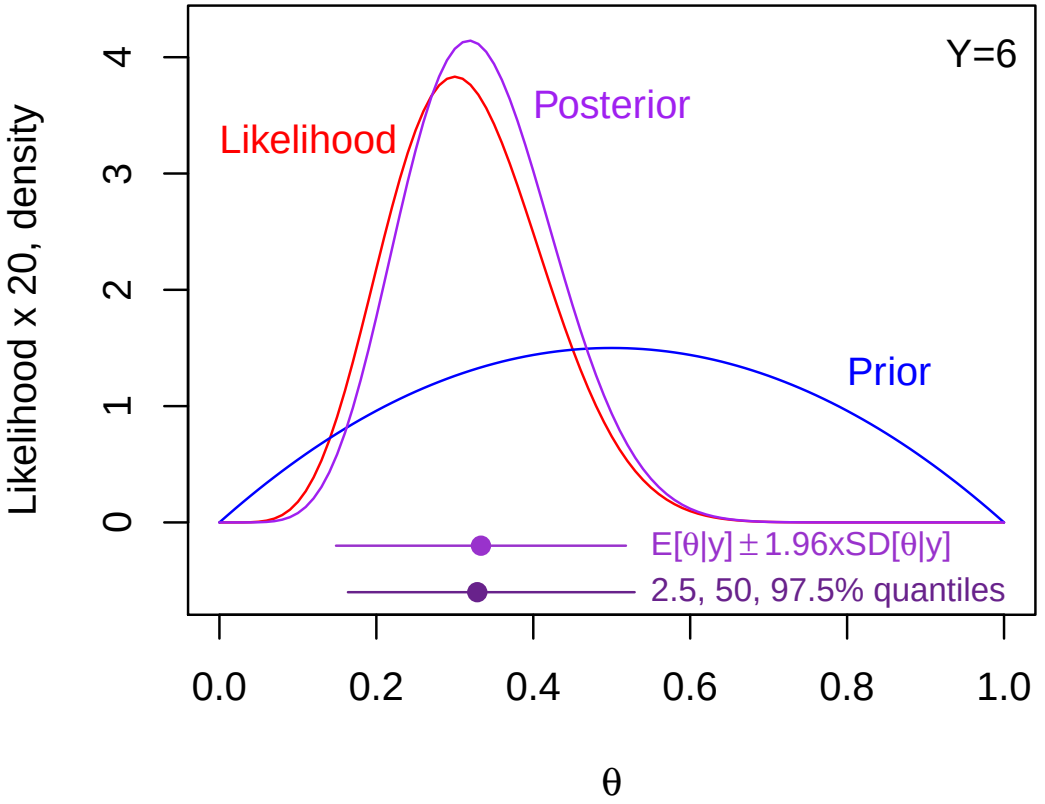
- For ≈symmetric unimodal posteriors, all 3 will be ≈similar. If in doubt, report the median
- Frequentist analysis typically uses the maximum likelihood estimate (MLE) that maximizes $p(y|\theta)$; same as posterior mode, if we have a flat prior

Bayesian inference: how to summarize a posterior?

To summarize posterior uncertainty, a natural analog of the standard error is the *posterior standard deviation*, $\text{StdDev}[\theta|Y=y] = \sqrt{\int(\theta - \mathbb{E}[\theta|y])^2 p(\theta|y)d\theta}$

If the posterior is ≈Normal, the interval $\mathbb{E}[\theta|Y=y] \pm 1.96\text{StdDev}[\theta|Y=y]$ contains approximately 95% of the posterior's support – an approximate 95% *credible interval*

More directly (and without relying on Normality) can calculate *central* 95% credible intervals as the 2.5%, 97.5% quantiles of the posterior.

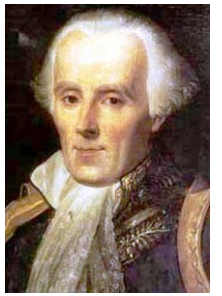


Bayesian inference: perhaps not so simple?

Bayesian inference can be made, er, transparent;

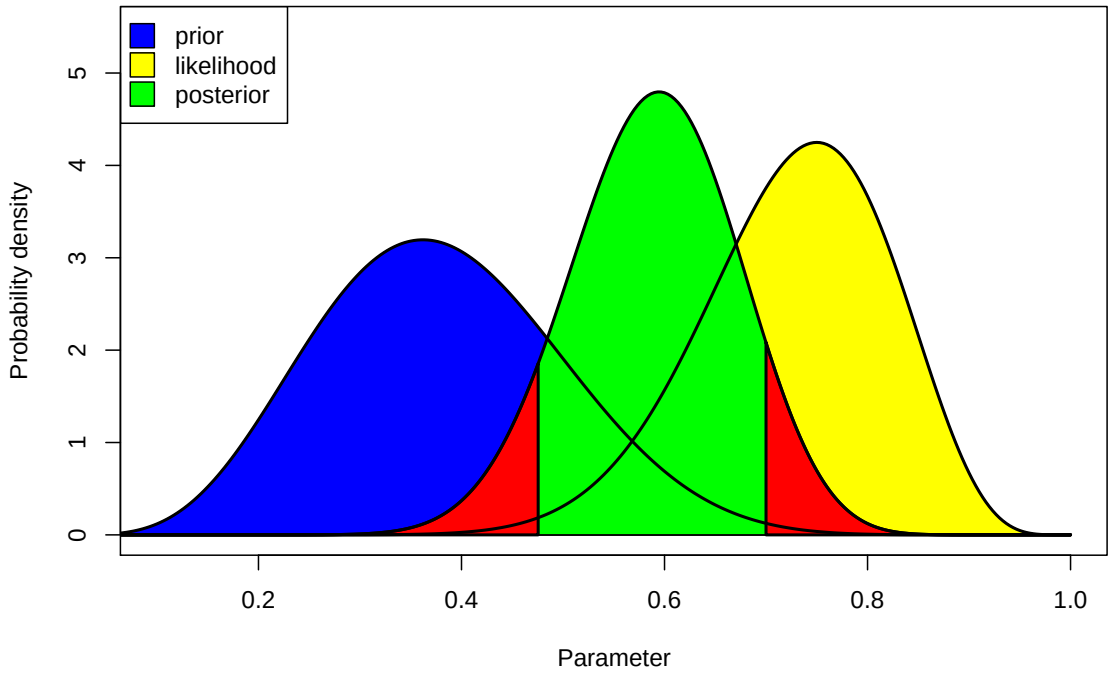
Common sense reduced to computation

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



Bayesian inference: perhaps not so simple?

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

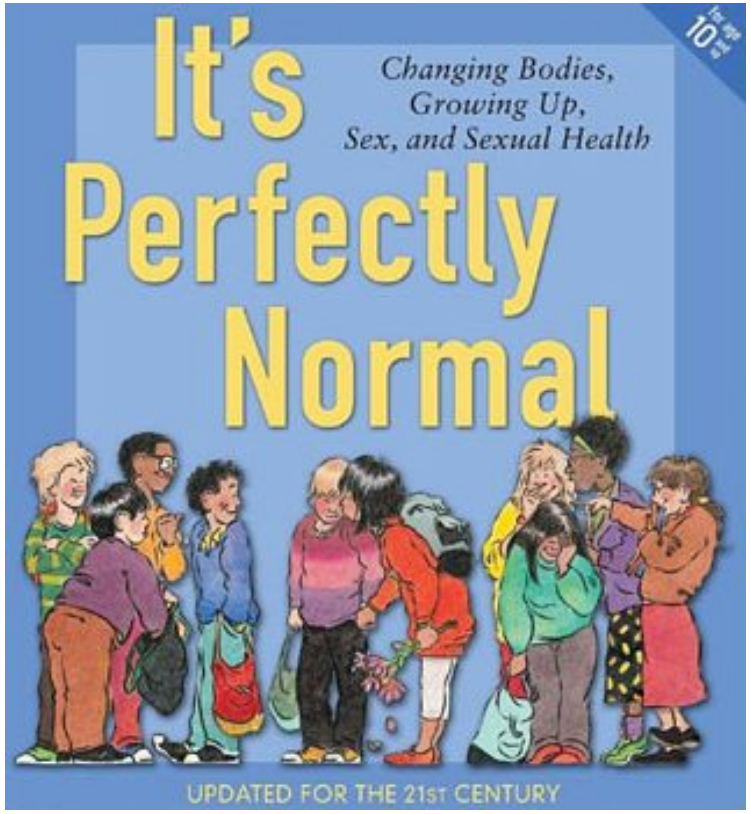


A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule

Stephen Senn, Statistician & Bayesian Skeptic (mostly)

Not so simple: where do priors come from?

An important day at statistician-school?



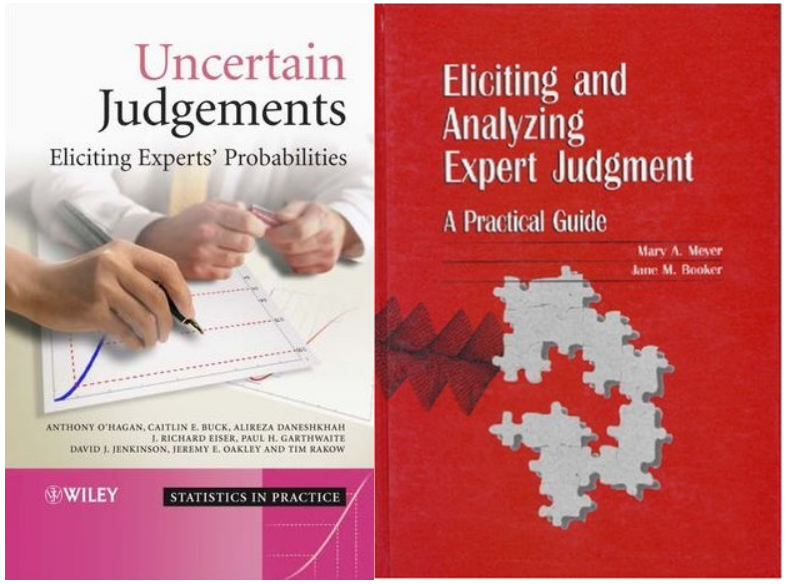
There's nothing wrong, dirty, unnatural or even *unusual* about making assumptions – carefully. Scientists & statisticians all make assumptions... even if they don't like to talk about them.

Not so simple: 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.

Like eliciting effect sizes for classical power calculations, it's not easy (see right) but here are some simple tips;

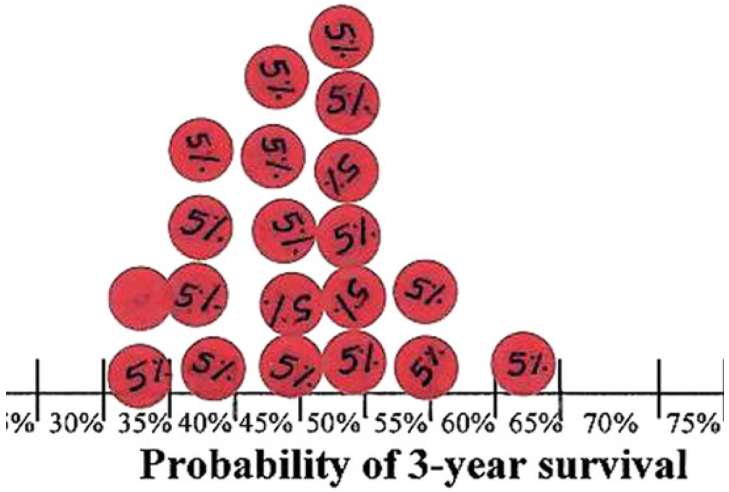


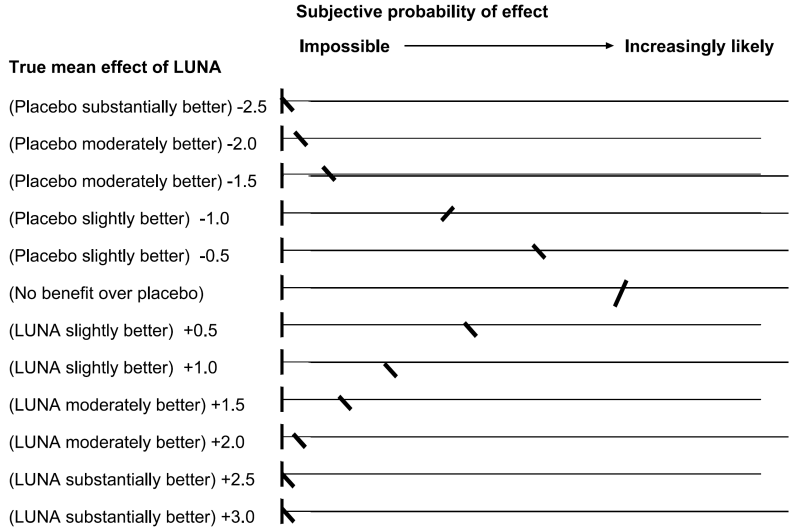
- Discuss parameters experts understand – e.g. code variables in familiar units, make comparisons relative to an easily-understood reference, *not* with age=height=IQ=0
- Avoid leading questions (just as in survey design)
- The 'language' of probability is unfamiliar; help users express their uncertainty

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

Not so simple: 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

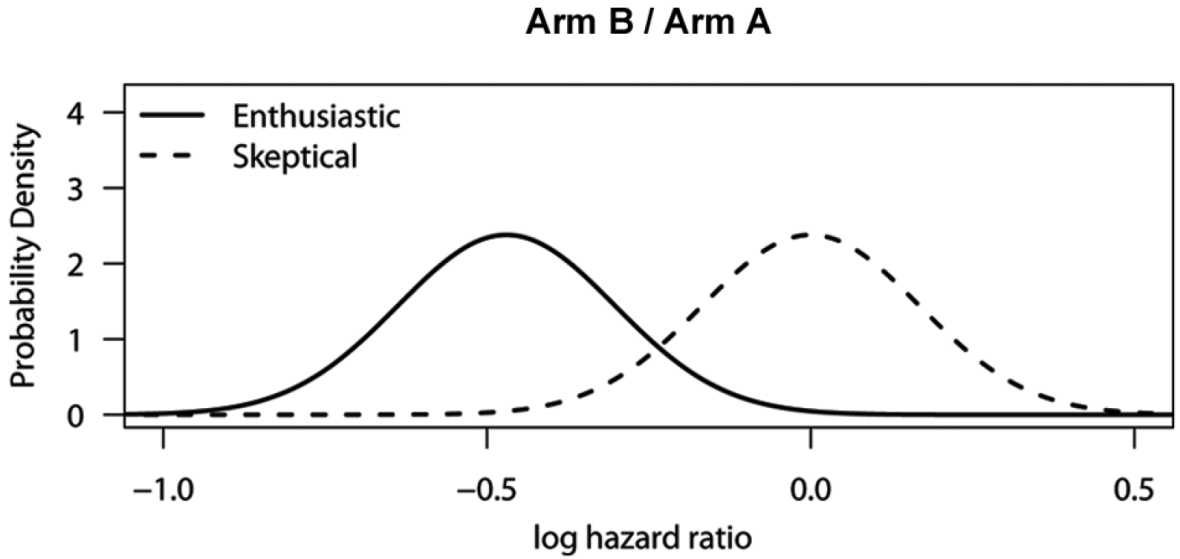
Normalize marks (Latthe *et al* 2005, J Obs Gync) for prior on pain effect of LUNA vs placebo

Typically these 'coarse' priors are smoothed. Providing the basic shape remains, exactly how much you smooth is unlikely to be critical in practice.

Not so simple: 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

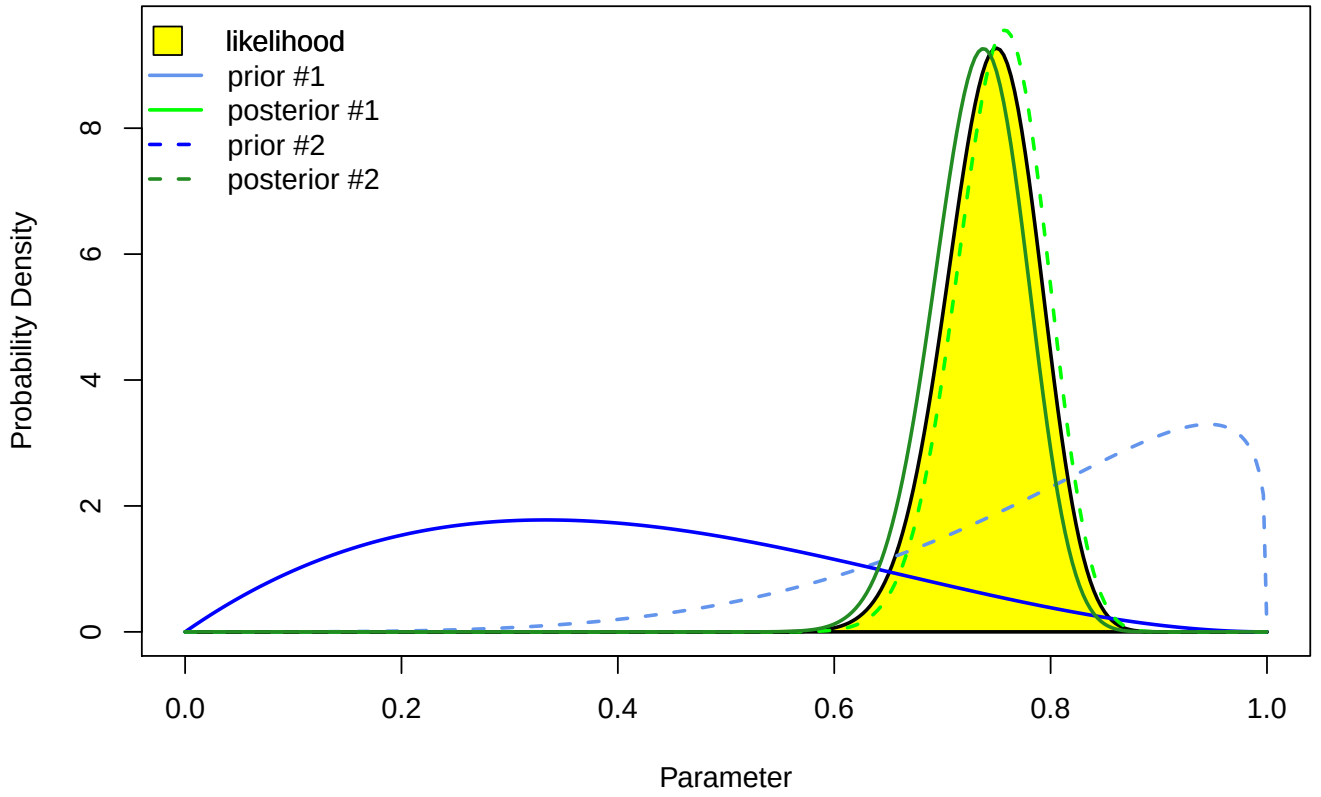


Known as 'Subjunctive Bayes'; if one had *this* prior and the data, *this* is the posterior one would have. If one had *that* prior... etc.

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*.

Not so simple: when don't priors matter? (*)

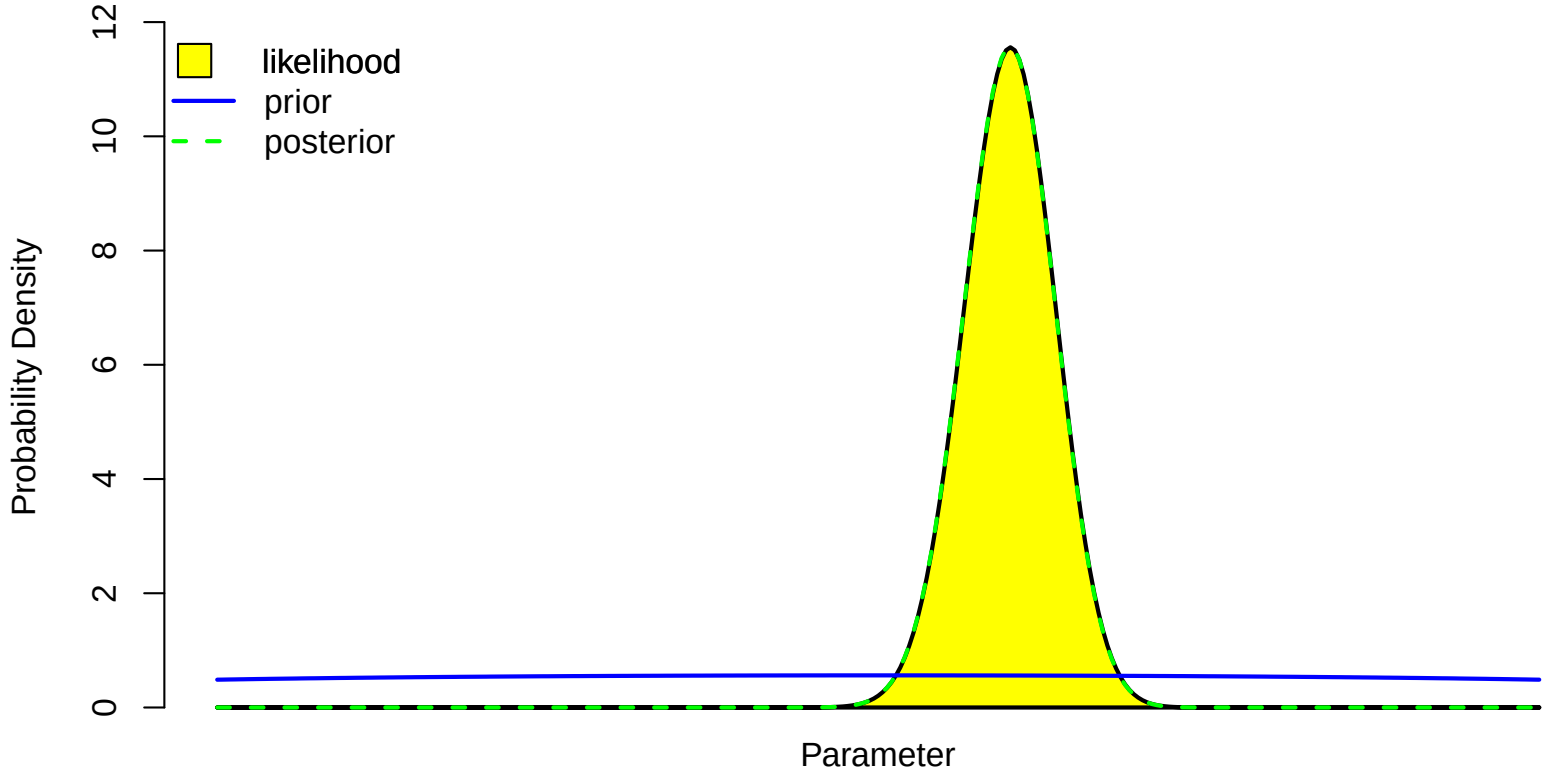
When the data provide a lot more information than the prior, this happens; (recall the stained glass color-scheme)



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

Not so simple: when don't priors matter? (*)

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

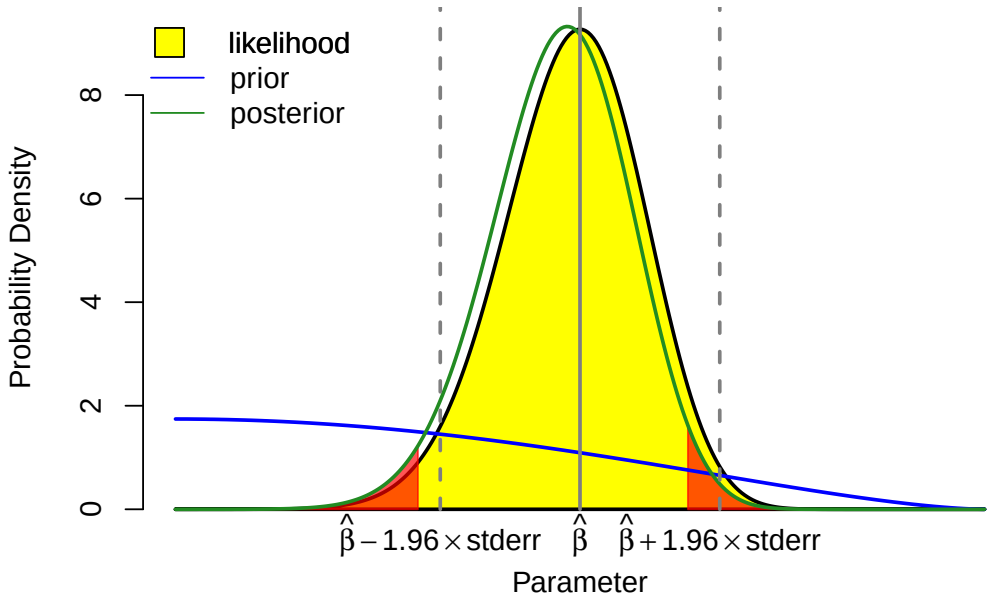


Not so simple: when don't priors matter? (*)

- Flat priors do NOT actually represent ignorance! Most of their support is for *very* extreme parameter values, and those can usually be ruled out with very rudimentary knowledge
- However, for parameters in 'famous' regression models, using flat priors to represent ignorance actually works okay. More generally, 'Objective Bayes' methods work to derive priors that are minimally-informative, though this is hard to define
- For many other situations, using flat priors works really badly – so be careful! (And also recall that prior elicitation is a useful exercise)

Not so simple: when don't priors matter? (*)

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*, Bayesian interpretations of 95% CIs are actually *okay*, i.e. saying we have ≈95% posterior belief that the true β lies within that range

* *and some regularity conditions*

Not so simple: when don't priors matter? (*)

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

One way to do this;

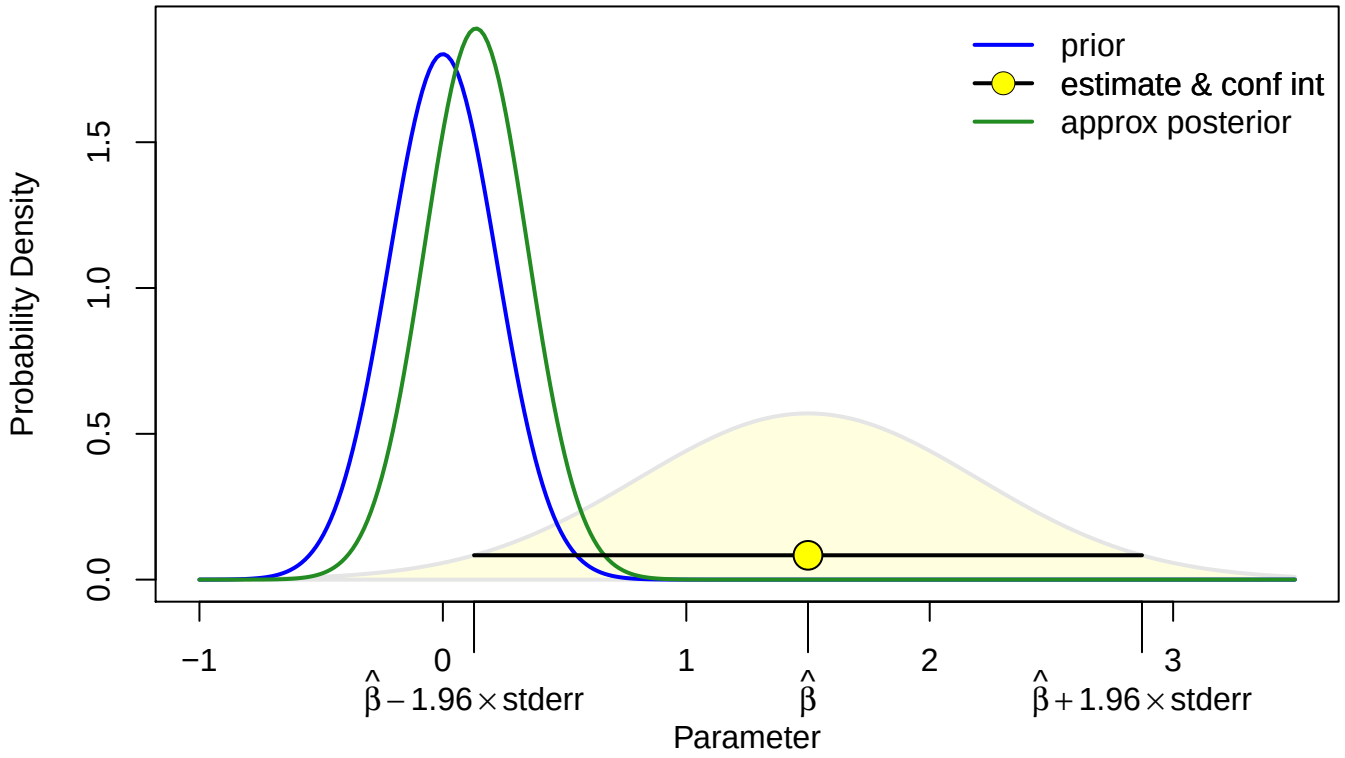
- Take point-estimate $\widehat{\beta}$ and corresponding standard error $stderr$, calculate precision $1/stderr^2$
- Elicit prior mean β_0 and prior standard deviation σ; calculate prior precision $1/\sigma^2$
- 'Posterior' precision $= 1/stderr^2 + 1/\sigma^2$ (which gives overall uncertainty
- 'Posterior'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

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

Not so simple: when don't priors matter? (*)

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

Not so simple: when don't priors matter? (*)

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

- Problems with the 'aggressive dissemination of noise' are a current hot topic...
- 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 $p < 0.05$

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

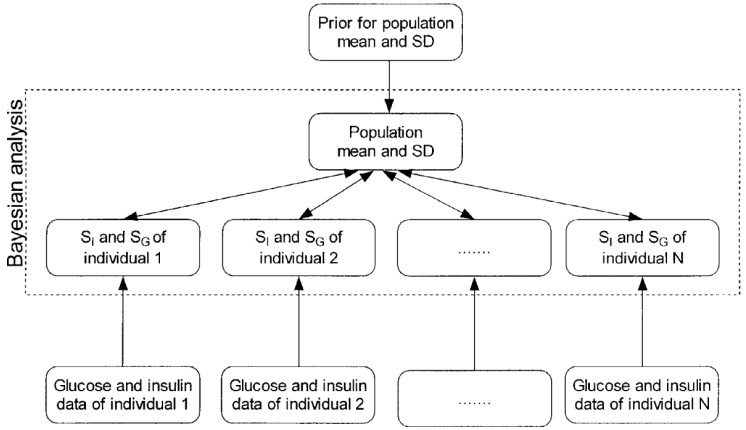
Many results that are rigorously proved and accepted start shrinking in later studies.

- 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? (*)

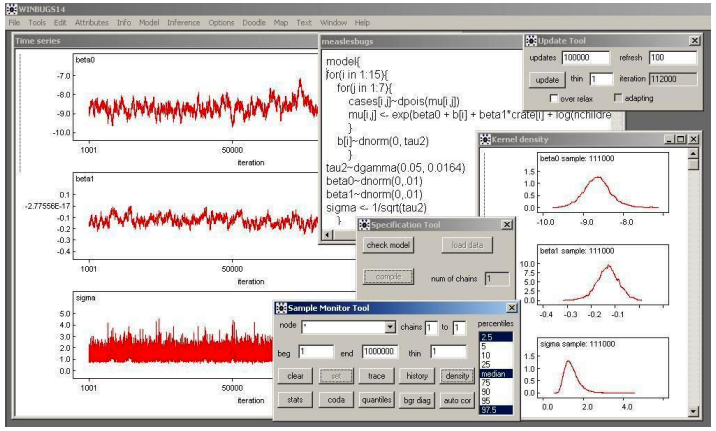
Allowing approximate Bayes, one answer is 'almost any analysis'. More-explicitly Bayesian arguments are often seen in;

Hierarchical modeling



One expert calls the classic frequentist version a "statistical no-man's land"

Complex 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 (Severini 1999, Rice 2004)
- Robust standard errors – like Bayesian analysis of a 'trend', at least for linear regression (Szpiro *et al* 2010)

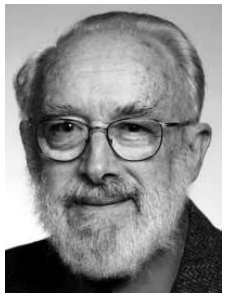
And some that can't;

- Many high-dimensional problems (shrinkage, machine-learning)
- Hypothesis tests ('Jeffrey's paradox') but NOT significance tests (Rice 2010)

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;



The only good statistics is Bayesian Statistics

Dennis Lindley (1923–2013)
writing about the future in 1975

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

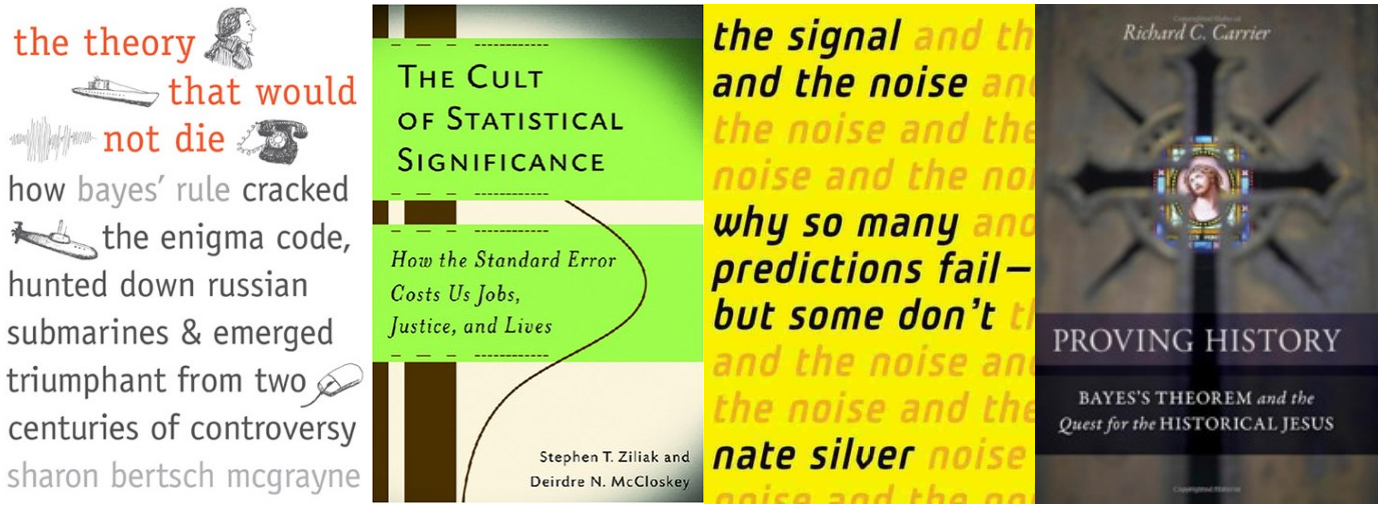
Maurice Kendall (1907–1983)
JRSSA 1968

- 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;

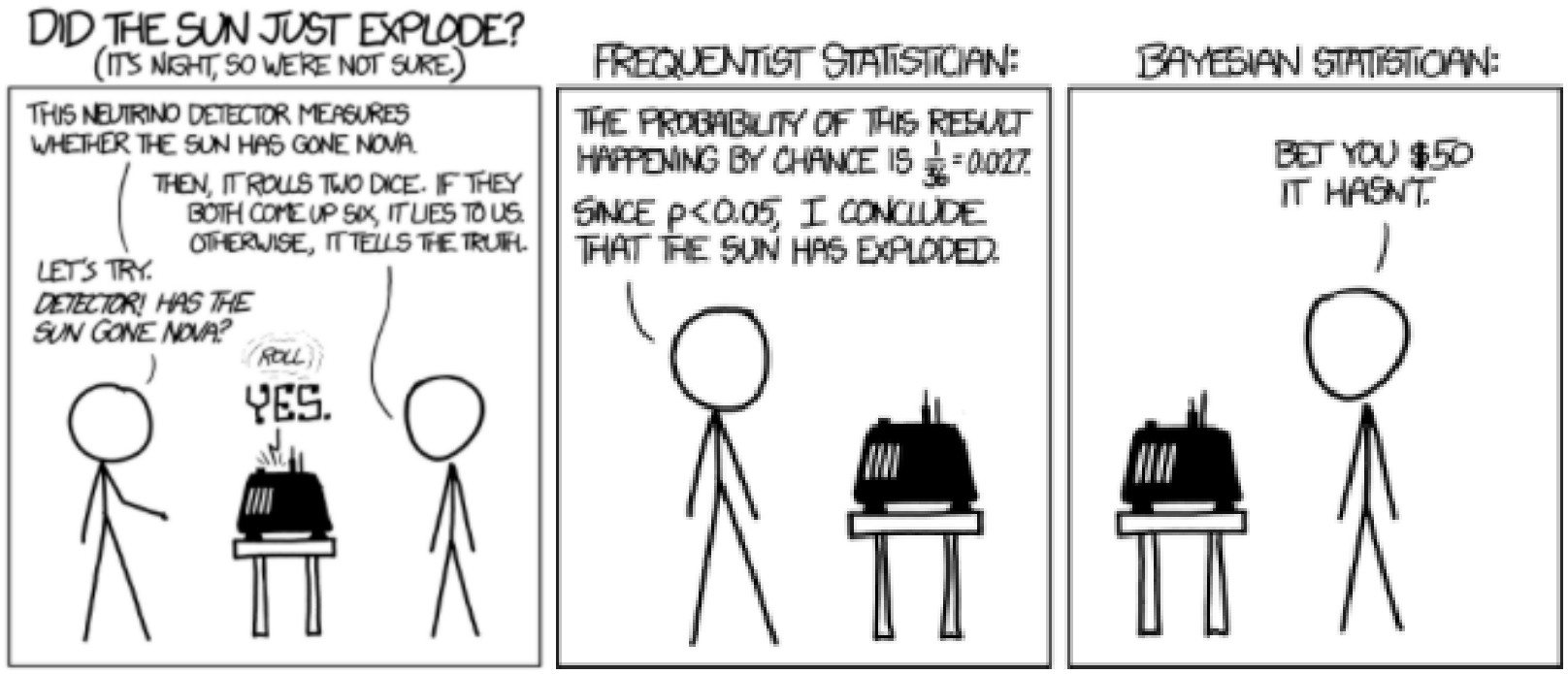


- Among those who actually understand both, it's hard to find people who totally dismiss either one
- Keen people: Vic Barnett's Comparative Statistical Inference provides the most even-handed exposition I know

Fight! Fight! Fight! (*)

XKCD on Frequentists vs Bayesians;

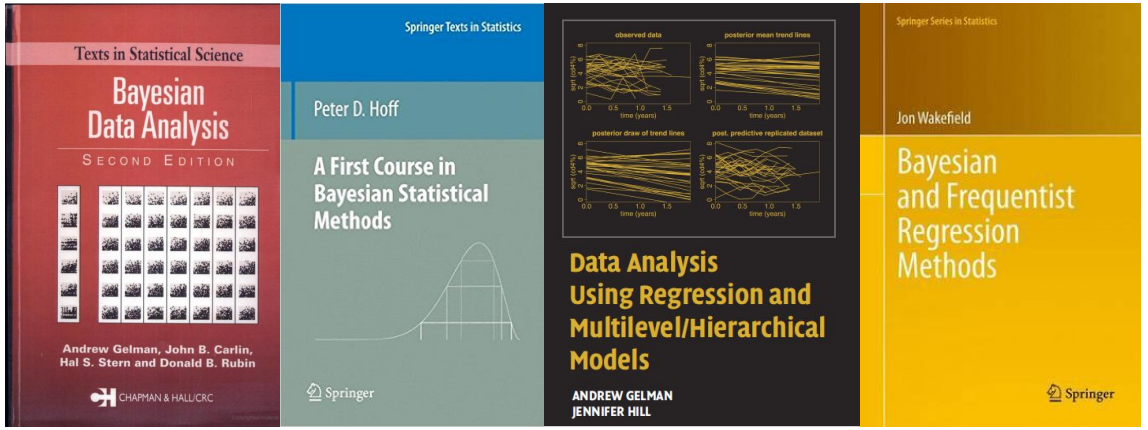
Here, the fun relies on setting up a straw-man; p-values are not the only tools used in a *skillful* frequentist analysis.



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