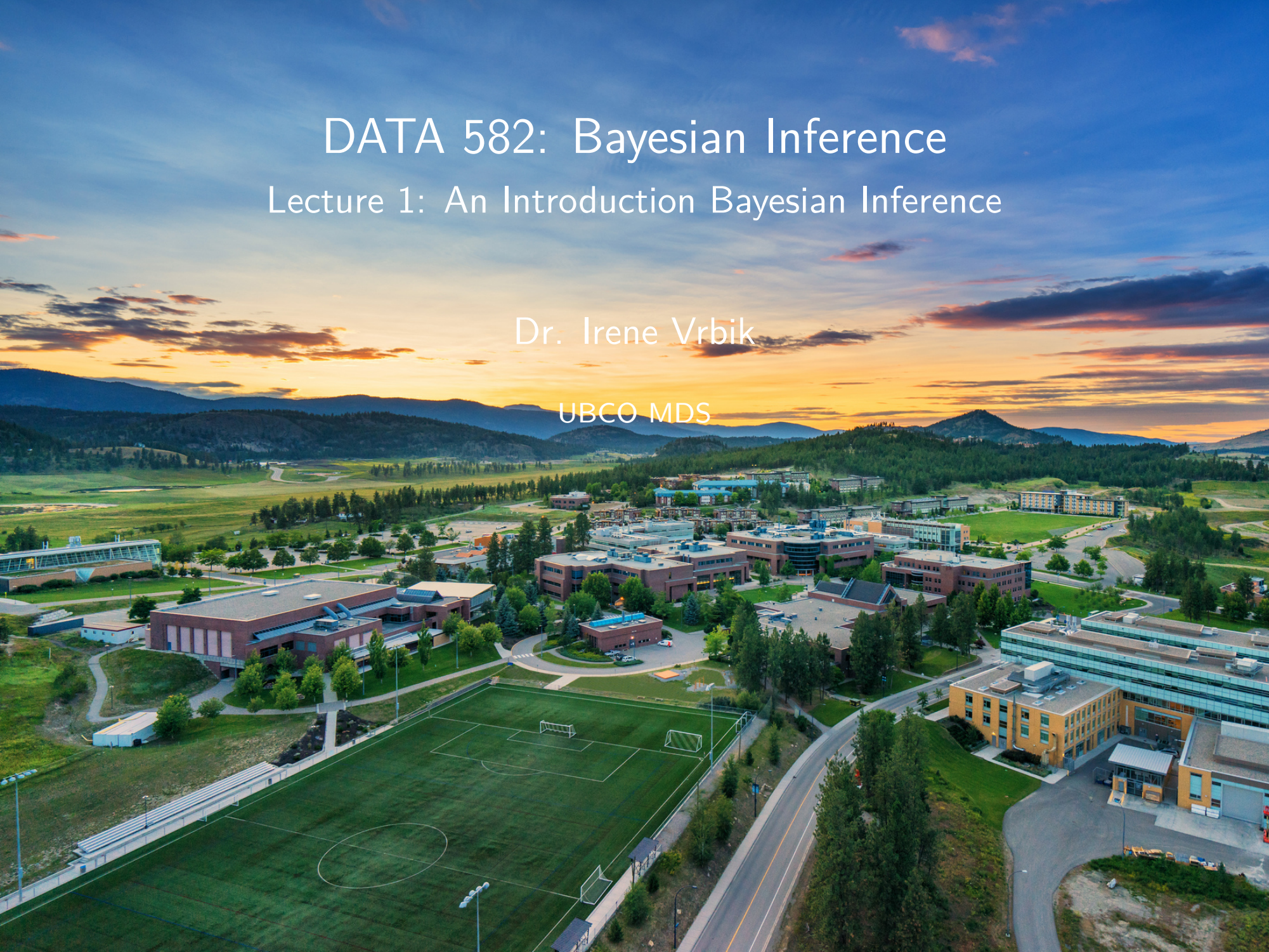
DATA 582: Bayesian Inference

Lecture 1: An Introduction Bayesian Inference

Dr. Irene Vrbik

UBCO MDS



Who am I?

Name:	**e-mail:**
Dr. Irene Vrbik	irene.vrbik@ubc.ca
Academic Career:	Undergrad 2009 (McMaster University), Masters 2010 (University of Guelph), PhD 2014 (University of Guelph), Postdoc 2014 (McGill University), NSERC[1] postdoctoral fellowship 2016 (UBCO), Instructor, 2018–2020 (UBCO) Assistant Professor of teaching, present (UBCO)

I have taught Statistics/Data Science/Probability at the University of Guelph, McGill, and UBCO.

[1] Natural Sciences and Engineering Research Council of Canada


UBC
adidas

Course Description

DATA 582 (1) Bayesian Inference:

Introduction to Bayesian paradigm and tools for Data Science. Topics include Bayes theorem, prior, likelihood and posterior. A detailed analysis of the cases of binomial, normal samples, normal linear regression models. A significant focus will be on computational aspects of Bayesian problems using software packages. Restricted to students in the MDS program.

Prerequisite: DATA 572 (Supervised Learning),
DATA 581 (Modelling and Simulation II).

Course Schedule: Block 6

Lecture: Mon 9:30 AM to 11:00 AM

- **Where:** EME 1153

Lab: Thursday 1:30 – 3:30 pm

- **TA:** Yining Zhou
- **Where:** EME 1153

Office Hours: Thursday 3:30–4:30

- **Where:** SCI 104

System and Tools:

- We will be using R and Rstudio
- Github will be used for accessing various course material lectures/labs/course outline/data sets, etc.
- Canvas will be used for posting grades, accessing Zoom links, and submitting assignments.
- Slack will act as a Discussion board and general communication platform. Please make sure you are signed in to the relevant workspace.

Optional e-books available at UBCO's Online Library

BR Alicia A. Johnson, Miles Q. Ott, Mine Dogucu. *Bayes Rules!* An Introduction to Applied Bayesian Modeling. Chapman and Hall/CRC, 2022. Available online:[2]

AG Gelman, Andrew, et al. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

PH Hoff, Peter D. *A first course in Bayesian statistical methods.* Vol. 580. New York: Springer, 2009.

WB Bolstad, William M., and James M. Curran. *Introduction to Bayesian statistics*, John Wiley & Sons, 2016.

SL Lynch, Scott M. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. Statistics for Social and Behavioral Sciences*, Springer Science & Business Media, 2007.

[2] online version https://www.bayesrulesbook.com/

- Suggested readings—in the form of relevant textbook chapters, interesting articles/videos/blog postings—will be posted on GitHub (see "Optional Reading" heading in the **Lectures** section) or as clickable links (in blue) throughout the lecture slides.

- These readings are intended to be supplementary and therefore not mandatory for you to complete to succeed in this course.

- While this course is intended to be self-contained, it is **not** intended to be a compilation of all things Bayesian. Rather a entry point to the basic tools of Bayesian statistical methods.

Marking and Evaluation

Block 6: March 25, 2024 to April 25, 2024[3]

This module will be 7 lectures, 4 labs, 2 quizzes (held during lecture time), and 2 assignments.

Item	**Weighting**	**Tentative Dates**
Assignments	40%	TBD
Quizes	60%	Quiz 1: Wednesday, April 10 Quiz 2: Wednesday April 24

[3]Note that Monday April 1 is a holiday (Easter) so there will be no class or lab.

Course Syllabus

- The course syllabus is a dynamic document that will live on https://github.com/ubco-mds-2023/Data-582
- Lecture material, assignment solutions, etc. can be found on the github repo and organized on the landing page (syllabus).
- Canvas will be most almost exclusive for assigning grades and submitting assignments.

Learning outcomes

By the end of this module students will be able to:

- Calculate inverse probabilities using Bayes' Theorem
- Calculated a likelihood function
- Specify and describe the role of prior distributions
- Combine prior distributions with likelihood to derive posterior distributions
- Summarize posterior distributions in the context of Bayesian Inference

What is Bayesian Inference?

Suppose I flip a coin and hide the outcome.
What is the probability that it lands heads?

Introduction

More generally, *inference* (*noun*) is the a conclusion reached on the basis of evidence and reasoning.

– Google dictionary

Statistical inference is the process of using data analysis to **infer** properties of an underlying distribution of probability. Inferential **statistical** analysis infers properties of a population, for example by testing hypotheses and deriving estimates.

– Wikipedia

There are two main paradigms within the realm of statistical inference:

Bayesian inference and *Frequentist inference*

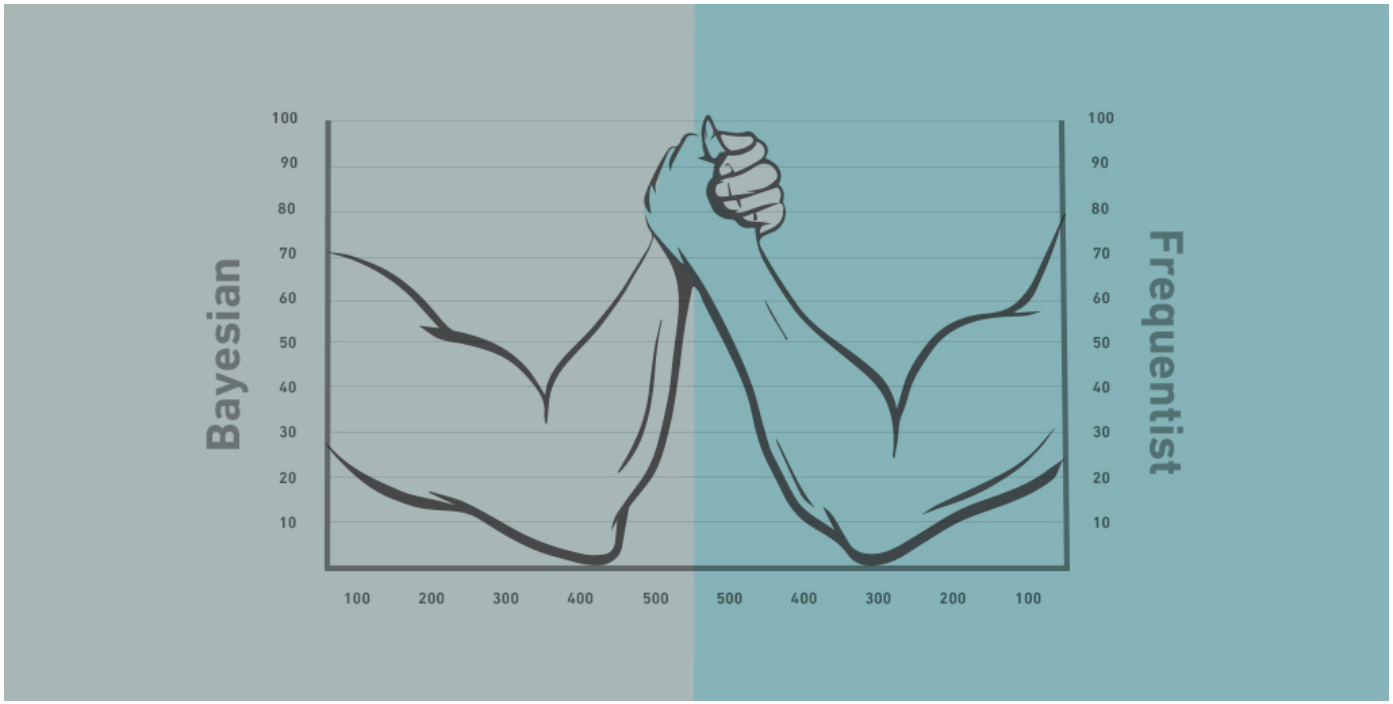


Image source: Pinnacle: *Assessing betting skill: Bayesian vs. Frequentist methods*

Frequentist vs.Bayesian Inference

A philosophical and fundamental difference between Frequentists and Bayesians concerns their treatment of *probability*.

Frequentist view probability as a limiting case of repeated measurements

- If you flip a coin many, many, many times, then roughly 50% will be *heads* (and the other $\sim$ 50% will be *tails*)

Bayesians use probabilities to quantify uncertainty about our beliefs.

- For example, we might talk about the probability of someone being guilty of a crime.

Frequentist vs. Bayesian Inference

Frequentist

- Probability explain in terms of *repeated* measurements.
- Fixed values cannot be assigned probabilities $P(X)$ no $p(x)$

Bayesian

- Probability is *subjective*
- Allows us to assign probabilities to facts, eg. $P(\text{person committed crime})$

Notice that our example involving the guilt/innocence of a person is not a repeatable event.

The frequentist says the person either committed the crime with 100% or 0% probability.

A Bayesian might assign an 80% probability that the person committed a crime (another Bayesian may say, 50%).

Frequentist vs. Bayesian Inference

- *Frequentists*[4], as the name suggests, view probabilities as long run *frequencies* with which events occur.
- Frequentist inference collects data to test a certain *hypothesis* (H_0) against some *competing hypothesis* (H_1).
- Conclusions require that the correct decision be made with a certain (high) probability which is determined by the significance level α (typically set to 0.05).
- These decisions are based on *p*-values
  - *p*-values $< \alpha$ are considered significant
  - i.e. small *p*-values lead us to reject the null hypothesis

[4] pronounced FREquent-tist or freQUENT-tist

Frequentist vs. Bayesian Inference

- Recall: a p-value calculates the probability of observing data at least as extreme as the data we collected, **assuming the null model is correct**., eg. assuming the hypothesis $H_0 : \theta = 5$ is true.

- Any given experiment can be considered as one of an infinite sequence of possible repetitions of the same experiment.

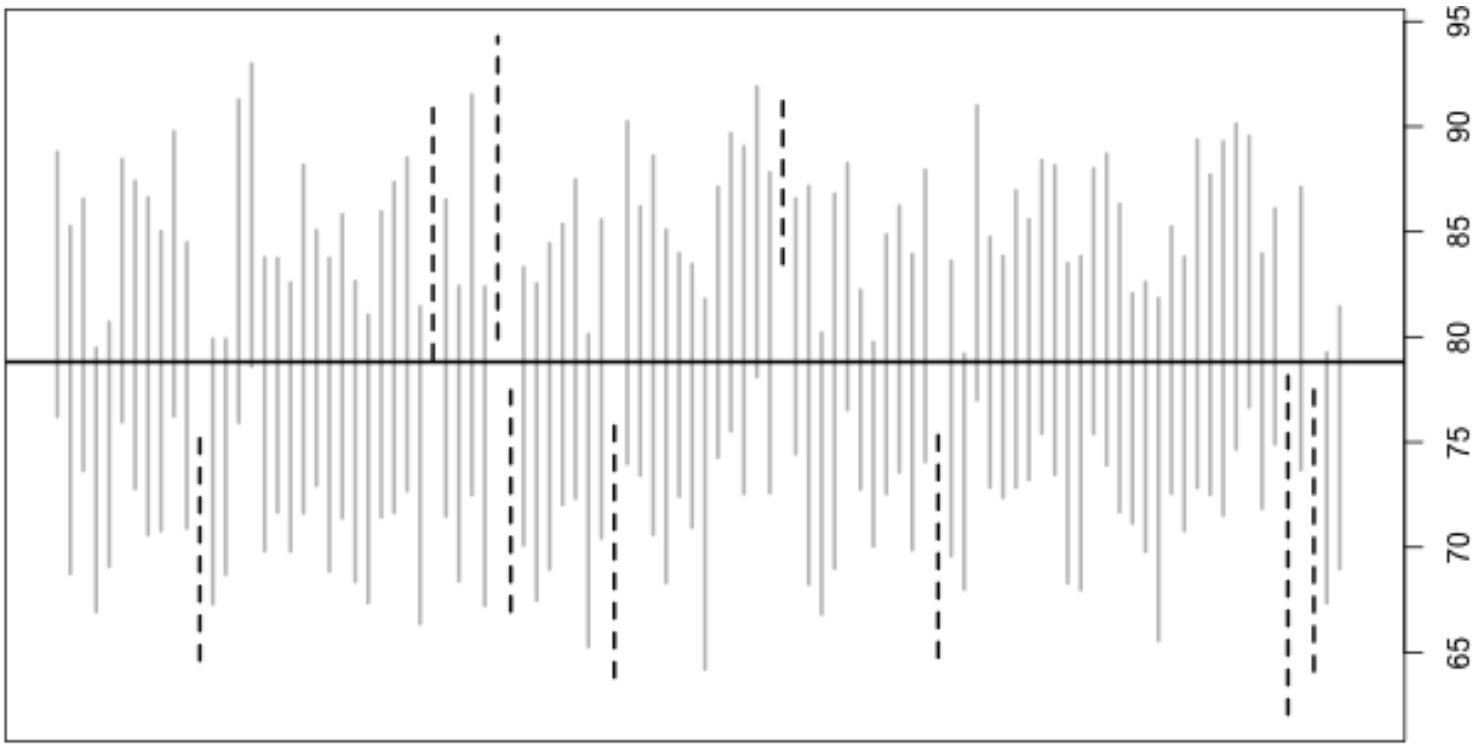
Frequentist Aim

Hence our model is our "truth" and our data is random.

$$P(data \mid model)$$

Frequentist vs. Bayesian Inference

- A frequentist would interpret a 95% confidence interval for unknown parameter θ as a range of values that would contain the true value of θ ninety-five percent of the time.
- Hence θ is a **fixed number** and the data is random.



Frequentist vs. Bayesian Inference

Some criticisms of the frequentist approach:

1. The frequentist approach relies on repeated experiments, which in practice we rarely do.
2. Inferences depend heavily on the sample size n
   - inaccurate estimates with small n
   - by increasing n any hypothesis can eventually be rejected
3. Two studies reporting the same p-value do not necessarily convey the same evidence.
4. Using an α-level of 0.05 means that 1 in 20 "significant results" would expected by random chance alone.

Frequentist vs. Bayesian Inference

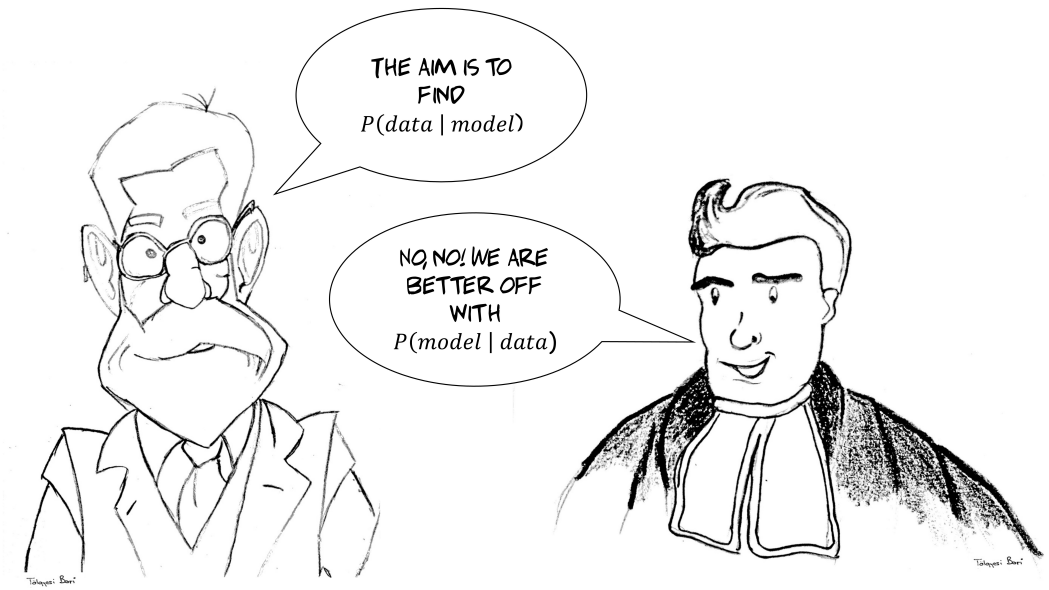
- In the context of Bayesian inference this subjectivity comes in the form of a *prior distribution*.
- In words, this involves assigning a probability distribution to an unknown population parameter, θ.
- It is useful, therefore, to treat parameters as *random variables* rather than unknown constants.
- In contrast to the Frequentist framework, the data is fixed and our model is what we are uncertain of.

Bayesian Aim

Bayesians consider the data to be our "truth" and we calculate the probability of the hypothesis (i.e. the model is uncertain).

$$P(model \mid data)$$

Image sourced from blog post: *Bayesianism vs Frequentism*, Agoston Torok

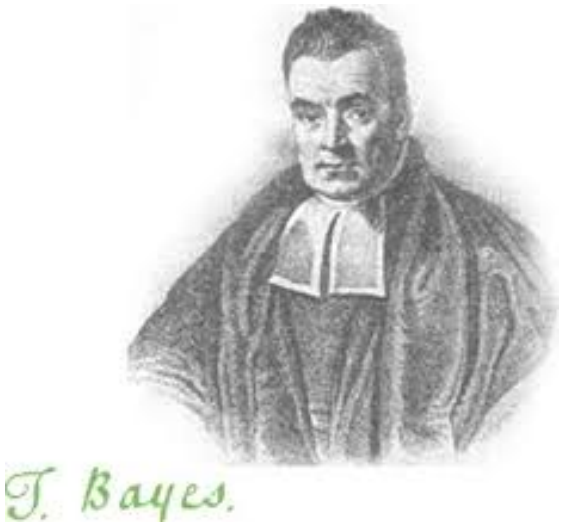


Who is Bayes?

A brief history

The Bayesian-Frequentist history

- Bayes' theorem is named after the English statistician, philosopher and Presbyterian minister Reverand Thomas Bayes (him $\rightarrow$).



- The name *Bayesian* comes from the fact that it relies on Bayes theorem.
- Bayesian statistics can be traced back to the late 1700 [source].
- Bayesian inference used to dominate statistical practices, but was underdeveloped for centuries in large part because Bayesian problems were computationally difficult.

The Bayesian-Frequentist history

- A famous adversary to Bayesian statistics in Sir Ronald Fisher. (him $\rightarrow$).
- In Fisher's 1925 handbook he said that Bayesian statistics (then called inverse probability) is "is founded upon an error, and must be wholly rejected".
- In the 20th century the favour flip-flopped over to Frequentist; however, with the advent of faster computers came the revolution in the use of Bayesian methods.

The Royal Society— a Fellowship of many of the world's most eminent scientists—discuss the contributions of Sir R.A Fisher and Bayes in this short YouTube video.

Why Bayesian Inference?

Pros to Bayesian

Pros:

- Bayesian Inference can be extremely useful where data is sparse (small sample size n)
- The resulting models are easy to interpret, eg. θ has a 95% probability of falling within the credible interval.
- Allows us to incorporate domain knowledge through the *prior distribution*
- A nice alternative when Frequentist model assumptions aren't met/verified.

Pros to Bayesian

Cons:

- The prior distribution is subjective and can heavily influence the end result
- Fitting a good model can be hard
- It involves hard to compute integral (MCMC will help with this)
- Models are often computationally expensive.
- Intimidating to get started

Comment

- Both Bayesian and Frequentist methods have their share of pros and cons.
- As Data Scientist, you should appreciate both "camps" as powerful tools to process data.
- There is no one-size-fits all model and some projects will benefit from one other the other.[5]
- As Bayesian methods are becoming more an more accessible (both with the advent of faster computers and the development of user-freindly(ish) packages/libraries) we can see Bayesian methods on the rise.

[5] My Master's Thesis was in the Bayesian Framework while my PhD thesis was in the Frequestist framework

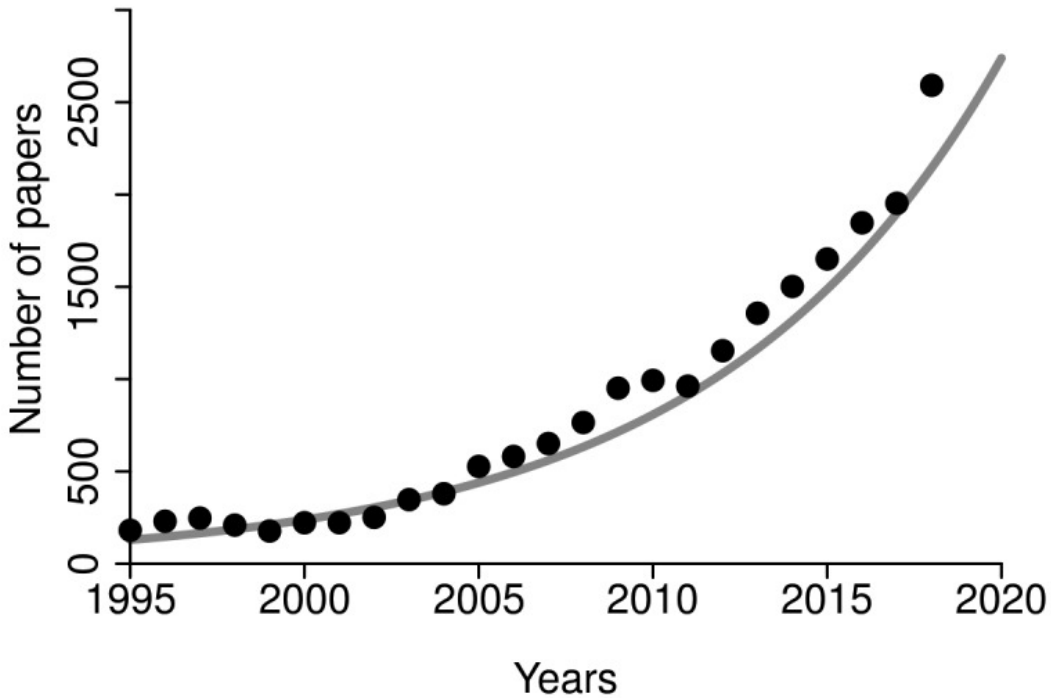


Figure: The number of published medical articles using Bayesian statistics in the period from 1995 to 2018 (sciencedirect.com, February 2019). [Source]

Terminology

Probability is a way to quantify the uncertainty surrounding the outcome of an experiment.

An *experiment* is a process by which we observe something random.

- eg. coin toss or the roll of a die.

An *outcome* is the result of an experiment.

- eg. experiment = rolling a die,
- outcome= number of dots facing up when it lands.

Terminology

We call the collection of all possible outcomes of an experiment the *sample space*, often denoted by S or Ω.

- eg. roll of a die, this is $S = \{1, 2, 3, 4, 5, 6\}$.
- S is comprised of subsets called events.

An *event* is any subset of the sample space, often denoted by the first letters of the alphabet

- For the roll of a die, the set odd numbers $E = \{1, 3, 5\}$.

We denote the probability of event E occurring by $P(E)$.

Probability Terminology

Events are said to be *disjoint* or *mutually exclusive* if they cannot happen at the same time.

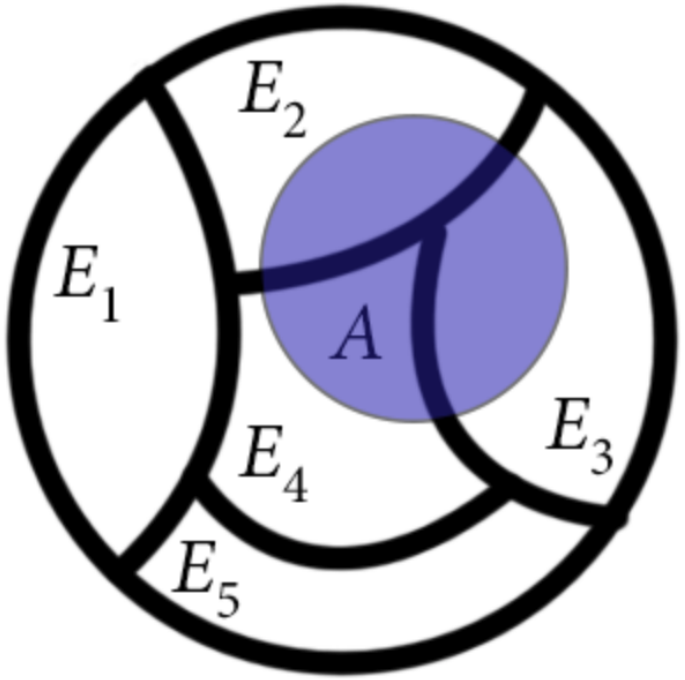
- eg. a coin cannot land both heads and tails

An event can be broken up into a collection of mutually exclusive and exhaustive subsets which make up a *partition*.

The subsets are exhaustive if every outcome in event is part of *exactly one* of these subsets.

- eg. Canada is partitioned into the provinces and territories.
- A visualization of this can be seen on the following slide . . .

You should think of a partition as a way to "cut up" a set into pieces. This colorful diagram is an example of a partition of a disc.



The area taken up by the set A is the same as the area taken up by the pieces of A which overlap the E's. That is, the E's give us a partition of A.

Figure: Image sourced from Math ∩ Programming - Jeremy Kun.

Axioms of Probability

For an experiment with possible outcomes $E_1, E_2, \ldots, E_n$, the probability $P(E_i)$ must obey the following rules:

Probabilities must be non-negative real numbers, $P(E_i) \geq 0 \ \forall i$

The probability of the entire sample space must be equal to one, i.e. $P(S) = 1$

If two events are disjoint the probability that either happens is the sum of their individual probabilities.

Probability Terminology

- Often events are combinations of two or more events formed by taking unions ($\cup$), intersections ($\cap$), and complements (E^C / $\overline{E}$).
- I find it useful to the think of the above operations as follows:
  - interchange union for "or",
  - interchange intersection for "and" and
  - interchange compliment for "not"

Die example

Consider rolling a die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Let A be the event of rolling an even number:

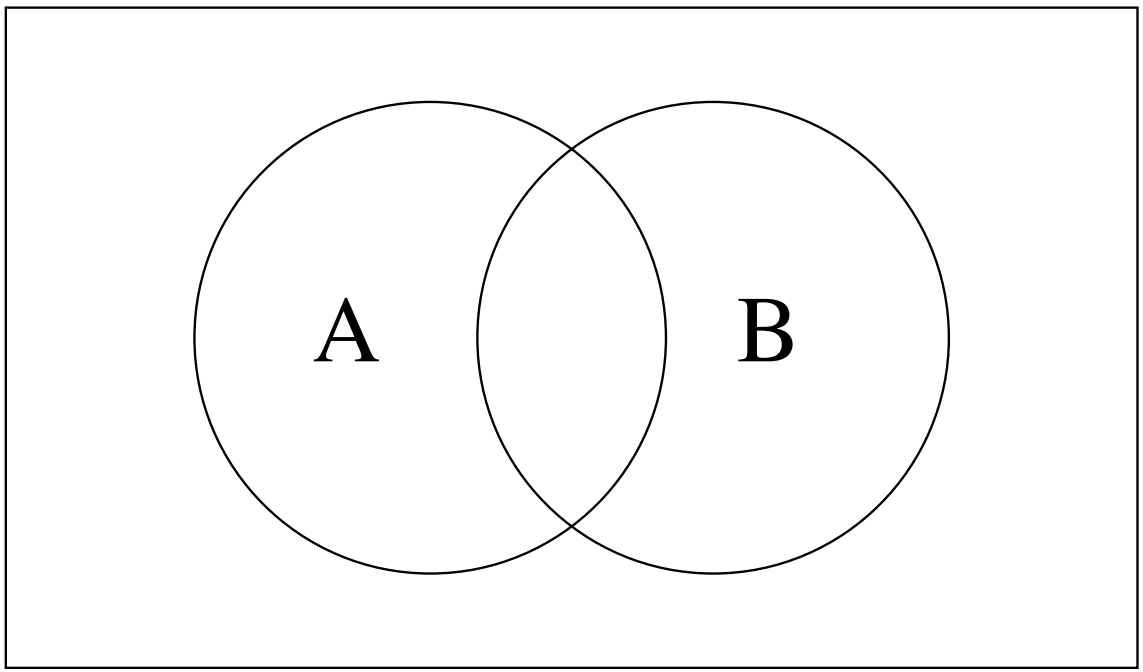
$$A = \{2, 4, 6\}$$

Let B be the event of rolling a value larger than 3:

$$B = \{4, 5, 6\}$$

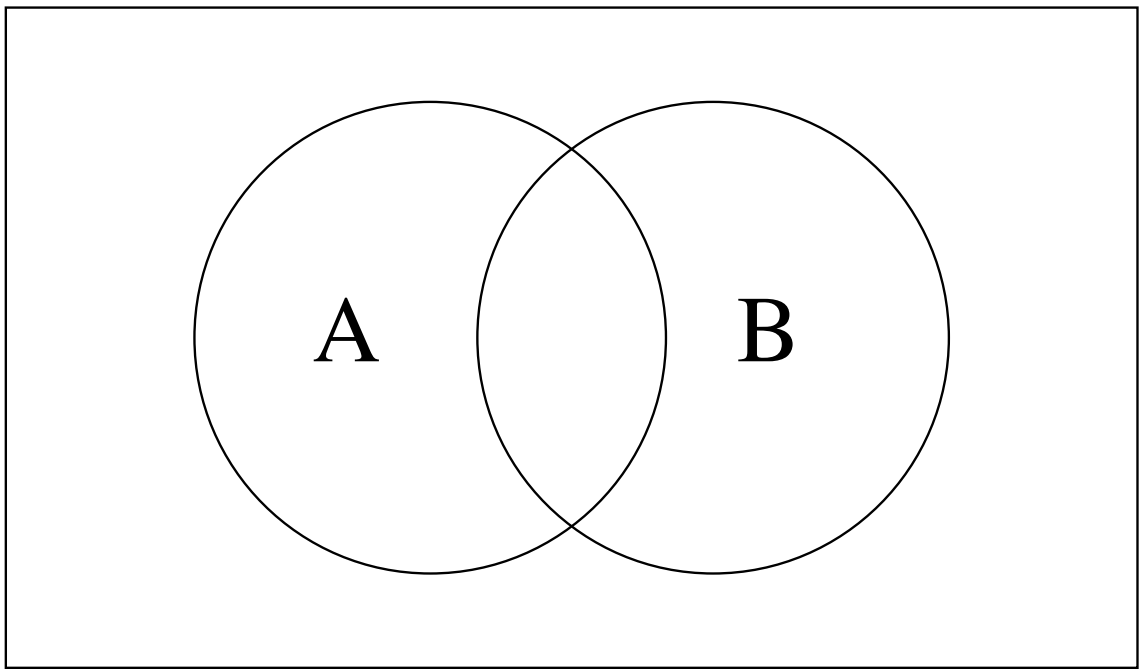
$S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$, $B = \{4, 5, 6\}$

The union of A and B, denoted $A \cup B$ and read "A or B", is the set of all elements that appear in A and/or B.



$S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$, $B = \{4, 5, 6\}$

The intersection of A and B denoted by $A \cap B$ and read "A and B', is the set of all elements that A and B have in common



Theorem ("OR"/addition Rule)

*For any two events E and F, the probability of E **or** F occurring is given by*

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{1}$$

Theorem (Partition Theorem)

Let $E_1, \ldots, E_n$ be a partition of S, and let A be an arbitrary event. Then

$$P(A) = \sum_{i=1}^{n} \mathrm{P}(E_i \cap A) = \sum_{i=1}^{n} \mathrm{P}(A \mid E_i)\mathrm{P}(E_i) \tag{2}$$

Probability Terminology

What is the probability of rolling an even number (define event $A = \{2, 4, 6\}$) or a number larger than 4 (define event $B = \{5, 6\}$)?

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 3/6 + 2/6 - 1/6 = 4/6 = 2/3 \end{aligned}$$

Conditional probability, aims at finding the probability of an event, given that a certain event has already occurred.

Definition

Provided $P(B) > 0$, the conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Rearranging the rule for conditional probability, we get:

Definition ("AND"/multiplication Rule)

If E and F are two events then

$$P(E \cap F) = P(E)P(F \mid E) = P(F)P(E \mid F)$$

Conditional Probability

Red Ace example

What is the probability that a randomly drawn card from a standard deck of cards is an ace, given that we know it is a heart.

$$P(A \mid \heartsuit) = \frac{P(A \cap \heartsuit)}{P(\heartsuit)} = \frac{1/52}{13/52} = \frac{1}{13}$$

Logically, there are 13 hearts in a deck, and only 1 of them is an ace. Hence $P(A \mid \heartsuit) = 1/13$

Bayes Theorem

Putting it all together we get *Bayes Theorem*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{def of cond prob}$$

$$= \frac{P(A)P(B \mid A)}{P(B)} \qquad \text{“AND” rule}$$

Definition (Bayes Theorem/Formula)

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{3}$$

This theorem forms the foundation of Bayesian inference.

Screening Test example

Suppose a screening test for a disease has a 2% false positive rate and a 99% true positive rate. Suppose also that the prevalence of the disease in the population is 0.5%. Finally suppose a randomly selected person tests positive. What is the probability that they have the disease?

See reading 11 (ex. 4) [6]

[6] Jeremy Orloff, and Jonathan Bloom. *18.05 Introduction to Probability and Statistics*. Spring 2014. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons .

Screening Test Example

- When a patient goes through screening there are two competing claims:
  - patient has the disease and
  - patient doesn't have the disease.

Question: *If a test yields a positive result, what is the probability that patient actually has the disease?*

- To help answer this question it will help to visualize this information in a *Tree Diagram*[7]

[7] a tool to organize outcomes and probabilities around the structure of the data.

Bayesian Updating

- Suppose we send the person to get retested.
- Our prior belief before seeing the results of the second test should be different from our beliefs of another individual taking the test for the first time.
- In addition, if the second test came back positive, surely that would heavier weight than the first positive test.
- To not loose the information from the first test, a Bayesian would use the posterior obtained from the first test and use that as the prior belief in the second experiment.
- In essence, this is the process *Bayesian updating*.

Data: the results of the experiment, i.e. positive test (event $+$).

Hypothesis: The person has the disease (event D)

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+)} \tag{4}$$

$P(D)$ our *prior* belief before looking at the evidence.
$P(+ \mid D)$ *likelihood* of evidence given they have disease.
$P(+)$ chance of seeing what we saw
$P(D \mid +)$ *posterior* belief upon considering the evidence.

Inverting Probability

- Notice that in this example we are given

$$P(+ \mid \text{disease})$$

but we are asked for

$$P(\text{disease} \mid +)$$

- Recall the competing paradigms:

$$\begin{array}{c} \text{(Bayesian)} \quad \text{vs (Frequentist)} \\ P(\mathit{model} \mid \mathit{data}) \text{ vs } P(\mathit{data} \mid \mathit{model}) \end{array}$$

- In the above example we used Bayes theorem in the context of single events and point probabilities.
- To enter into the world of Bayesian inference, we must extend this theorem to *distributions* of θ (which we treat as a random variable). Notice how probability rules above can all be extended:

Events & point probabilities	**Distributions**
$P(A \cap B) = P(B \mid A)P(A)$	joint pdf $p(x, y) = p(y \mid x)p(x)$
$P(B) = P(B \mid A)P(A)$ $+P(B \mid \overline{A})P(\overline{A})$	marginal pdf $p(y) = \int p(y \mid x)p(x)$ marginal pmf $p(y) = \sum p(y \mid x)p(x)$
$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$	conditional pdf $p(x \mid y) = \dfrac{p(x, y)}{p(y)}$

Probability distributions

- Recall that a *probability distribution* is simply a function that gives the probabilities of occurrence of different possible outcomes for an experiment.
- For a discrete random variable X, these are referred to as probability mass functions (pmf)
  - eg. Bernoulli, Binomial, Poison distribution ...
- For a continuous random variable X, these are referred to as probability density functions (pdf)
  - normal (Gaussian), beta, t distribution, ...
- A review of some popular distributions are provided in the `UsefulDistributions.pdf`. For a more see SL Ch. 2.3.

Probability mass function (pmf)

Definition (probability mass function (pmf))

If X is a discrete random variable, the function given by $p(x) = P(X = x)$ for each x within the range (or support) of X is called the probability distribution or probability mass function of X.

- The pmf may take the form of a table, a function, or a plot.
- The key components of all of these pmf forms is that they have:
  1. The possible values that random variable X can take on (the support/range)
  2. The probabilities, $p(x)$, for each value in the support.
  3. The $\sum p(x) = 1$

Probability density function (pdf)

Definition (probability density function (pdf))

A function with values $p(x)$, defined over the set of all real numbers, is called a **probability density function (pdf)** of continuous random variable X if and only if

$$P(a \leq X \leq b) = \int_a^b p(x)dx.$$

for any real constants a and b with $a \leq b$.

- A function can serve as a probability density of a continuous random variable X if its values, $p(x)$, satisfy the conditions
  1. $p(x) \geq 0$, for $-\infty < x < \infty$.
  2. $\int_{-\infty}^{\infty} p(x)dx = 1$.

Treating θ as a random variable we restate Bayes Theorem as follows:

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

$p(\theta)$ is our *prior*	our belief before looking at the data
$p(y \mid \theta)$ is our *likelihood*[8]	chance observing our data given our hypothesis is true
$p(y)$ is the *evidence*	total plausibility of seeing our data
$p(\theta \mid y)$ is our *posterior*	our updated beliefs after we have considered all the evidence

[8] referred to as a "*sampling model*" in the PH textbook.

This allows us to summarize our prior belief on a model parameter using a distribution rather than point probabilities.

More commonly, we will express the above as follows:

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

$$\implies p(\theta \mid y) \propto p(\theta)p(y \mid \theta)$$

$$\textit{posterior} \propto \textit{prior} \times \textit{likelihood}$$

Notes

- As we will see in an upcoming lecture, $p(y)$ is of little importance to us, making the prior, likelihood and posterior our main functions of interest.

- While we will discuss likelihoods in more detail next lecture, it is worth pointing out that likelihood is not a probability distribution.

- The interplay between the prior, likelihood, and posterior is one which we will study in detail over the next few lectures.

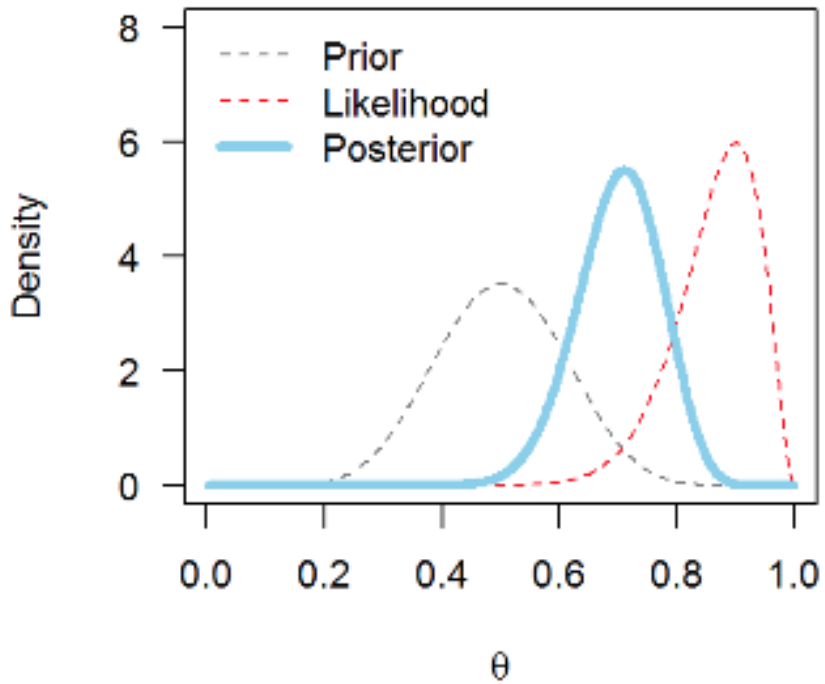


Figure: Image sourced from: *(Pesky?) Priors*, a blog post written by Jim Grange

The Bayesian-Frequentist debate

- It is still widely debated as which method is superior and I don't expect to answer that in this module.
- Both have its advantages and disadvantages and generally the strengths (resp. weaknesses) in Frequentist paradigm can be viewed a weaknesses (resp. strengths) in the Bayesian paradigm.
- This module aims to provide you with the knowledge to perform and understand this alternative school of thought on the how inference should proceed when the data are subject to random variation.