1 Theory of Inference

1.1 Goals of week 1

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.

1.2 Materials for week 1

1. Chapters 2 and 3 of *Student's Guide to Bayesian Statistics*
2. Lambert's Lecture 1 slides.
3. **Bonus** Video lecture on the *Bayes' Rule: The Theory That Would Not Die* by Sharon McGrayne, first 45 minutes of the video.

2 First things, 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, more particularly, the scientist constructs hypotheses, or systems of theories, and tests 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.

Now that we have gotten our existential house in order, we can proceed with inferring things about the world from observations of it.

3 Purpose of Statistical inference

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? We develop ideas about how the world works, gather data, construct a model, test the model, then see what we can say about our original query. 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 scientists have spent the last few hundred years developing a mathematical framework within which we can answer such questions. From *Student's Guide* page 17

> Statistical inference is the logical framework which we can use to test our beliefs about the noisy world against data. We formalize our beliefs in models of probability.

Put slightly differently:

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

And, finally, 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

In other words we can assemble the building blocks of probability into models that enable us to draw conclusions about things that we do not observe. For example, we do not know how an MBA will affect someone's salary. There could be a recession, salaries vary widely across industries, within the same industry the level of effort or ability vary by person, or discriminatory practices could be present. All of these factors are likely to affect the change in earnings from acquiring an MBA.

A key component of statistical inference (and particularly Bayesian inference) are probability distributions. We build the model with them, and the model answers us with a probability distribution. Before we turn to that, we present our proposed Bayesian inference engine and a note of caution from prominent social scientists.

4 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. Furthermore, these steps minimize the chance that your model produces a result for the wrong reason or fails to capture key features in your data. 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.

1. **Develop the question**. Iterate until it achieves the right balance of feasibility and usefulness.
2. **Gather data, research current literature**. What's been done before? How much data is available?
3. **Construct model**. For a Bayesian, the model is the prior distribution and the likelihood, "each of which represents some compromise among scientific knowledge, mathematical convenience and computational tractability".
4. **Test your code**. Generate fake data with the model and see if you can recover the known parameters.
5. **Perform prior predictive checks**. How much do you need to constrain priors to rule out outrageous values? Is the model capable of generating data resembling your observed data? Are your priors unduly constraining your model, ie is it capable of detecting potentially competing possibilities?
6. **Estimate the model**. Did the estimation converge? Do you need to tune the estimation?
7. **Carry out posterior predictive checks**. Dream up enough test statistics to convince you and others that the model is useful. What features of the data does the model fail to produce? Can you expand the model somehow to capture this? If not, how fatal is this failure? If it is indeed a failure, re-evaluate your modeling choices and cycle through the process again.
8. **Write up your results**. Decide what your model really tells you about the world and write it up. Ensure that you include your degree of certainty in every model estimate you produce.

Put succinctly, in following this process,

> 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.

5 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.

For more motivation of our Bayesian inference engine see Philosophy and the practice of Bayesian statistics and references listed therein.

5.1 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.

5.2 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 (Gelman, 2003). When we find such particular failures, they tell us how the model must be improved; when severe tests cannot find them, the inferences we draw about those aspects of the real world from our fitted model become more credible. In designing a good test for model checking, we are interested in finding particular errors which, if present, would mess up particular inferences, and devise a test statistic which is sensitive to this sort of misspecification.

Put succinctly, what we're attempting to do with model checking is to determine:

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

5.3 Do not let the perfect be the enemy of the good

A very famous and influential group of political scientists—King, Keohane, and Verba—caution us not to let the perfect be the enemy of the good while conducting statistical inference in the social sciences in their book *Designing Social Inquiry*:

> 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.
>
> Limited information is often a necessary feature of social inquiry. Because the social world changes rapidly, analyses that help us understand those changes require 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.
>
> Joseph Schumpeter once cited Albert Einstein, who said 'as far as our propositions are certain, they do not say anything about reality, and as far as they do say anything about reality, they are not certain' (Schumpeter [1936] 1991:298-99). Yet even though certainty is unattainable, we can improve the reliability, validity, certainty, and honesty of our conclusions by paying attention to the rules of scientific inference.

Now that we're sufficiently philosophically motivated, let's get down to brass tacks.

6 Frequentist World

Let's start with the frequentist view of the 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.

- Frequentists assume the data is random and results from sampling from a fixed and defined population.
- To a frequentist noise that obscures the real population process is attributable to sampling variation.

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.

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

7 Bayesian World

Wash your hands, take a walk, get a sip of coffee, clear your mind. In a Bayesian world fictitious repetitions do not exist. Probabilities here are abstractions which we use to express our uncertainty. Probability is a scale from 0—where we are certain an event will not occur—to 1—where we are certain that it will. In this world we witness the data, and we never perfectly know the value of an unknown parameter. We have epistemic uncertainty.

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?

Take your pick. Either way the math is the same!

8 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.

These differing perspectives set the two on diverging methods for 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.

Put differently:

- Frequentist inference conditions on a null hypothesis to assess the plausibility of the data one observes, with another step of reasoning required to either reject or fail to reject the null hypothesis.
- Bayesian inference involves conditioning on the data at hand to produce posterior probability statements about parameters and hypotheses.

9 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.

9.1 Random variables

Lambert's YouTube lecture on random variables and probability distributions, 7 minutes.

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: 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

- Let Y be the number of tails. How do we define Y in terms of X?
- Suppose I is 1 if the first toss lands Heads, 0 otherwise. This is an example of an indicator random variable.

9.2 Discrete probability distributions

Lambert's YouTube lecture on discrete distributions, 6 minutes.

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.

We have a random variable which simplifies our lives, but we want to know what it behaves like using probability. For example, what is the probability that a particular random variable falls in a range? If M is the number of earthquakes in California in the next 10 years, what is the probability that M equals 0? 5?

To answer these questions, we put a distribution around the random variable. The distribution places probabilities on each value of the random variable.

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 $\{s \in S : X(s) = x\}$. If X is the number of heads in a coin toss example, $P(X = 1)$ is the probability that the coin lands heads one time.

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

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

What is the PMF of I?

- $p_I(0) = P(I = 0) = \frac{1}{2}$
- $p_I(1) = P(I = 1) = \frac{1}{2}$

Let's plot the PMF: Draw on the board the values of p_X and p_I in separate histograms. Note the sum of the bars equals 1 in each plot, and that values outside the support equal 0.

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: $S = \{(1,1),(1,2),\ldots,(6,6)\}$

What is the PMF of T?

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

Draw this on the board.

Our first probability distributions!

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 $X \sim$ 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.

Now suppose that we have a series of N Bernoulli trials, how is that distributed? First another definition.

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

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

Defintion Suppose that n indepensdent 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 $X \sim$ Binomial(n, p), where n is a positive integer and $0 < p < 1$.

Definition (Binomial PMF) If $X \sim$ 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.

Handout or show on the board various examples of the binomial distribution. Binomial(10,1/2) Binomial(10,1/8) Binomial(100,0.03) Binomial(9,4/5)

Another useful way to describe the distribution of a random variable is the cumulative distribution function, CDF.

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: Let $X \sim$ Binomial$(4, \frac{1}{2})$. What are the values of the PMF?

- $P(X = 0) = 0.062$
- $P(X = 1) = 0.25$
- $P(X = 2) = 0.375$
- $P(X = 3) = 0.25$
- $P(X = 4) = 0.062$

The values of the CDF are:

- $P(X \leq 0) = 0.062$
- $P(X \leq 1) = 0.312$
- $P(X \leq 2) = 0.688$
- $P(X \leq 3) = 0.938$
- $P(X \leq 4) = 1$

9.3 Continuous random variables

Lambert's YouTube lecture on continuous distributions, 8 minutes.

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.

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 value of the PDF at a particular value of x is not a probability. To obtain a probability we have to integrate the PDF.

$$F(x) = \int_{-\infty}^{+\infty} f(t)dt \tag{3}$$

And we're just about done with calculus!

This is analogous to how we went from the PMF to the CDF for a discrete random variable, we summed up all values of the PMF that were less than or equal to x. For continuous random variables we have to integrate to sum up all values of the PDF less than or equal to x.

9.4 Uniform Distributions

A uniform random variable on the interval (a,b) is a completely random number between a and b, ie the PDF is constant over the interval.

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

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The area under the curve is a rectangle with a width of $(b - a)$ and height $\frac{1}{b-a}$, so the area equals 1.

The CDF of a uniform random variable is:

$$F(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases} \tag{5}$$

Draw a uniform PDF and CDF on the interval (0,1).

9.5 Normal 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.

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

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right), \text{for} - \infty < z < \infty \tag{6}$$

$\frac{1}{\sqrt{2\pi}}$ is a normalizing constant that makes the PDF sum to 1. Show plots of a normal PDF and CDF.

A couple of notable properties:

1. The normal distribution is symmetric, ie $f(z) = f(-z)$.
2. Area under the PDF to the left of -2, $P(Z \leq -2) = \phi(-2)$, equals the area to the right of 2, $P(Z \geq 2) = 1 - \phi(2)$.

9.5.1 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.

In this situation we are summing a large number of independent factors, and the *Central Limit Theorem* claims that the more times you sample, the more your distribution will look like a Normal distribution.

9.5.2 Why might we use the normal distribution as a base case?

1. *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.
2. *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.

9.6 Gamma Distribution

9.7 Poisson Distribution

10 Apply Bayes Rule to a real data point

Watch Lambert's YouTube lecture on the intuition of Bayes' rule, 8 minutes.

11 Lab

11.1 Experimenting with probability mass functions

Sample from binomial and Poisson distributions.

11.2 Roll some dice

Simulate rolling from two-dice.

11.3 Central limit theorem

Simulate samples from various probability distributions and compute the mean. Does the distribution of the mean approximate a normal distribution as you increase the number of experiments you conduct?

11.4 Intuition of Bayesian reasoning with coin flips

Work through various coin flipping scenarios to see how Bayesian updates compare with your intuition.

11.5 Decisions about elevators

Material: slides 130 to 233 of Lecture 1

Imagine we want to create a model for the frequency a lift (elevator) breaks down in a given year, X . This model will be used to plan expenditure on lift repairs over the following few years.

An aside: how to survive a falling lift

1. Assume a range of unpredictable and uncorrelated factors (temperature,lift usage, etc.) affect the functioning of the lift, so we say $X \sim \text{Poisson}(\lambda)$, where λ is the mean number of times the lift breaks in one year.

2. By specifying that X is Poisson-distributed we assume that there is a continuous probability of failure over time.

3. *Important*: we don't a priori know the true value of therefore our model defines collection of probability models; one for each value of . We call this collection of models the Likelihood.

By specifying a model framework X Poisson() we defined the boundaries of the "Small World". The Small World contains a collection of probability distributions known as the Likelihood.

11.5.1 First question

Compute the posterior of the average number of failures per year given the data for the last 10 years. Compute the 90 percent credible interval. (`np.quantile`)

11.5.2 Second question

Using the posterior you computed in the first question, it costs $1000 per repair visit, and we budget $10,000 a year for lift repairs. Over the next 5 years, how many times do we expect to face a repair bill of $15,000 or greater?

11.5.3 Third question

Suppose you can sign a service contract with an elevator repair company. You can pre-pay for as many repairs in a year as you want for $1000 per repair visit. If the repair is not prepaid, a repair visit costs $1500. How many visits should you prepay? Write a paragraph justifying to your boss why you chose this number and a prediction for the number of times in the next 5 years you expect to exceed the prepaid budget.

11.5.4 Challenge problem

Suppose you have 4 friends in other buildings who face the same problem as you, and you are able to collectively buy a service contract. Compute the combined Poisson parameter and its credible interval. Would you save money over the next 5 years if you collectively bought a contract?

11.6 Extra material

11.6.1 Where does all this fit? Inductive vs Deductive research

7 minute video lecture on inductive and deductive research approaches in social science.

There are a couple of paradigms for how we go about creating knowledge: * **Inductive Research** begins with a research question and the collection of empirical data, which are used to generate hypotheses and theory. The question drives data collection and tests, which we can use to form theories about the world. * **Deductive Research** approaches begin with a theory-driven hypothesis, which guide data collection and analysis. Theory provides the hypothesis, and we use observations to test.

11.7 Exponential Distribution

The exponential distribution is used to model the waiting time for a successful event when successes arrive at a rate of lambda successes per unit of time. The average number of successes in time interval t is $t\lambda$.

Definition A continuous random variable X is said to have the exponential distribution with parameter lambda, where $\lambda > 0$, if its PDF is

$$f(x) = \lambda \exp(-\lambda x). \tag{7}$$

The corresponding CDF is

$$F(x) = 1 - \exp(-\lambda x) \tag{8}$$

and its support is $(0, \infty)$.