Bayes Time

Week 01

T

2 October 2019

Outline

Goals

1. Understand the motivation behind inference.
2. Appreciate the similarities and differences between Frequentist and Bayesian approaches to inference.
3. Know how to manipulate probability distributions.

Who are we?

We are scientists who, borrowing from Karl Popper, put forward:

> *Statements, or systems of statements, and test them step by step. In the field of empirical sciences, we construct hypotheses and test them against experience by observation and experiment.*

And what do we as scientists spend our time doing? As the famous physicist Richard Feynman declares

> *Experimenters search most diligently, and with the greatest effort, in exactly those places where it seems most likely that we can prove our theories wrong. In other words we are trying to prove ourselves wrong as quickly as possible, because only in that way can we find progress.*

Questions

Let's start with some questions:

1. How much will a student earn after receiving an MBA?
2. Will the Democrats win the next US Presidential election?
3. Did the casino give me weighted dice to play craps?
4. How much do school-provided breakfasts increase student achievement?

How do we answer these questions?

1. We develop ideas about how the world works,
2. gather data,
3. construct a model,
4. test the model,
5. then see what we can say about our original query.

But this is hard…

It is difficult to test those theories because we have uncertainty about how the world works (epistemological) and there are many factors outside of our control that introduce noise (ontological). Some say it is like,

> *listening to a classical orchestra which is playing on the side of a busy street, while we fly overhead in a plane.*

Luckily we have statistics

> *Statistical inference is the logical framework which we can use to test our beliefs about the noisy world against data.*

Put slightly differently:

> *Statistical inference is concerned with drawing conclusions, from numerical data, about quantities that are not observed. Bayesian Data Analysis, 3rd edition*

What is Bayesian inference?

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations. Bayesian Data Analysis, 3rd edition

Bayesian Inference Engine

- You can answer almost any question using the Bayesian inference engine that we outline below.
- If you follow these steps, you will produce a set of empirically tested insights from a transparent and repeatable model that you and others can build upon.
- Talking to colleagues and seeking out diverse perspectives is central to your Bayesian inference efforts, and each step below presents an opportunity to consult with others.

Bayesian Inference Engine

1. Develop the question
2. Gather data, research current literature
3. Construct model
4. Test your code
5. Perform prior predictive checks
6. Estimate the model
7. Carry out posterior predictive checks
8. Write up your results

Way shorter way of putting it

We build a statistical model out of available parts and drive it as far as it can take us, and then a little farther. When the model breaks down, we dissect it and figure out what went wrong.

Why is this process important?

Take it from Gelman and Shalizi in their published article *Philosophy and the practice of Bayesian statistics*

> *Likelihood and Bayesian inference are powerful, and with great power comes great responsibility. Complex models can and should be checked and falsified. This is how we can learn from our mistakes.*

What do we mean by falsify?

We know our model is wrong. It is a simplification of the complex, unwieldy world we live in. So, what do we really mean by falsify?

> *We are not interested in falsifying our model for its own sake—among other things, having built it ourselves—we know all the shortcuts taken in doing so, and can already be morally certain it is false.*

If not falsifying, then what is the point of model checking?

> *The goal of model checking, then, is not to demonstrate the foregone conclusion of falsity as such, but rather to learn how, in particular, this model fails.*

Also, we would like to be able to have some faith in our results:

> *If the data resembles the model closely enough, in the respects which matter to us, then reasoning based on the model will be reliable.*

But, don't let the perfect be the enemy of the good!

Nothing in our set of rules implies that we must run the perfect experiment (if such a thing existed) or collect all the relevant data before we can make valid social scientific inferences. An important topic is worth studying even if very little information is available.

The result of applying any research design in this situation will be relatively uncertain conclusions, but so long as we honestly report our uncertainty, this kind of study can be very useful.

And, you know, things change

> *The social world changes rapidly. To understand those changes requires that we describe them and seek to understand them contemporaneously, even when uncertainty about our conclusions is high. The urgency of a problem may be so great that data gathered by the most useful scientific methods might be obsolete before it can be accumulated.*
>
> *If a distraught person is running at us swinging an ax, administering a five-page questionnaire on psychopathy may not be the best strategy.*

Frequentist World

- Frequentists suppose that the probability of flipping a coin is fixed by God.
- We observe noisy data that is the result of one of an infinite number of repeated experiments.
- Events in this infinite series of experiments occur with probabilities which represent the long-run frequencies of those events occurring in an infinite series of experimental repetitions.

Example: We flip a coin 10 times, and it lands heads 7 times. To a frequentist this is because the universe dealt us an odd sample. See "Rosencrantz and Guildenstern Are Dead" by Stoppard.

Frequentists summed up

Frequentist view the parameters of the probability model as being fixed and the known parts of the system—the data—as varying.

Bayes' World

- Probabilities here are abstractions which we use to express our uncertainty.
- In a Bayesian world fictitious repetitions do not exist.
- In this world we witness the data and never perfectly know the value of an unknown parameter.

Philosophical aside

- Is the parameter truly fixed, but our beliefs are uncertain?
- Or is there no definitive true probability of an event occurring, because the world is ever changing, and the system incessantly evolves?

Compare and Contrast

- Frequentists assume data is random and originates from a series of infinite repetitions. Bayesians assume that we are witnesses to the data, which is fixed.
- Frequentists view parameters as fixed features of the universe that represent a long-run average. Bayesian conceive of probabilities as an expression of subjective beliefs that can be updated in light of new data.

Which leads to diverging paths of inference

- Frequentists use inverse probability as evidence for a given hypothesis and compute the probability of obtaining a sample as extreme as or more than the one actually obtained assuming a hypothesis is true.
- Bayes' formula circumvents the problem of generating fictious samples by inverting the Frequentist probability to obtain the probability of the hypothesis given the data we obtained.

Probability Distributions are central to Bayesian inference

In Bayesian models we quantify uncertainty with probability distributions.

We input probability distributions to Bayes' rule and the output we receive is a probability distribution, so we need to be comfortable with these objects.

Random variables

Definition Given an experiment with sample space, S, a *random variable* is a function from the sample space S to the real numbers $\mathbb{R}$.

A random variable X assigns a numerical value $X(s)$ to each possible outcome of the experiment. Randomness stems from the fact that the outcome of our experiment is random and described by our probability function P.

Example random variable

Example: Suppose we toss a fair coin twice. The sample space $S = \{HH, HT, TH, TT\}$. Let X be the number of heads. This is a random variable with possible values 0, 1, and 2.

X assigns values to each outcome:

- X(HH) = 2
- X(HT) = 1
- X(TH) = 1
- X(TT) = 0

Discrete Probability Distributions

Definition Given an experiment with sample space, S, a random variable is a *discrete random variable* if there is a finite list of values $a_1, a_2, \ldots, a_n$ or an infinite list of values $a_1, a_2, \ldots$ such that $P(X = a_j \,\text{for some}\, j) = 1$.

The set of values x such that $P(X = x) > 0$ is called the support of X.

Probability Mass Function

Definition The *probability mass function* of a discrete random variable X is the function p_X given by $p_X(x) = P(X = x)$. This function is positive if x is in the support of X, and 0 otherwise.

1. PMFs are non-negative, $p_X(x) > 0$ if $x = x_j$ for some j, and $p_X(x) = 0$ otherwise.
2. PMFs sum to 1: $\sum_{j=1}^{\infty} p_X(x_j) = 1$.

Note, $X = x$ is an event consisting of all outcomes s to which X assigns the number x. Formally, we write this as $\{s \in S : X(s) = x\}$.

Example

Suppose we flip a fair coin twice. What is the PMF?

- $p_X(0) = P(X = 0) = \frac{1}{4}$
- $p_X(1) = P(X = 1) = \frac{1}{2}$
- $p_X(2) = P(X = 2) = \frac{1}{4}$

Another example

Example: (if time) We roll two fair 6-sided dice. Let $T = X + Y$ be the total of the two rolls, where X and Y are the individual rolls. The sample space has 36 equally likely outcomes: $S = \{(1,1),(1,2),\dots,(6,6)\}$

What is the PMF of T?

- $P(T = 2) = P(T = 12) = \frac{1}{36}$
- $P(T = 3) = P(T = 11) = \frac{2}{36}$
- $P(T = 4) = P(T = 10) = \frac{3}{36}$
- $P(T = 5) = P(T = 9) = \frac{4}{36}$
- $P(T = 6) = P(T = 8) = \frac{5}{36}$
- $P(T = 7) = \frac{6}{36}$

Our first probability distribution

Definition A random variable X has the Bernoulli distribution with parameter p, if $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 < p < 1$. We write this as $X \sim \text{Bernouli}(p)$. ($\sim$ is read as “distributed as”)

Any event has a Bernoulli random variable associated with it, equal to 1 if the event happens and 0 otherwise. If the event is a “success” or a “failure” we can call it a Bernoulli trial.

Aa quick detour to count things. . .

Definition For any nonnegative integers n and k, the binomial coefficient $\binom{n}{k}$, said “n choose k”, is the number of subsets of size k for a set of size n and equals

$$\frac{n!}{(n-k)! * k!} \tag{1}$$

Expanding our world a bit

Suppose that we have a series of N Bernoulli trials, how is that distributed?

Your second probability distribution

Defintion Suppose that n independent Bernoulli trials are performed, each with the same success probability p. Let X be the number of successes.

The distribution of X is called the binomial distribution with parameters n and p. We write $X \sim \text{Binomial}(n, p)$, where n is a positive integer and $0 < p < 1$.

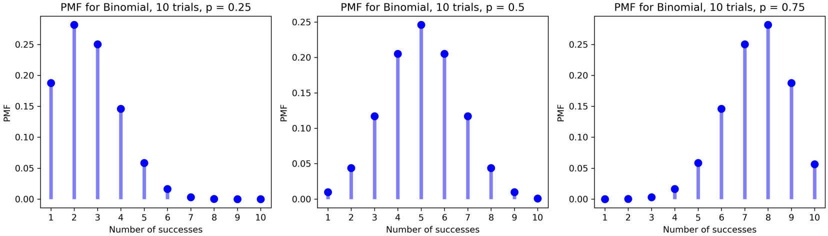
Binomial Probability Mass Function

Definition (Binomial PMF) If $X \sim \text{Binomial}(n, p)$, then the PMF of X is

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{(n-k)} \tag{2}$$

for k = 0, 1, . . . , n; and $P(X = k) = 0$ otherwise.

Example binomial PMFs


PMF for Binomial, 10 trials, p = 0.25

PMF for Binomial, 10 trials, p = 0.5

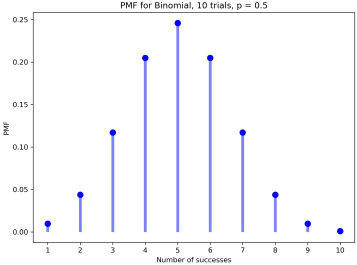
PMF for Binomial, 10 trials, p = 0.75

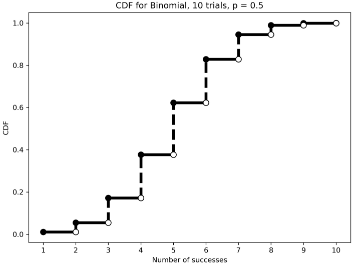
Cumulative Distribution Function

Definition The cumulative distribution function of a random variable X is the function F_X given by $F_X(x) = P(X \leq x)$.

1. The CDF is increasing in x, if $x_1 \leq x_2$ then $F_X(x_1) \leq F(x_2)$.
2. The CDF converges to 0 in the limit as x approaches $-\infty$; and converges to 1 in the limit as x approaches ∞.

Example Binomial PMF and CDF





Continuous random variables

Continuous random variables can take on any real value in an interval. For example: the wing span of an eagle, the height of a person, the length of time until a light bulb burns out.

Definition A random variable has a continuous distribution if the CDF is differentiable.

Cumulative Distribution Function, Continuous Version

Definition For a continuous random variable X with CDF F, the probability density function of X is the derivative of the CDF, given by $f_X(x) = F'(x)$. The support of X, and of its distribution, is the set of all x where $f(x) > 0$. Valid PDFs are nonnegative, $f(x) \geq 0$; and integrate to 1, $\int_{-\infty}^{+\infty} f(x)dx = 1$.

A key difference between discrete and continuous random variables is that for a continuous random variable, the PMF, $P(X = x) = 0$ for all x. $P(X = x)$ is the height of a jump in the CDF at x, but the CDF of X has no jumps!

The Uniform Distribution

Definition A continuous random variable U is said to have the uniform distribution on the interval (a, b) if its PDF is

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Normal (Gaussian) Distribution

The normal distribution is the “bell-shaped curve” everyone always talks about. It is widely used in statistics, because of the Central Limit Theorem which says that under weak assumptions, the sum of a large number of independent and identically distributed random variables has an approximately normal distribution.

Standard Normal PDF

Definition A continuous random variable Z is said to have the standard Normal distribution if its PDF is given by:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right), \text{for} -\infty < z < \infty \tag{4}$$

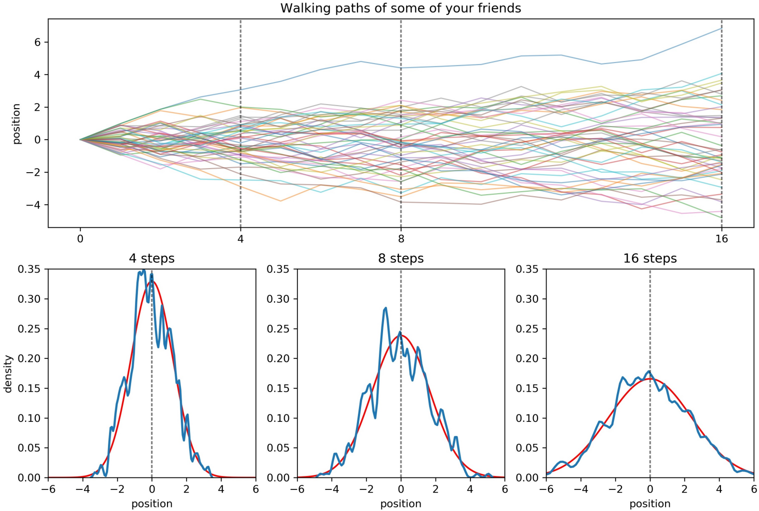
$\frac{1}{\sqrt{2\pi}}$ is a normalizing constant that makes the PDF sum to 1.

Why is the normal distribution so common?

Suppose you and a thousand of your closest friends line up on the halfway line of a soccer field. Each of you has a coin in your hand. At the sound of a whistle, you begin flipping the coins. Each time a coin comes up heads, that person moves one step towards the left-hand goal. Each time a coin comes up tails, that person moves one step towards the right-hand goal.

Each person flips the coin 16 times, follows the implied moves, and then stands still. Now we measure the distance of each person from the halfway line. Can you predict the proportion of people who are standing on the halfway line? How about the proportion 5 yards away? We claim that the distribution of people around the halfway line will be approximately normal.

Did you imagine something like this?



Why might we use the Normal distribution as a base case?

Ontological justification

The world is full of Gaussian distributions, approximately. It is a widespread pattern appearing in measurement errors, variations in growth, and the velocities of molecules. At their heart these processes add together fluctuations, and repeatedly adding fluctuations results in a distribution of sums that have shed all information about the underlying process except for the mean and spread.

Why might we use the Normal distribution as a base case?

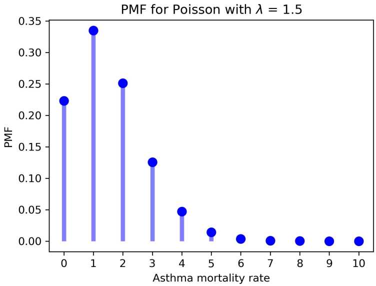
Epistemological Justification

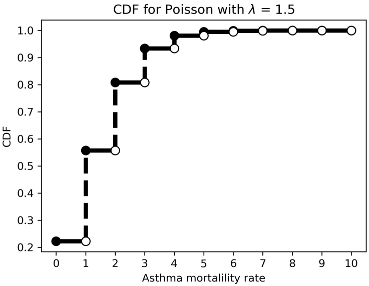
If all we are willing to say about the distribution of a measure is the mean and variance, then the Gaussian distribution is most consistent with our assumptions. If we assume the measure has finite variance, then the Gaussian distribution is the shape that can be realized in the largest number of ways and does not introduce any new assumptions, ie it is the least surprising and least informative assumption we can make.

Example: estimating a rate from Poisson data

- In one US county, It is found that 3 people out of a population of 200,000 died of asthma in a single year.
- A Poisson sampling model is often used for data of this form.
- The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate ($\lambda > 0$) and independently of the time since the last event.
- How can we build a Bayesian model to help us estimate the actual rate of disease?

Example of a Poisson PMF and CDF





Bayes' Rule

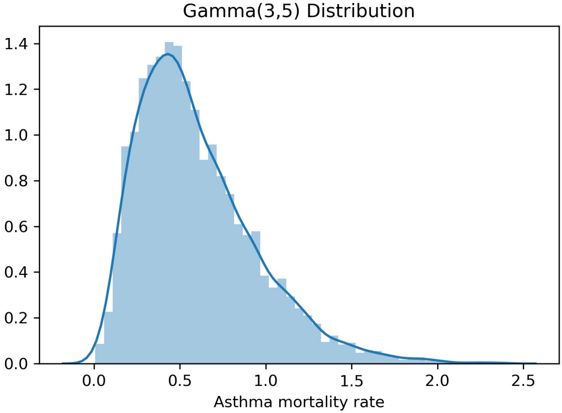
$$P(\lambda|x = 1.5) = \frac{P(x = 1.5|\lambda)P(\lambda)}{P(x = 1.5)} \tag{5}$$

What is $P(\lambda)$?

- $P(\lambda)$ is our prior distribution for the rate of asthma mortality rate.
- We know that a Poisson rate must be greater than zero. So what kind of probability distribution can we place on λ to reflect our prior knowledge?
- Suppose we know that mortality rates in countries similar to the US are rarely above 1.5 per 100,000 people, with typical rates around 0.6.

A reasonable prior

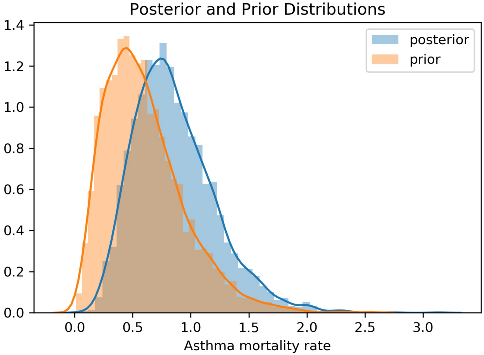
We pick a prior of Gamma(3.0,5.0), because it matches our prior knowledge, and its support is greater than zero.



What's next?

- We specificed our prior, Gamma(3,5).
- We also specified our likelihood, $\mathrm{Poisson}(x = 1.5|\lambda)$.
- Now we can run our black box inference machine to compute our posterior density.
- We will learn how this inference machine works next week, but, for now, trust us that the posterior distribution looks like this. . .

Posterior on asthma mortality rates


Posterior and Prior Distributions
posterior
prior
Asthma mortality rate