WEEK 10 – INTRO TO BAYES

BAYES' THEOREM

- For two events A and B, the conditional probability of A given B is:

  $\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$ where ∩ means“the intersection of” (both A and B occur)

- Let A^c be the complement of A. Bayes' theorem allows us to compute $\Pr(A \mid B)$ via

$$\Pr(A \mid B) \ = \frac{\Pr(B|\ A)\Pr(A)}{\Pr(B\ |A)\Pr(A) + \Pr(B\ |A^c)\Pr(Ac)}$$

- Which simplifies to, when $\Pr(B) > 0$

$$\Pr(A\ |B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

A is some proposition about the world *B* is some data/evidence

$$\Pr(A\,|B) = \frac{\Pr(B\mid A)\Pr(A)}{\Pr(B)}$$

Example: *A* represents the proposition that it rained today, and *B* represents the evidence that the grass outside is wet

Pr(rain | wet grass) = "What is the probability that it rained today given that the grass outside is wet"

Before looking at the ground, what is the probability that it rained, Pr(rain)?

Think of this as the plausibility of an assumption about the world.

- We then ask “how likely the observation that the grass is wet outside is under that assumption?” Pr(wet grass | rain)
- This updates our initial beliefs about the proposition (that it rained today) with some observation (that the grass is wet).

This is called Bayesian *Inference*

- Our *initial* beliefs are represented by the **prior distribution**

Pr(rain)

- and our *final* beliefs are represented by the **posterior distribution**

Pr(rain |wet grass)

- The denominator simply asks,

"What is the total plausibility of the evidence?"

BAYESIAN THINKING

We provide our understanding of a problem and some data, and in return get a quantitative measure of how certain we are of a particular fact.

This approach to modeling uncertainty is particularly useful when:

- Data is limited
- We have reason to believe that some facts are more likely than others, but that information is not contained in the data we model on
- We're interested in precisely knowing how likely certain facts are, as opposed to just picking the most likely fact

FREQUENTISTS VERSUS BAYESIANS

FREQUENTIST

- For frequentists, probabilities are fundamentally related to frequencies of events. Probability only has meaning in terms of a limiting case of repeated measurements.
- Example: you are standing in a wind tunnel and you measure wind speed repeatedly
- Each time the result will be slightly different due to the statistical error of the measuring device.
- In the limit of many measurements, the frequency of any given value indicates the probability of measuring that value.
- the true wind speed is, by definition, a single fixed value, not a distribution

FREQUENTISTS VERSUS BAYESIANS

BAYESIAN

- It's ok to be uncertain about what is true because we can use data as evidence that certain facts are more likely than others
- A Bayesian might claim to know the wind speed, U with some probability Pr(U): that probability can certainly be estimated from frequencies in the limit of a large number of repeated experiments, but this is not fundamental
- Probability is related to your own knowledge about an event. We can meaningfully talk about the probability that the true wind speed lies in a given range.
- That probability codifies our knowledge of the value based on prior information and available data.

FREQUENTIST

Sampling from a particular model (the likelihood) repeatedly. The likelihood defines the distribution of the observed data conditional on the unknown parameter(s).

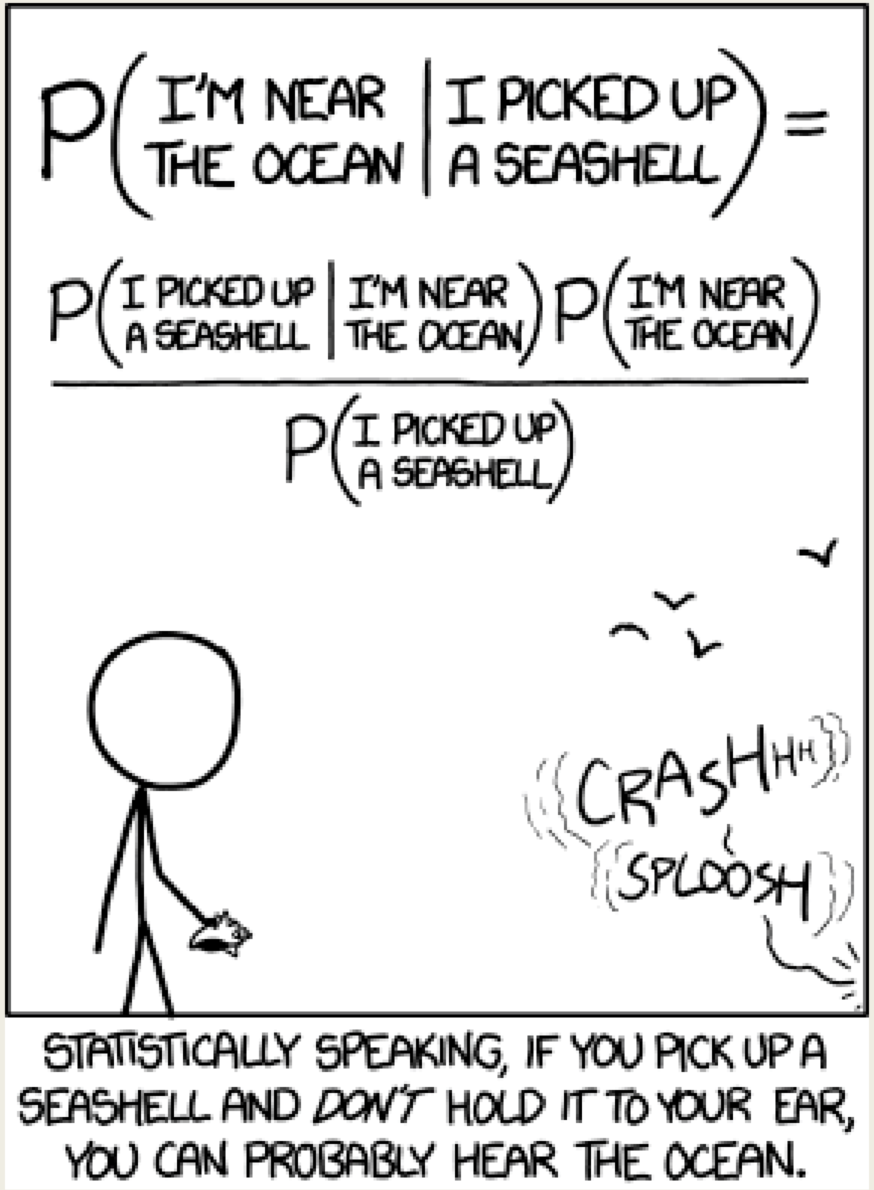
- A parameter is fixed but an unknown constant.
- The probabilities are always interpreted as long-run relative frequencies.
- Statistical procedures are judged by how well they perform in the long-run over some infinite number of repetitions of the experiment

BAYESIAN

- sampling a model (likelihood) and also knowing a prior distribution on all unknown parameters in the model
- the likelihood and prior are combined in such a way that we compute the distribution of the unknown parameter given the data (posterior distribution)

- A parameter is considered to be a random variable and has a distribution.
- The prior distribution placed on the unknown parameter quantifies our beliefs regarding the unknown parameter.
- We use the laws of probability to make inferences about the unknown parameter of interest.
- We update our beliefs about the unknown parameter after getting data (likelihood). This yields the posterior distribution which reweights things according to the prior distribution and the data (likelihood).

But we have to ensure that the posterior is a proper probability distribution.


P(I'M NEAR THE OCEAN | I PICKED UP A SEASHELL) =
P(I PICKED UP A SEASHELL | I'M NEAR THE OCEAN) P(I'M NEAR THE OCEAN)
P(I PICKED UP A SEASHELL)
CRASHHH
SPLOOSH
STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND DON'T HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

https://xkcd.com/1236/

Mosasaur in Kansas

- (example from Davis, 1986)
- Fragment of an unknown species of mosasaur has been found in western Kansas
- We want to send a student party to search for more remains
- Unfortunately, the source of the fragment cannot be identified precisely because it was found in a stream bed downstream of 2 tributaries
- The drainage area of the large stream (L) is 18 square miles, and the other (S) is 10 square miles
- $P(L) = 18/28 = 0.64$
- $P(S) = 10/28 = 0.36$

- However, 35% of Cretaceous rocks in large basin are marine, and 80% of smaller basin are marine
- Conditional probability of a marine fossil given it is found in either basin
- P(marine |L) = 0.35
- P(marine |S) = 0.80
- Using Bayes' theorem: $\Pr(A\,|B) = \dfrac{\Pr(B|\,A)\Pr(A)}{\Pr(B\,|A)\Pr(A) + \Pr(B\,|A^c)\Pr(Ac)}$
- We ask, "what is the probability that the fossil came from basin L, given that the fossil is marine?"
- $\Pr(\,L|marine) = \dfrac{\Pr(marine\,|L)\Pr(L)}{\Pr(marine\,|L)\Pr(B)+\Pr(marine\,|S)\Pr(S)}$
- $\Pr(L\,|marine) = \dfrac{(0.35)(0.64)}{(0.35)(0.64)+(0.80)(0.36)}$ = 0.44
- $\Pr(S\,|marine) = \dfrac{(0.8)(0.36)}{(0.35)(0.64)+(0.80)(0.36)}$ = 0.56

$$\Pr(A\,|B) = \frac{\Pr(B \mid A)\,\Pr(A)}{\Pr(B)}$$

- A = some proposition about the world
- B = some evidence

$$\Pr(\theta\,|y) = \frac{\Pr(y \mid \theta)\,\Pr(\theta)}{\Pr(y)}$$

- θ = some parameters of a statistical model
- y = some data

$$\Pr(\theta) = prior$$

$\Pr(y \mid \theta)$ = likelihood

$$\Pr(\theta \mid y) = posterior$$

Posterior ∝ Likelihood x Prior

Posterior = Prior x Evidence / constant

Are you getting enough sleep?

- What proportion of American college students get at least eight hours of sleep?
- *p* = the (unknown) proportion of students who sleep at least 8 hours.
- In the Bayesian viewpoint, your beliefs about the uncertainty in this proportion are represented by a **prior probability distribution** placed on this parameter. This distribution reflects your subjective prior opinion about plausible values of p.

Defining our prior

- We **believe** that college students generally get less than eight hours of sleep and so *p* is likely smaller than 0.5. Our best guess at the value of *p* is 0.3, but we think it is plausible that this proportion could be any value in the interval from 0 to 0.5

Our <u>prior</u> density for *p* is *g(p)*:

- A ***success*** is sleeping 8 hours or more. A random sample of *s* successes and *f* failures has a likelihood function:

$$L(p) \propto ps\ (1-p)^f$$

Our <u>posterior</u> density for *p* given some data:

$$g(p|data) \propto g(p)L(p)$$

Posterior ∝ Likelihood x Prior

Posterior = Prior x Evidence / constant

β prior

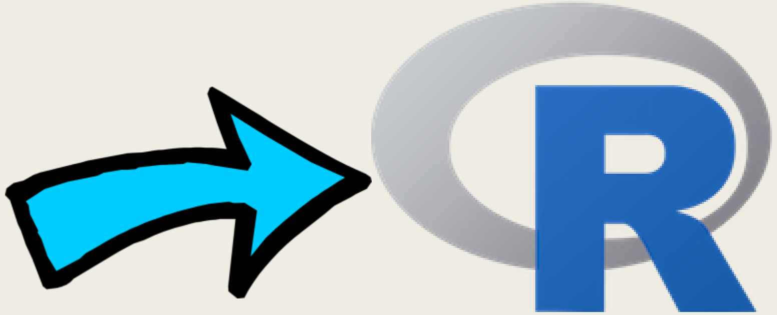
- Since *p* is a continuous parameter, we can construct density *g(p)*
- we think that *p* is equally likely to be more than or less than 0.3, but we're 90% confident that *p* < .9
- We use the beta distribution

$$g(p) \propto p^{a-1}(1-p)^{b-1}$$

we obtain hyperparameters *a* and *b* indirectly through statements about the percentiles of the distribution. We need to solve the following equations:

$$\int_0^{0.3} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} d\theta = 0.5$$

$$\int_0^{0.5} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} d\theta = 0.9$$



Making predictions

- We’ve learnt about our proportion of heavy sleepers, *p*, but now we want to predict the number of sleepers, *y*, in our next sample
- Sample size of *m*

$$f(y) = \int f(y \mid p)\, g(p)\, dp$$

- *g* and *f* are prior and posterior predictive densities

MARKOV CHAIN MONTE CARLO

- MCMC is simply an algorithm for sampling from a distribution.
- It is a type of “Monte Carlo” (i.e., a random) method that uses “Markov chains”
- Used is to draw samples from the **posterior probability distribution** of some model in Bayesian inference. With these samples, you can then ask things like “what is the mean and range for a parameter?”

- The MCMC sampling strategy sets up an Markov chain for which the stationary distribution equals the posterior distribution of interest.
- A general way of constructing a Markov chain is by using a Metropolis-Hastings algorithm.

Quick Markov Chain recap

METROPOLIS-HASTINGS ALGORITHM

- We have some posterior distribution that we want to sample from, and we're going to evaluate some function

$$f(x) \propto p(x)$$

- We also need a probability density function, P, that we can draw samples from.

1. Start in some state *x(t)*
2. Propose a new state *x'*
3. Compute the "acceptance probability"
4. Draw some uniformly distributed random number *u* from [0,1]
5. if *u* < *α* accept the point, setting *x(t+1)* = *x'*. Otherwise reject it and set *x(t+1)* = *x(t)*

METROPOLIS-HASTINGS ALGORITHM

- This will generate a series of samples $[x_0, x_1, \ldots]$
- Note that where the proposed sample is rejected, the same value will be present in consecutive samples.
- Also note that these are not independent samples from the target distribution; they are dependent samples; that is, sample *x(t)* depends on *x(t-1)* and so on.
- However, because the chain approaches a stationary distribution, this dependence will not matter so long as we sample enough points.