Bayesian Inference

Philippe Rast

PSC 204B
UC Davis, Winter 2018

Topics

Single-level Regression:

Week 1 Linear Regression (G&H: 3,4)
Week 2 Multiple Regression
Week 3 Violation of Assumptions
Week 4 Logistic Regression and GLM (G&H: 5, 6)
Week 5 Over-fitting, Information Criteria and Model comparison (McE: 6)
Week 6 Regression inference via simulations (G&H: 7–10)

Multilevel Regression:

Week 7 Multilevel Linear Models (G&H: 11–13)
Week 8 Multilevel Models (G&H: 14, 15)
Week 9 Multilevel Models & Bayesian Inference
Week 10 Fitting Models in Stan and brms (G&H: 16, 17 / McE: 11)

Overview

- Example: Globe Tossing
- Example: Globe Tossing

Bayesian data analysis

- Many ways to use the term "Bayesian"
  - Mainly it denotes a particular interpretation of probability.
  - ▷ Bayesian inference is no more than counting the numbers of ways things can happen, given our assumptions

  cf. "Garden of Forking Data"
- Use probability to describe uncertainty
  - Extends discrete logic (true/false) to continuous plausibility
- Computationally difficult
  - Markov chain Monte Carlo (MCMC) to the rescue
- Used to be controversial
  - Ronald Fisher: Bayesian analysis "must be wholly rejected."

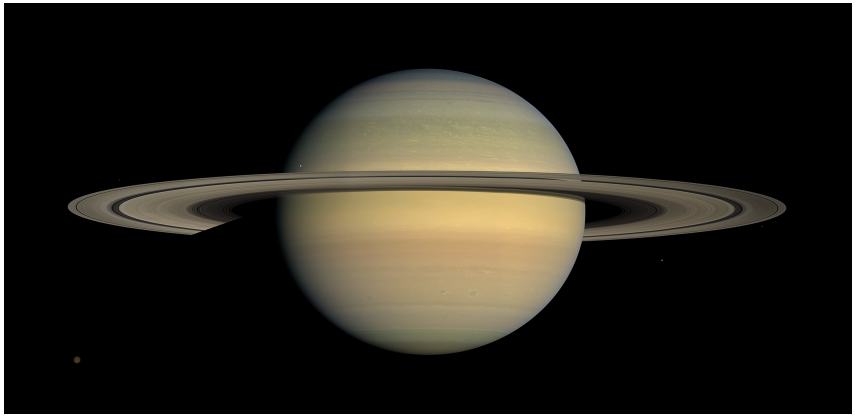
Bayesian data analysis

Frequentist approach to probability:

- Requires that all probabilities be defined by connection to countable events and their frequencies in very large samples.
- Leads to frequentist uncertainty being premised on imaginary resampling of data
  - ▷ if we were to repeat the measurement many many times, we would end up collecting a list of values that will have some pattern to it.
- Implication: Parameters and models cannot have probability distributions, only measurements can
- Distribution of these measurements: *sampling distribution*.
- Resampling is hardly ever done – and hardly feasable
- "the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination" (Ronald Fisher, 1956, quoted in McElrath)

Bayesian data analysis

Some thoughts about uncertainty



Saturn as viewed by Cassini spacecraft in 2008

Saturn as viewed by Galileo (maybe) in 1610

Bayesian data analysis

Contrast with frequentist view

- Probability is just limiting frequency
- Uncertainty arises from sampling variation
- Uncertainty about the planet's shape and rings,
  - *none* of the uncertainty is a result of variation in repeat measurements
- Sampling distribution of any measurement is constant, because the measurement is deterministic!
- There is nothing "random" about it.
- Frequentist statistical inference needs to relegate variation to uncertainty arising from repeated sampling

Bayesian data analysis

Bayesian probability much more general

- Probability is in the model, not in the world
- Coins are not random, but our ignorance makes them so
- Deterministic "noise" can still be modeled using probability, as long as probability is not identified with frequency

i.e. Variation is part of the parameter, part of the model

- **Important:** Even if a Bayesian procedure and frequentist procedure give exactly the same answer, Bayesian inference is not based on imagined repeat sampling.
  - ▷ More generally, in Bayesian models "randomness" is as a property of information, not of the world.

What is Bayesian Statistics?

- A philosophy of statistics.
- A generalization of classical statistics.
- An approach to statistics that explicitly incorporates previous knowledge in modeling data.
- A different method for fitting models.
- An easier approach to statistics.
- Generally: A more natural way to interpret probabilities
- But: Only because Bayesian approach is justifiable, does not mean that all other approaches are invalidated!
- It's just a model after all...

Globe Tossing Example

McElrath

- Toss a model of a globe into air
- Catch it - record whether index finger lands on water (W) or land (L)
- Repeat toss a number of times
- Write down sequence
- ▷ Data: `W L W W W L W L W`
- Obtain probability of how much of the earth is covered in water

- For WLWWWLWLW:
  - Some true proportion of water, p
  - Toss globe, probability p of observing W, $1 - p$ of L
  - Each toss therefore independent of other tosses

Globe Tossing Example

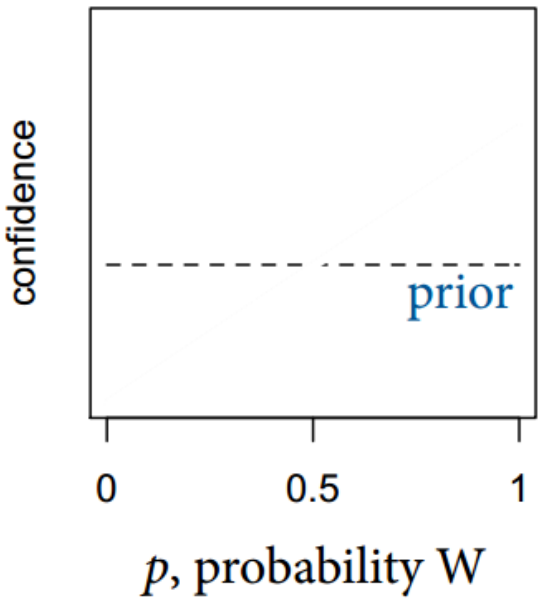
McElrath

Approach probability - from toss to toss

- *Bayesian updating*
- ▷ Defines optimal learning in model – converts prior into posterior
- Give our model an information state, before the data:
- Here, an initial confidence in each possible value of p between zero and one
- Condition on data to update information state:
- New confidence in each value of p, conditional on observed data

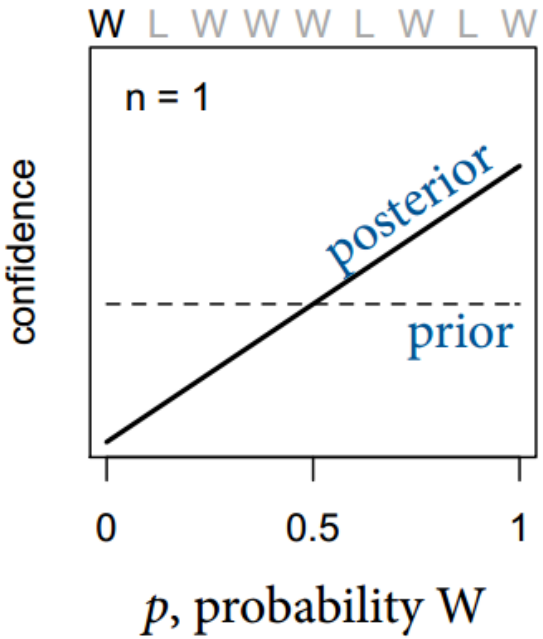
Globe Tossing Example

McElrath



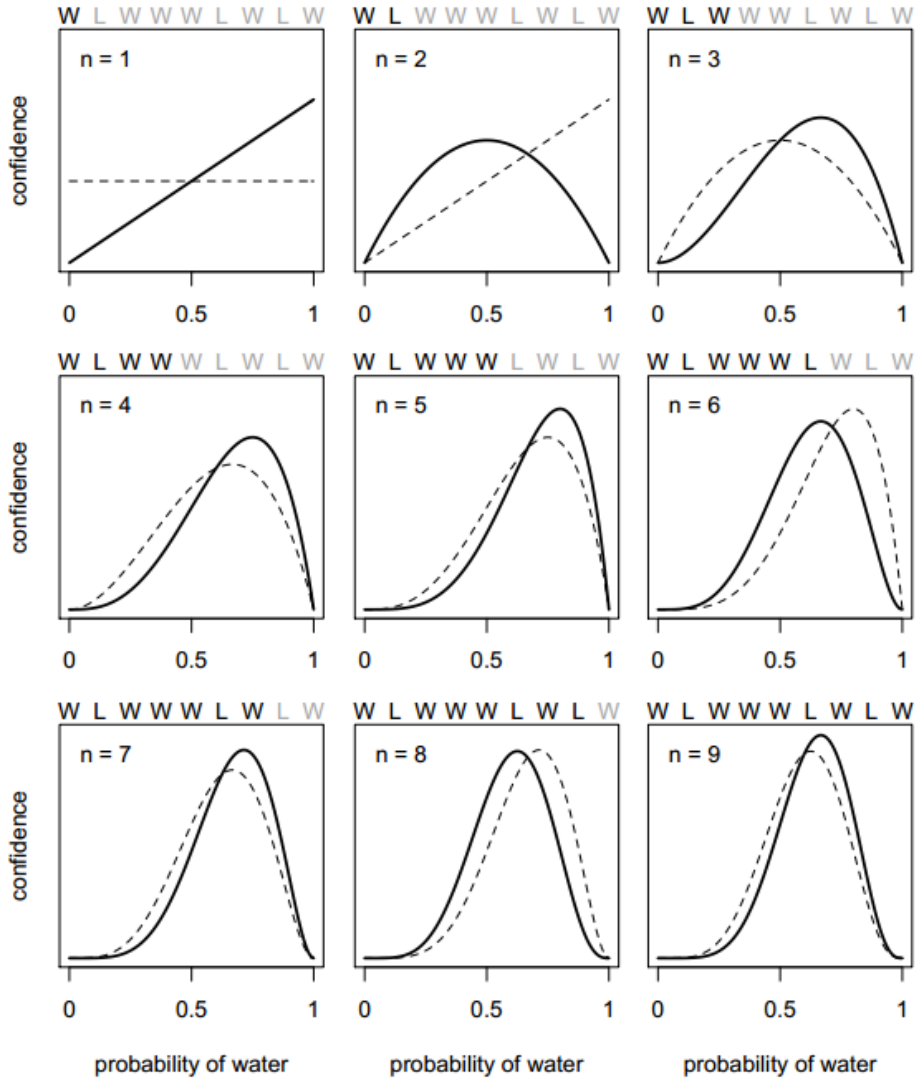
Globe Tossing Example

McElrath



Globe Tossing Example

McElrath



Globe Tossing Example

McElrath

- Assume:

1. Likelihood
2. Parameters
3. Prior

- Deduce: Posterior

Prior to Posterior

Bayes Theorem

- Prior – which parameter values you think are likely and unlikely.
- Collect data.
- Data gives us *Likelihood* – which parameter values the data consider likely
- Update prior to *Posterior* – what values you think are likely and unlikely given prior info and data.
- Prior and posterior are both probability distributions
- Bayes Theorem: Posterior $=$ Likelihood $\times$ Prior $\times$ constant

Bayes' Theorem

- In 1763, Thomas Bayes published a paper on the problem of induction
  - arguing from the specific to the general.
- In modern language and notation, Bayes wanted to use Binomial data comprising r successes out of n attempts to learn about the underlying chance θ of each attempt succeeding.
- Bayes' key contribution was to use a probability distribution to represent uncertainty about θ.
- This distribution represents 'epistemological' uncertainty, due to lack of knowledge about the world, rather than 'aleatory' probability arising from the essential unpredictability of future events..
- Modern 'Bayesian statistics' is still based on formulating probability distributions to express uncertainty about unknown quantities.
- These can be underlying parameters of a system (induction) or future observations (prediction).

Bayes' Theorem

$$\Pr[\theta|y] = \frac{\Pr[y|\theta] \times \Pr[\theta]}{\Pr[y]}$$

- $\Pr(\cdot)$ denotes the probability distribution
- $\Pr(\cdot|\cdot)$ denotes a conditional probability
- When y represents data and θ represents parameters in a statistical model, Bayes Theorem provides the basis for Bayesian inference.
- The 'prior' distribution $\Pr[\theta]$ (epistemological uncertainty) is combined with 'likelihood' $\Pr[y|\theta]$ to provide a 'posterior' distribution $\Pr[\theta|y]$ (updated epistemological uncertainty)
  - ▷ The likelihood is derived from an aleatory sampling model $\Pr[y|\theta]$ but considered as function of θ for fixed y.

Bayes' Theorem

- Practical use of the Bayesian approach requires consideration of several practical issues
- Source of the prior distribution
- Choice of a likelihood function
- Computation and summary of the posterior distribution in high-dimensional problems

Bayes' Theorem For Two Events

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

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]}$$

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.

Little League

- $\Pr[A|B]$ Probability of hitting the ball (A), given that the ball is in the strike zone (B)
- $\Pr[B|A]$ Probability of ball being in strike zone, given that it was hit
  - In 10 good pitches, about 7 hits (7/10)
- $\Pr[A]$ probability of hitting the ball:
  - Of 10 pitches, about 1 hit on any pitch (1/10)
- $\Pr[B]$ probability of good pitch:
  - Of 10 pitches, about 7 foul and 3 good (3/10)
- A terrible night...

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} = \frac{7/10 \times 1/10}{3/10} = \frac{7/100}{3/10} = \frac{7}{30} = .23$$

Little League

Good hitter:

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} = \frac{4/10 \times 7/10}{3/10} = \frac{28}{30} = .93$$

Good pitcher:

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} = \frac{8/10 \times 7/10}{9/10} = \frac{56}{90} = .62$$

Bayes' Theorem

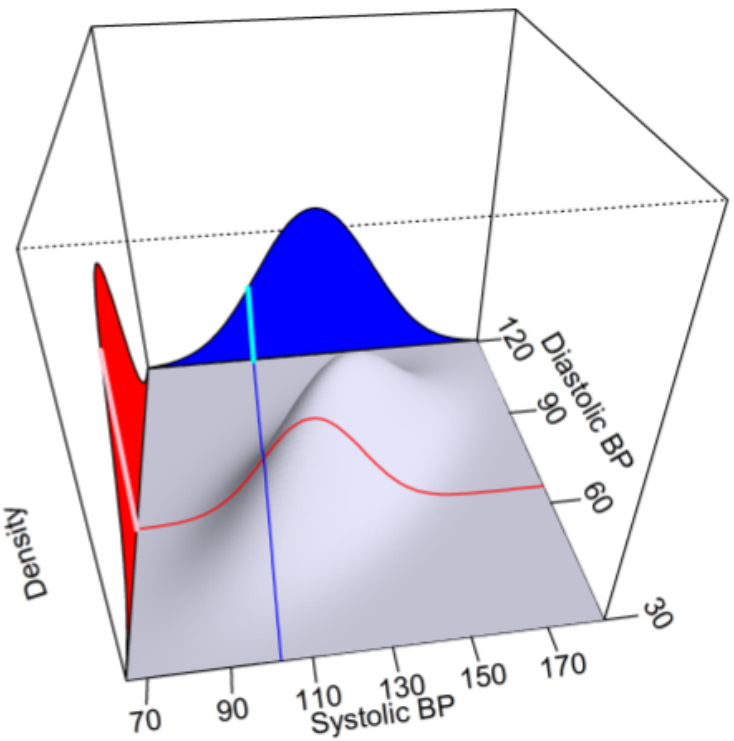
Why does it matter? If 1% of a population have cancer, for a screening test with 80% sensitivity (probability of detection) and 95% specificity (true-negatives, percentage of healthy people who are correctly identified as not having the condition);

- $\Pr[\text{Tested positive}|\text{Cancer}] = .80$
- $\Pr[\text{Tested positive}] = .05$
- $\Pr[\text{Cancer}] = 0.01$
- $\Pr[\text{Cancer}|\text{Tested Positive}] = \frac{.80 \times 0.01}{0.05} = .16$
- In words: Probability of having cancer when tested positive is 16%, i.e., 84% are false positives!

- Prosecutor's fallacy:
  $\Pr[\text{Tested positive}|\text{Cancer}] = \Pr[\text{Cancer}|\text{Tested Positive}]$
  - ▷ a small probability of evidence given innocence need NOT mean a small probability of innocence given evidence.

Bayes' Theorem

Bayes' theorem also applies to continuous variables – say Systolic and Diastolic blood pressure
preface



The conditional densities of the random variables are related this way

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

which can be written 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'.

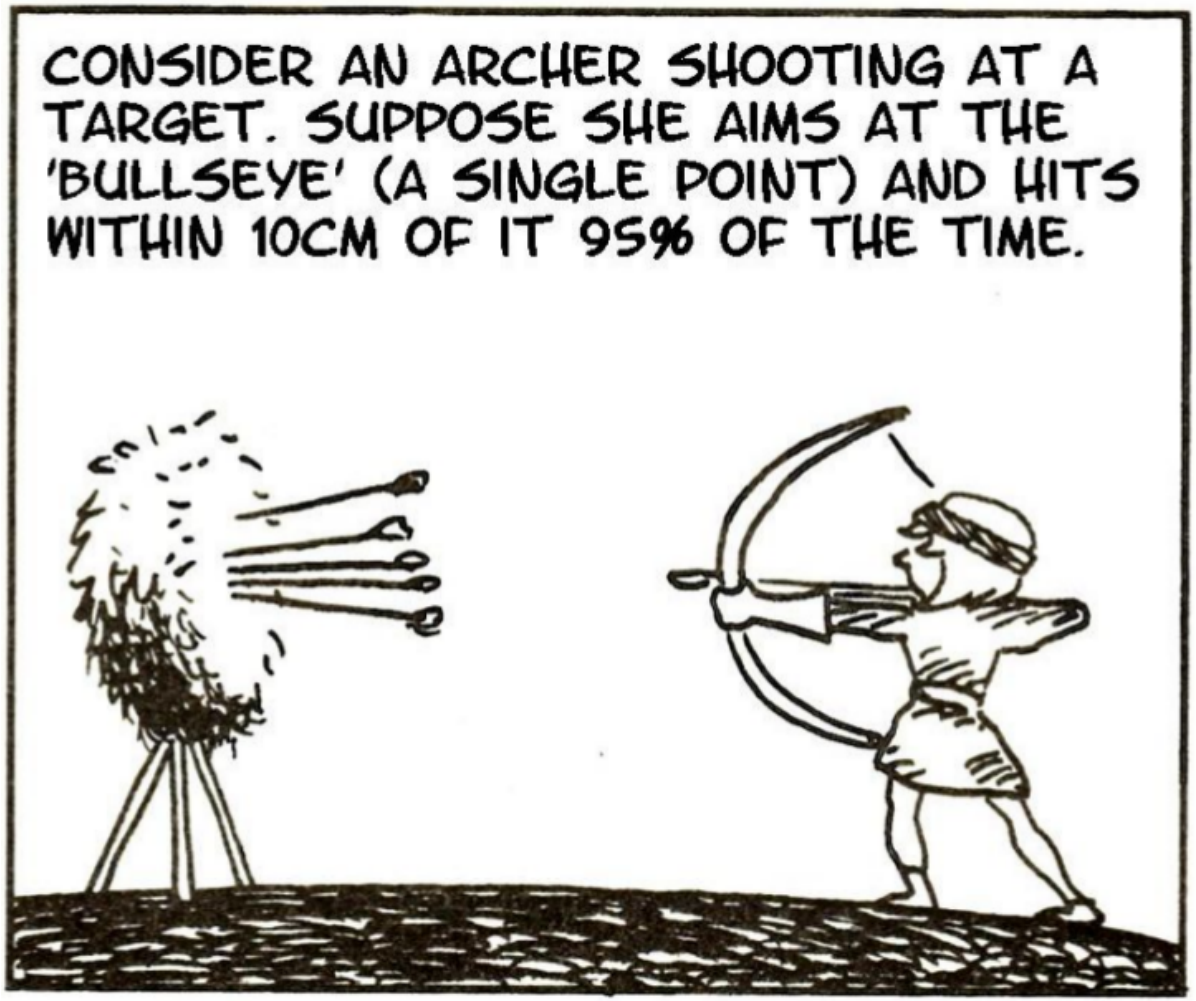
Bayesian inference

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.

- So why all the fuss?
- 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.

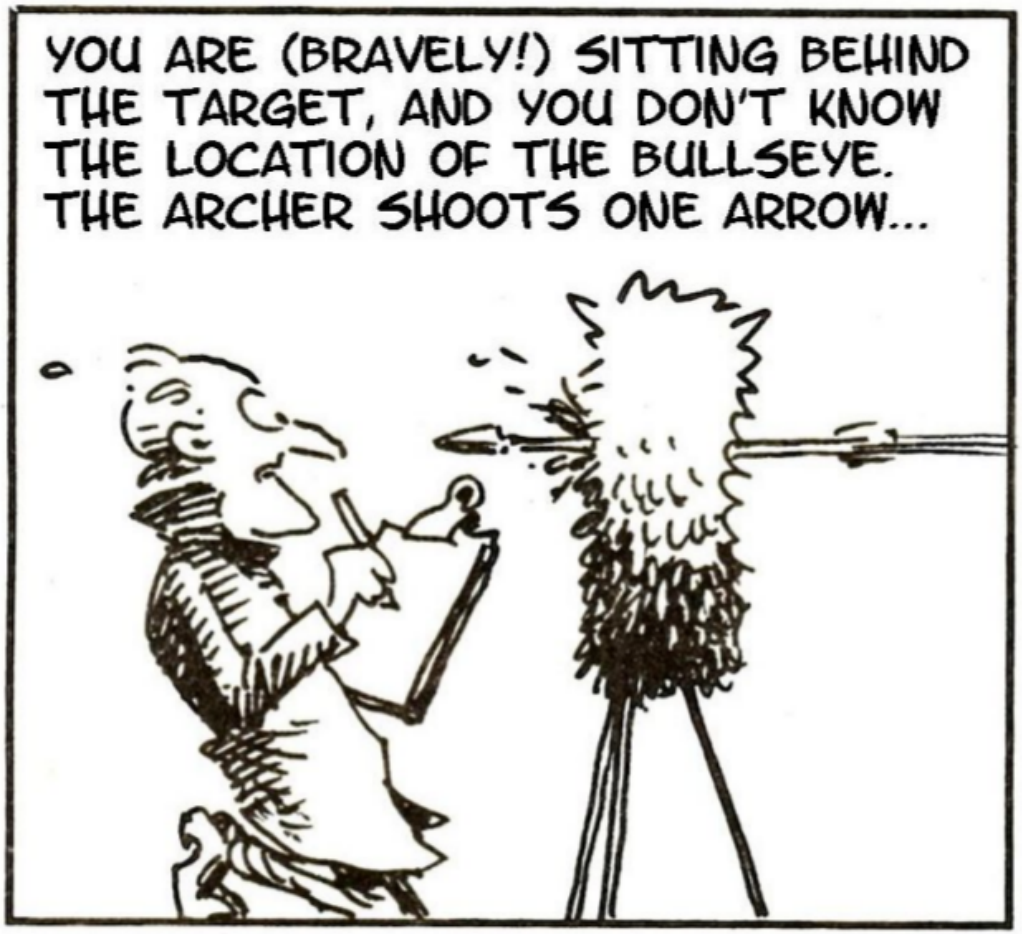
- Note: Frequentist inference, e.g. using p-values & confidence intervals, does *not* quantify what is known about parameters.

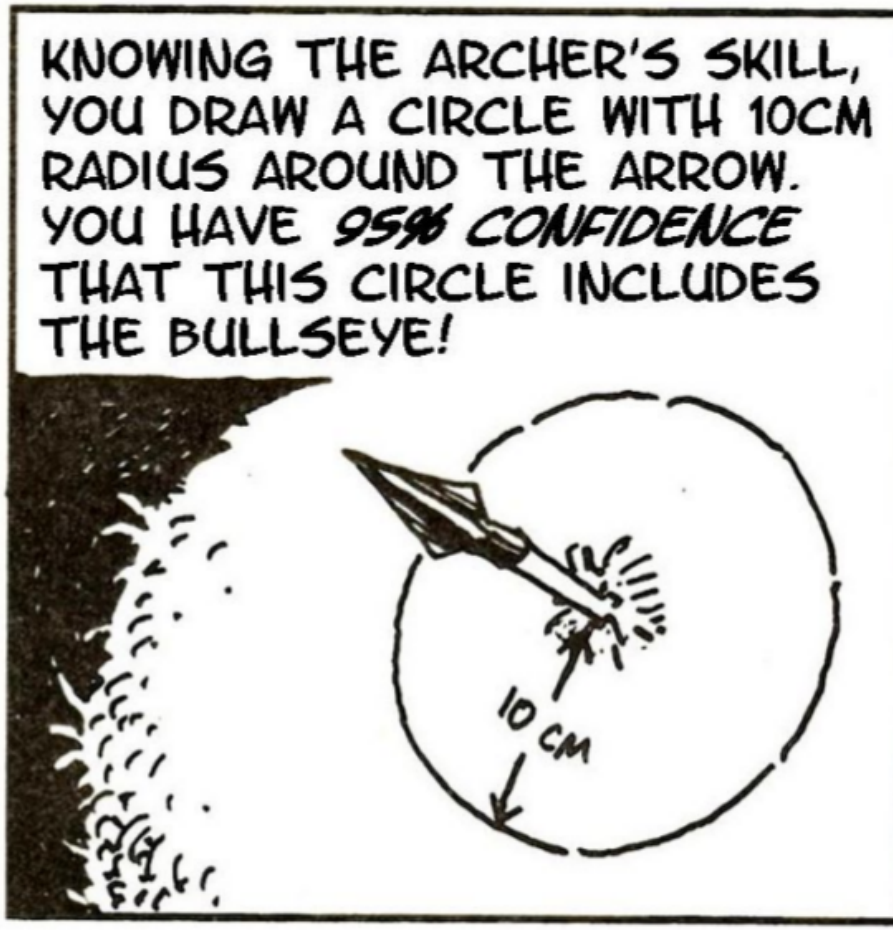
Frequentist Inference



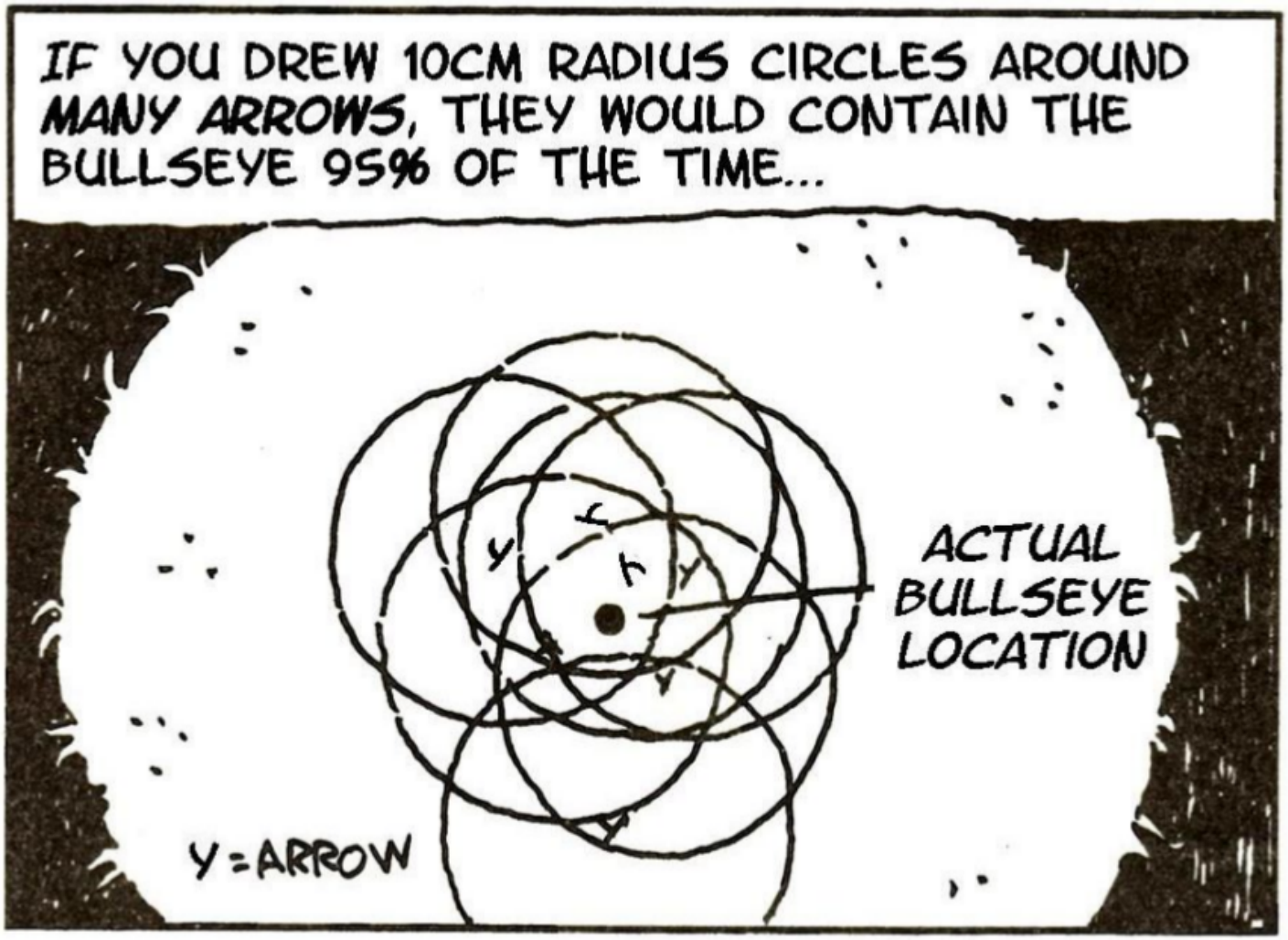
Adapted from Gonick & Smith, The Cartoon Guide to Statistics and Ken Rice at UW

Frequentist Inference



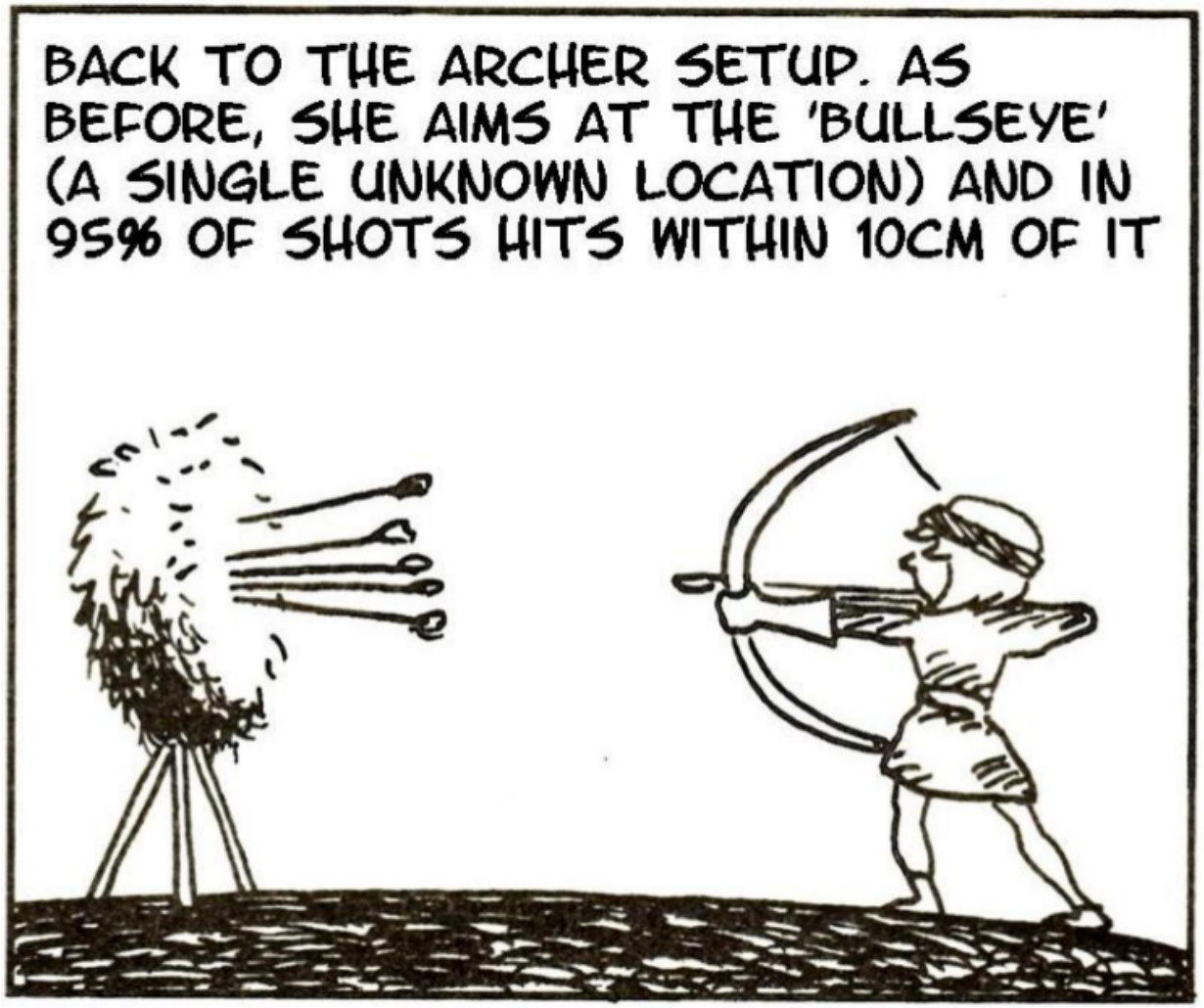


Frequentist Inference



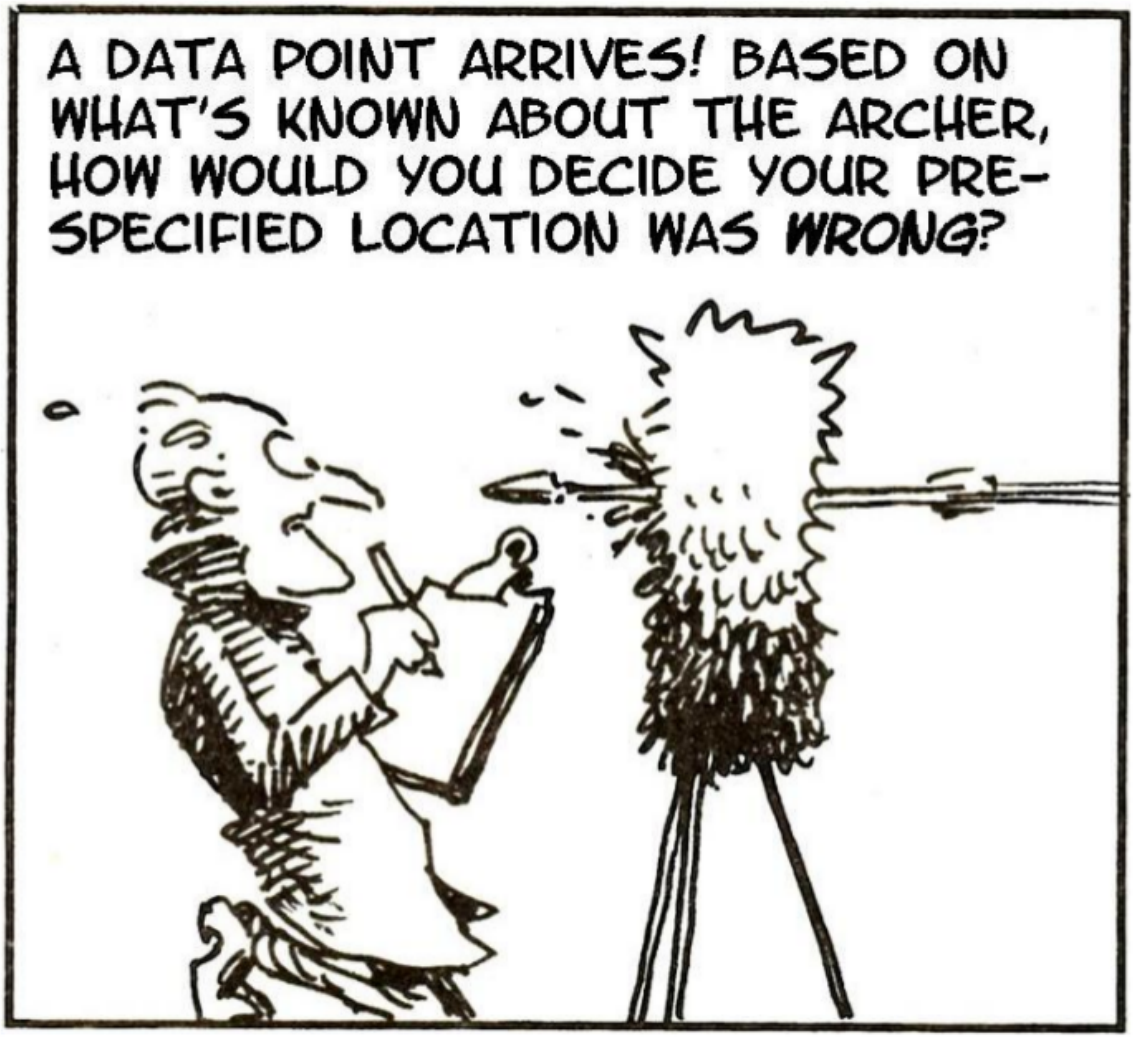
The interval traps the truth in 95% of experiments. To define anything frequentist, you have to imagine repeated experiments.

Frequentist Inference

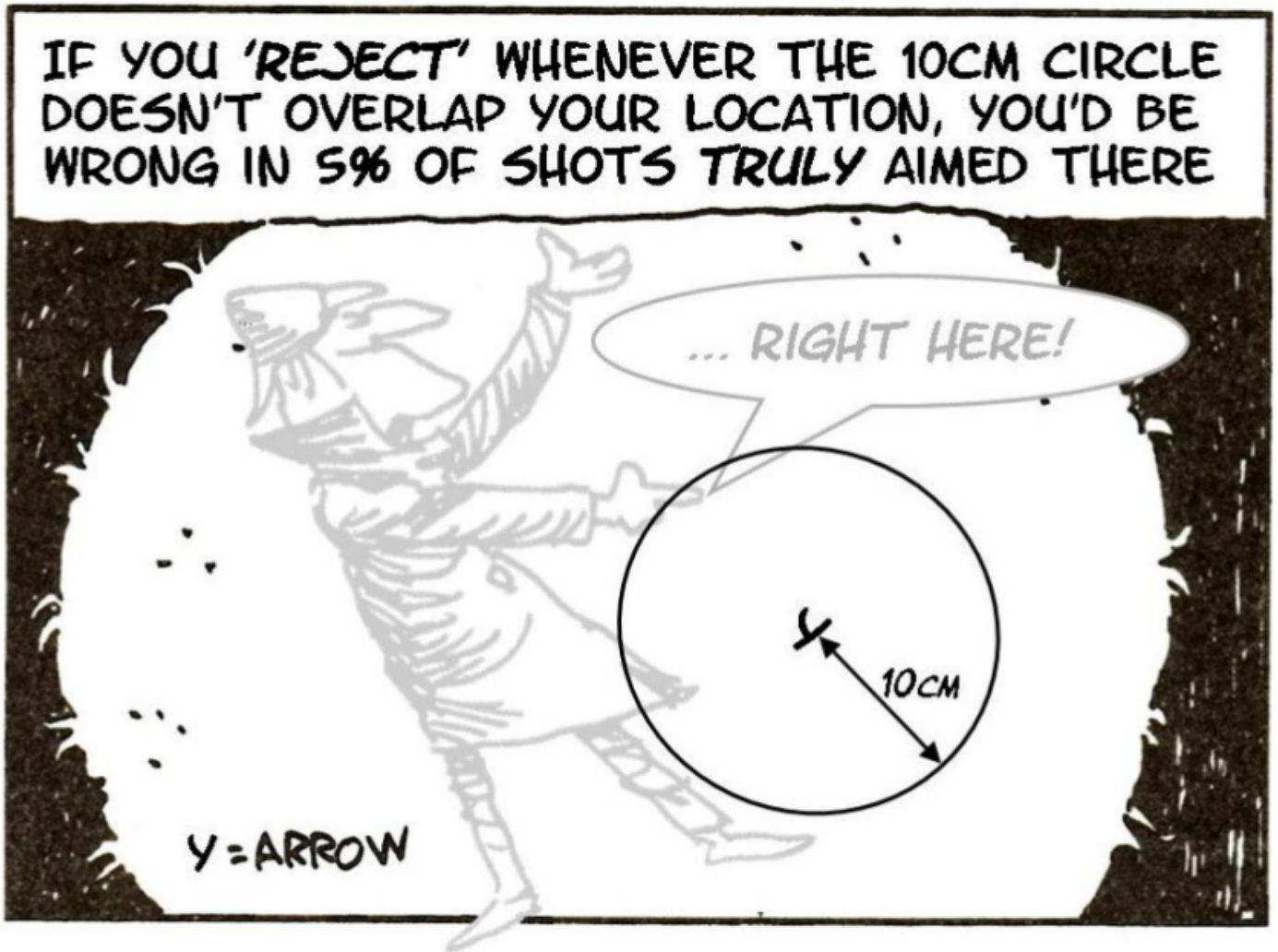


Frequentist Inference

Frequentist Inference



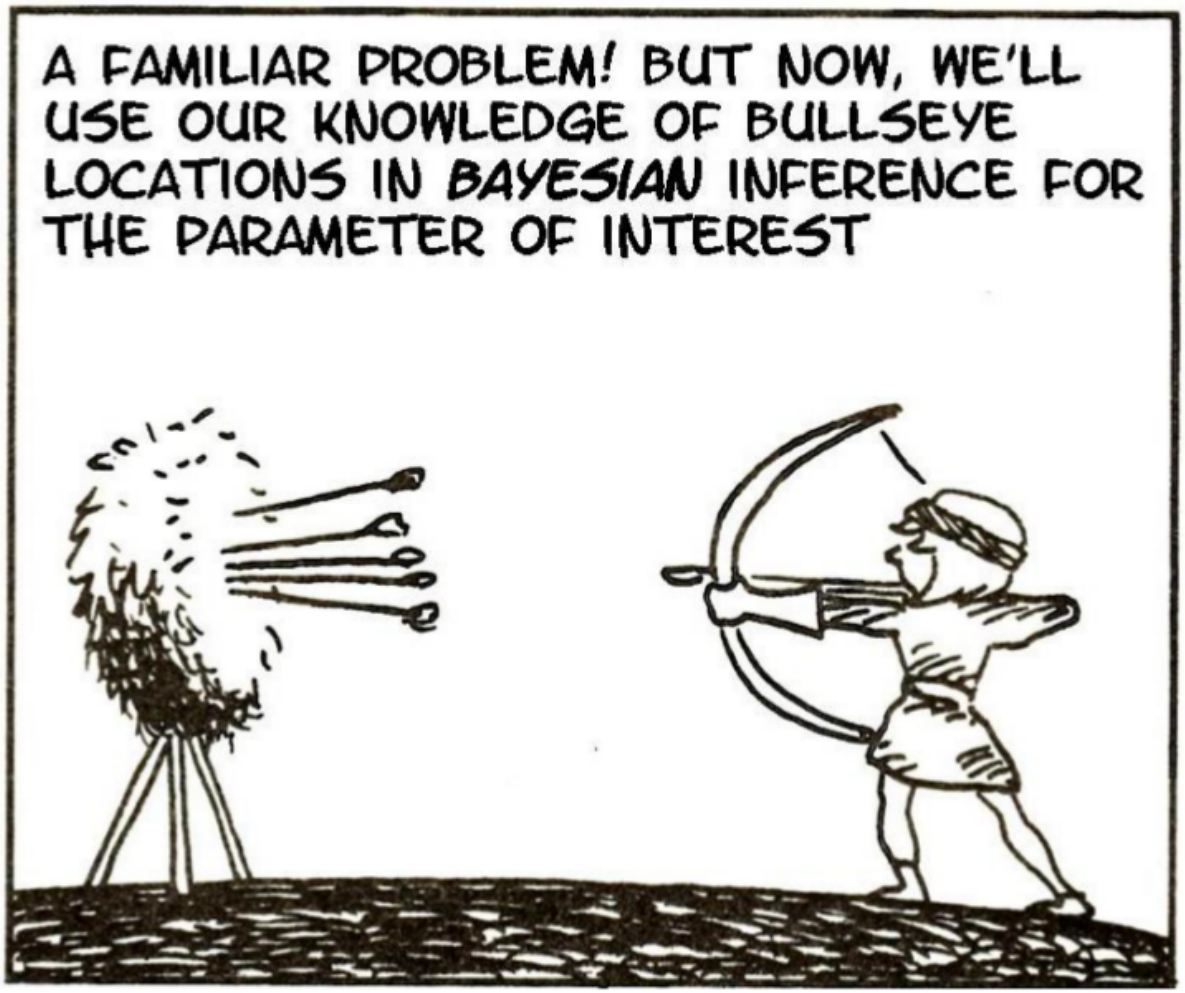
Frequentist Inference



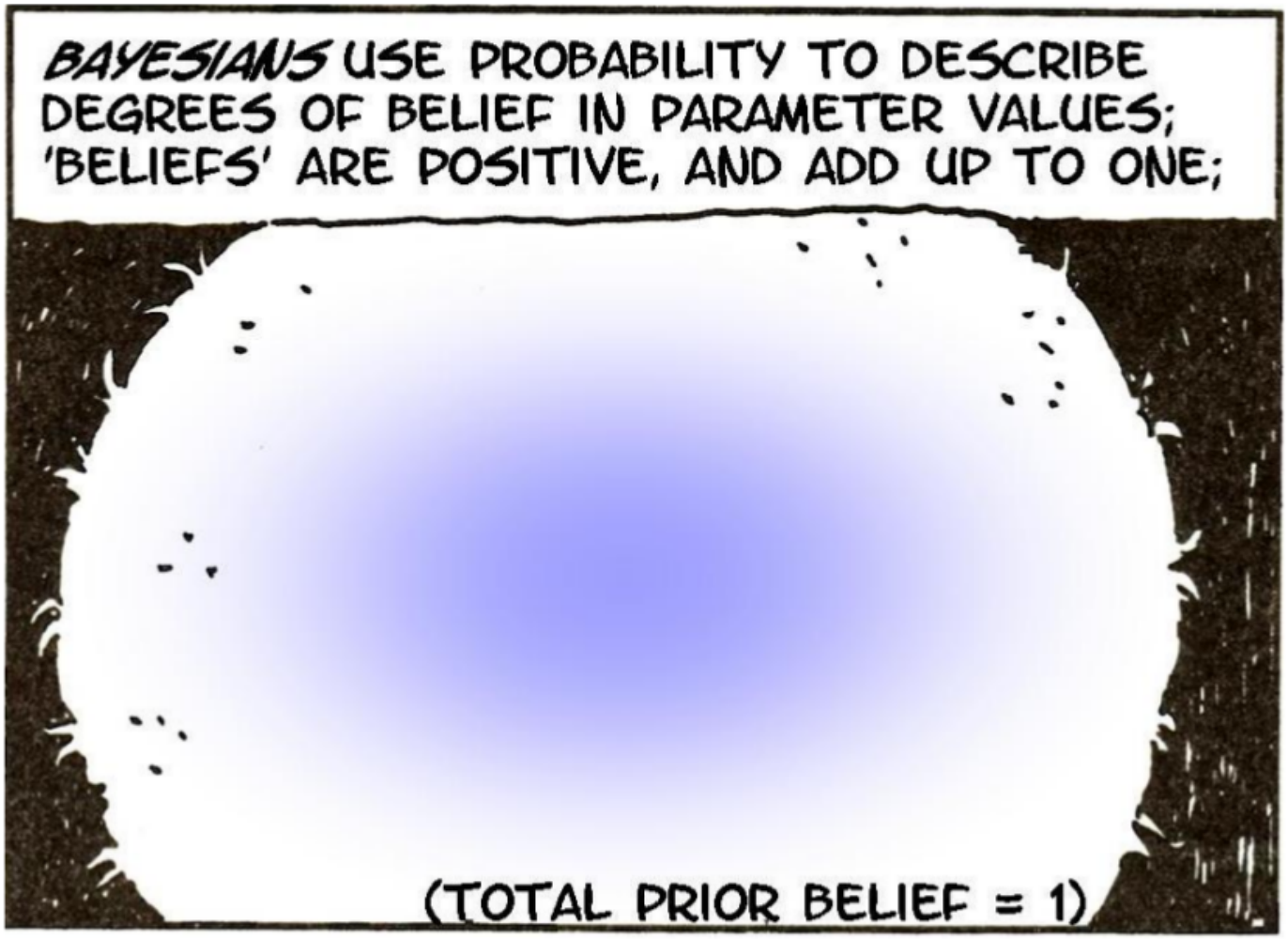
Frequentist Inference

- For testing or estimating, imagine running your experiment again and again. Or, perhaps, make an argument like this (cf. Larry Wasserman, *All of Statistics*):
- 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
- 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?

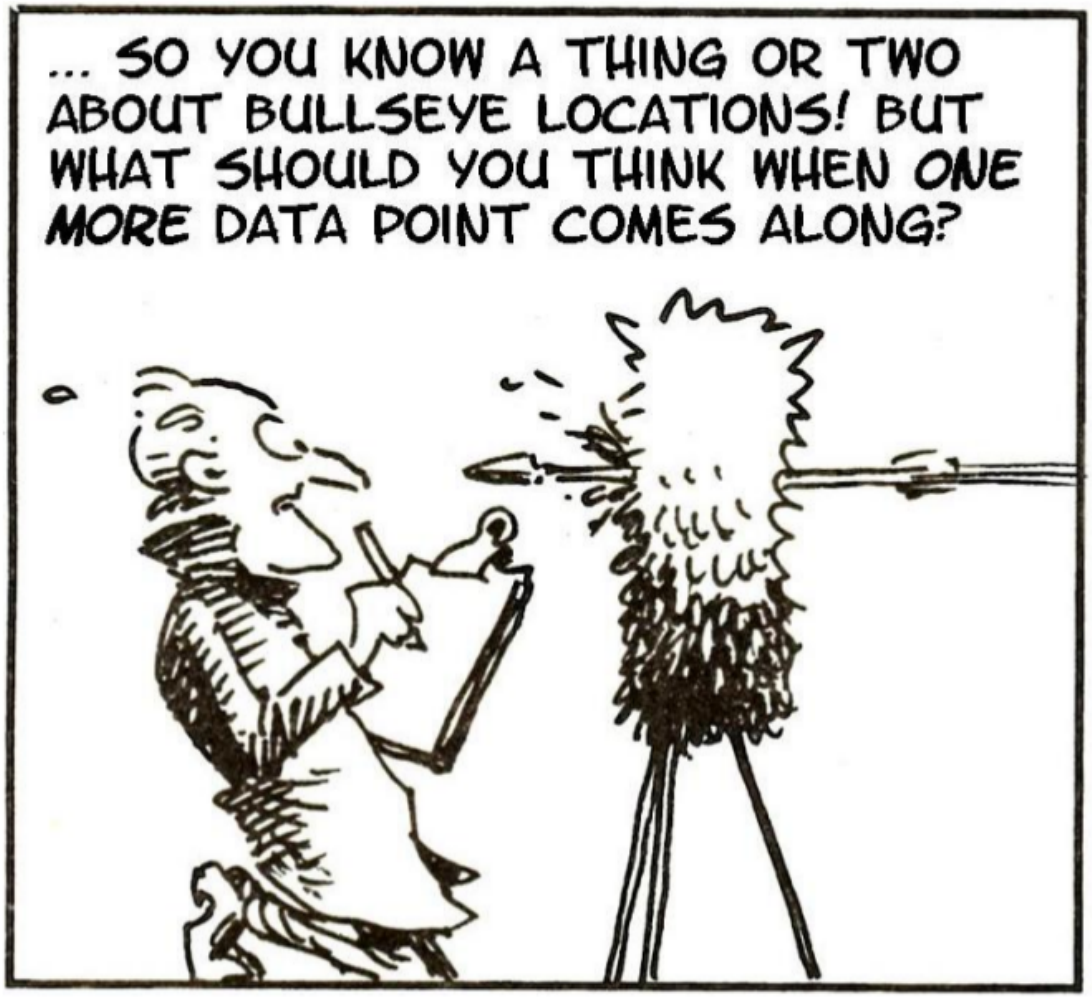
Bayesian Inference



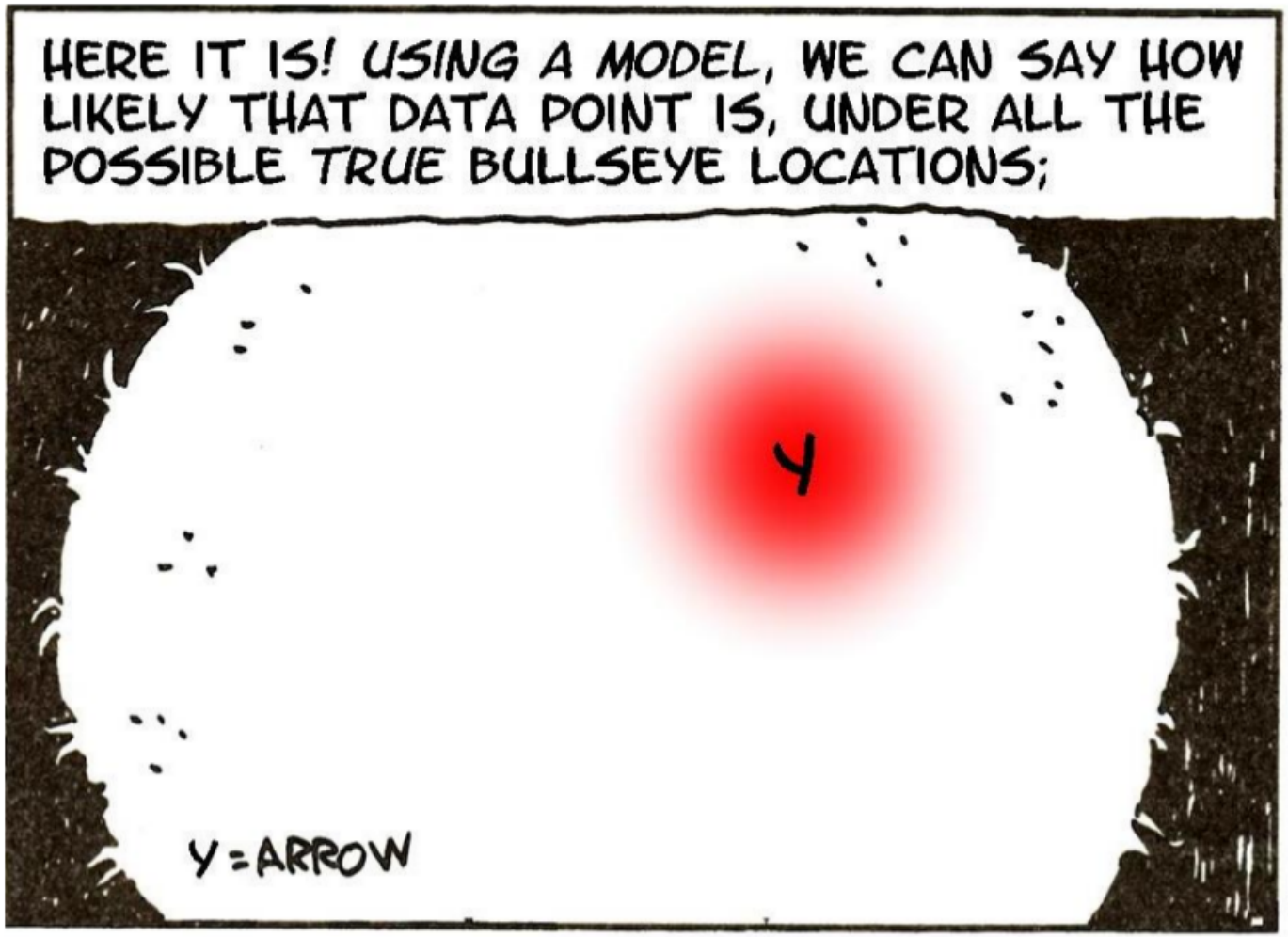
Bayesian Inference



Bayesian Inference



Bayesian Inference



Bayesian Inference

How to update knowledge, as data comes in? We use;

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

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

$$p(\beta|\mathbf{Y}) \propto f(\mathbf{Y}|\beta) \times \pi\beta$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

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

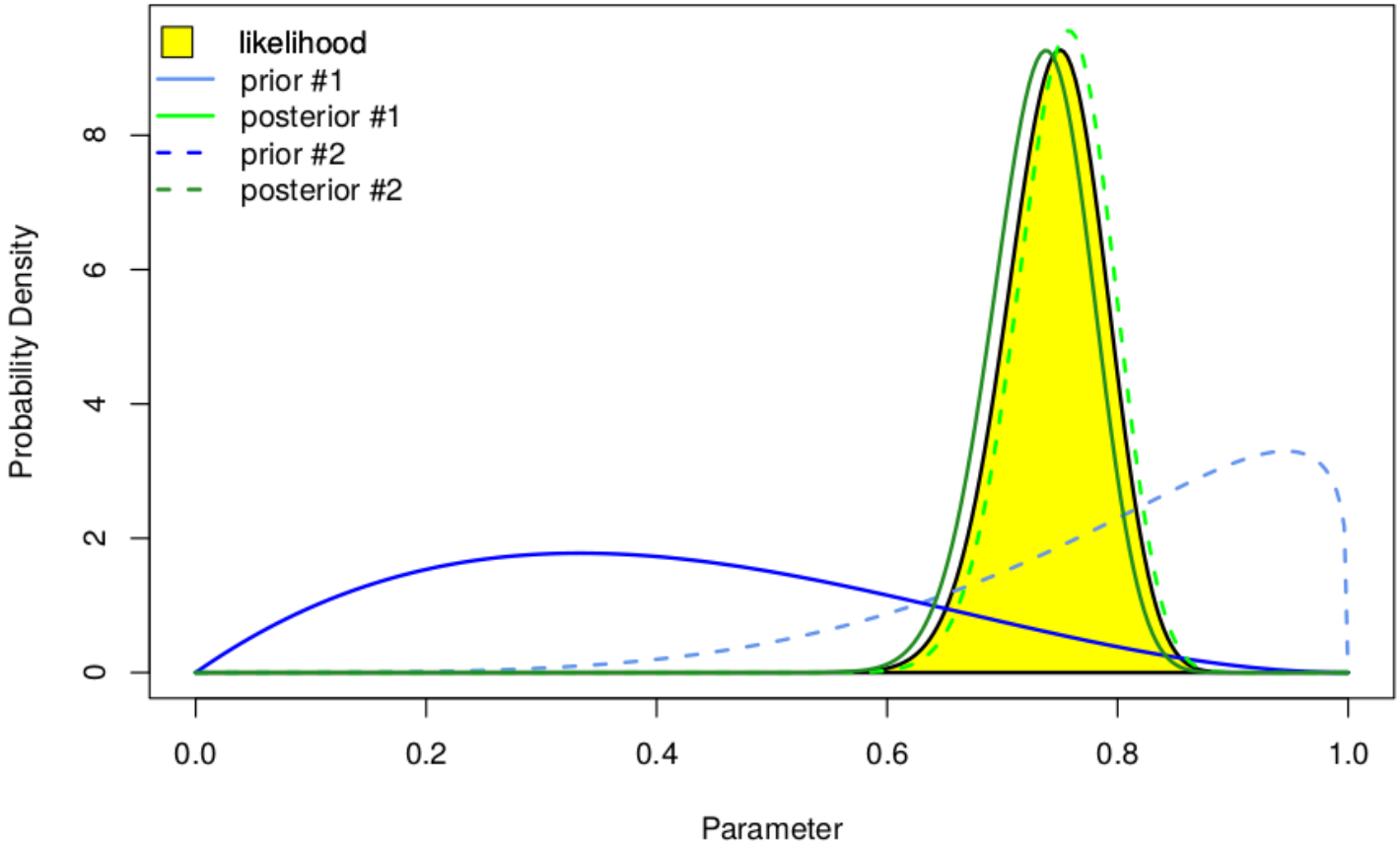
- No replications – but we can still replicate and investigate frequentist properties
- Given modeling assumptions & prior, process is automatic
- Keep adding data, and updating knowledge, as data becomes available... knowledge will concentrate around true β.

Where do priors come from?

- Priors are assumptions
- Priors come from all data external to the current study, i.e. everything else.
- E.g. incorporating what subject-matter experts know/think is known as eliciting a prior.
- Also, knowledge of what is typical in a field.
- In Psychology it is uncommon to obtain very large effect sizes
- Standardizing variables can help figuring out what is reasonable – and what not

When don't priors matter (much)?

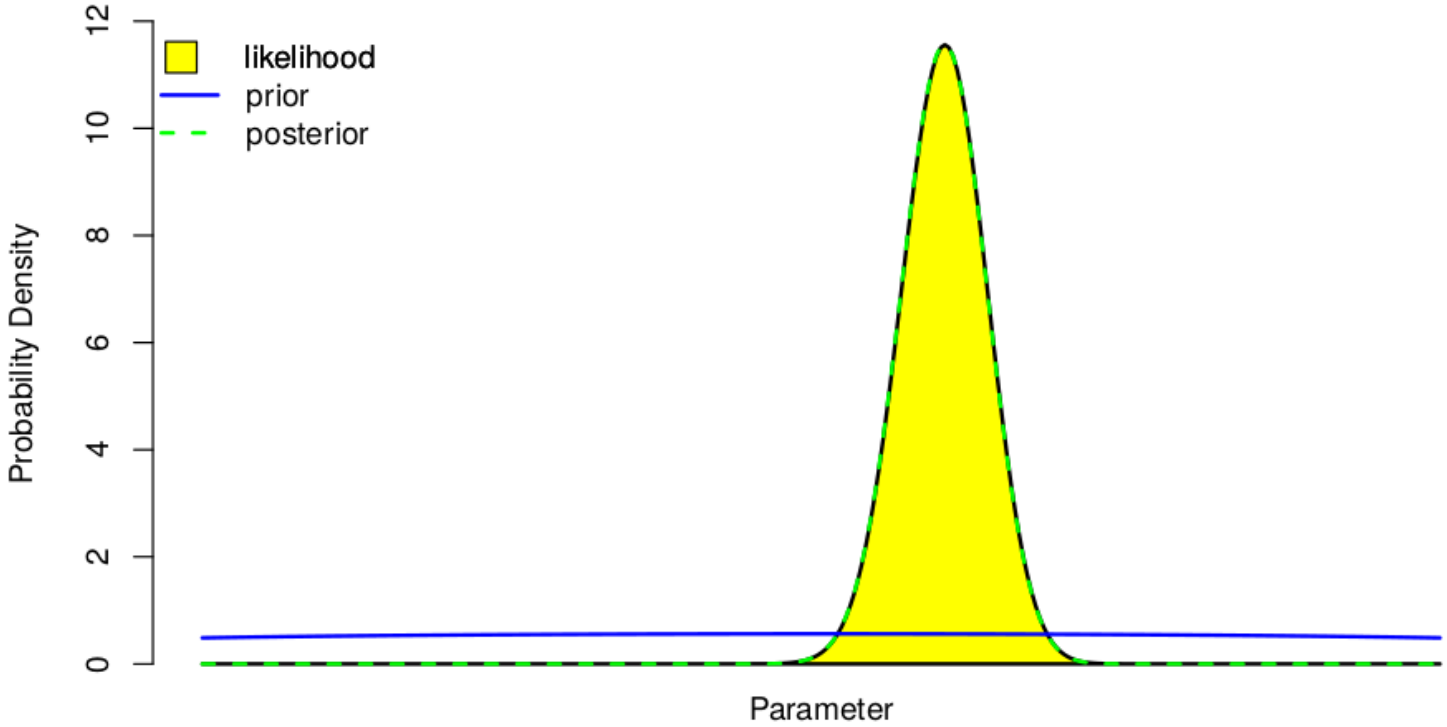
When the data provide a lot more information than the prior, this happens;



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

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 simple 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.

Computation for Bayesian statistics

- Bayesian analysis requires evaluating expectations of functions of random quantities as a basis for inference
- These quantities may have posterior distributions which are multivariate or of complex form or often both.
- For many years Bayesian statistics was essentially restricted to conjugate analysis, where the mathematical form of the prior and likelihood are jointly chosen to ensure that the posterior may be evaluated with ease.
- Revolutionary change occurred in the early 1990s with the adoption of indirect methods, notably Monte Carlo Markov Chain (MCMC).

The Monte Carlo method

- Posterior distribution $\Pr(\theta|y)$ is approximated by taking a very large random sample of realizations of θ from $\Pr(\theta|y)$
- The approximate properties of $\Pr(\theta|y)$ by the respective summaries of the realizations.
- For example, the posterior mean and variance of θ may be approximated by the mean and variance of a large number of realizations from $\Pr(\theta|y)$.
- Similarly, quantiles, modes, etc may be obtained from realizations from $\Pr(\theta|y)$
- Some similarity with bootstrapping and simulations
- Samples from the posterior can be generated in several ways, without exact knowledge of $\Pr(\theta|y)$
  - Rejection and importance sampling

Markov Chain Monte Carlo (MCMC)

- Realizations from the posterior used in Monte Carlo methods need not be independent, or generated directly.
- If the conditional distribution of each parameter is known (conditional on all other parameters), one simple way to generate a possibly-dependent sample of data points is via Gibbs Sampling.
- This algorithm generates one parameter at a time; as it sequentially updates each parameter, the entire parameter space is explored.
- It is appropriate to start from multiple starting points in order to check convergence, and in the long-run, the 'chains' of realizations produced will reflect the posterior of interest.
- More efficient methods are now available
- Hamiltonian Monte Carlo