Bayesian Bootcamp

Astrostatistics Interest Group
North Carolina State University
Reetam Majumder and Eric Yanchenko

Statistics: big picture

There is some process in the world (universe?) that we want to understand

We assume that this process has a **data-generating mechanism**

Collect **data**

Make **inference**

Estimation process

$$\theta \rightarrow Y \rightarrow L(y;\theta) \rightarrow \hat{\theta}$$

Data generating process

Defined by *model* and parameters *theta*

Unknown

Observed data

Defined by *Y*

Known

Assume model

Defined by *likelihood*

Assumed (hopefully close to reality)

Estimate parameters

Theta hat

Estimated, with **uncertainty**

INFERENCE

Always think about what is known, unknown and what you want to know.

Two paradigms

Inference is where statisticians make their money

Two primary philosophical schools of thought:

Frequentist **Bayesian**

Frequentist

The **data (Y)** is treated as *random*

Parameters (theta) are treated as *fixed*

Statistical procedures have properties in the *long run*, i.e., high **frequency**

R. A. Fisher
20th century

Frequentist estimation

How does a frequentist estimate theta?

Maximum likelihood estimation:

Given your model, what values of theta make the data you saw the *most likely*

Yields point estimates, confidence intervals, hypothesis testing, etc.

Explaining estimates is not always intuitive

Bayesian

The **data (Y)** is treated as *fixed*

Parameters (theta) are treated as *random*

Subjective belief about parameters

Belief about parameters is updated as you observe data

Arguably, how a rational person operates

Thomas Bayes
18th century

Toy example

A new ice cream shop opens in Durham

You want to determine how good it is.

The “goodness” of the restaurant is some unknown parameter ϴ which takes values between 0 and 10.

Goal is to estimate ϴ

Toy example

You are a tough critic so without knowing anything else, you say that there is a 95% chance that Θ is between 2 and 6

Your friends went and said it is the best ice cream they have ever had. Your still skeptical but now you think Θ is likely between 3 and 8.

You go to the parlor and are very impressed with the ice cream. You think that Θ is probably 8.5. But you only went once so it could have been a fluke. You think that there is a 95% chance that Θ is between 8 and 9.

Notice that your belief about the “goodness” of the ice cream (as measured by Θ) is updated each time you get new data

Bayes Theorem

How can we do this belief updating in a principled way? **Bayes theorem**

Likelihood of seeing the data (same as in frequentist)

Prior distribution (reflects your current information about Θ)

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

Posterior distribution of Θ

(reflects information about Θ after observing the data; what we want)

Conditional probability

Bayes theorem relies heavily on **conditional probability**

In life (and statistics), often times we have some information to inform our beliefs (probability) about an outcome (event)

This is **conditional probability**

Let’s look at the ice cream example for conditional probability

Conditional probability

Before talking with anyone, there is a 95% chance that Θ is between 2 and 6

This is an **unconditional** probability: *P(Θ)*

After talking to your friends, you have some new information. Now your probability statement is **conditional** on this new information: *P(Θ | friend's recommendation)*

After you taste the ice cream, you have even more information. Again, the probability is now conditional on all previous information:

P(Θ | your experience and friend's recommendation)

Conditional probability and Bayes theorem

Let's look at Bayes theorem again

Think of the probability as summarizing our information about Θ, which is ultimately what we are after.

Distribution of *Y* is *conditional* on the value of Θ

Prior distribution of Θ is *unconditional* on the data. Could also think of it as conditional on previous studies

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

Distribution of Θ is *conditional* on the observed data *Y*

Key point

Frequentists: all information about Θ comes from the data

Bayesians: information for Θ comes from the data AND prior

Choosing the prior is important and a lively research area and is certainly relevant for your work

Simple example

Let's *mathematically* work through a simple example.

Let Θ be the "goodness" of the ice cream parlor. Θ can take values 1, 2 or 3

Before tasting the ice cream, you think that all possible values of Θ are equally likely, i.e., $P(\Theta=1) = P(\Theta=2) = P(\Theta=3) = ⅓$. This is your prior distribution

Let *Y* be the observed tastiness of the ice cream. *Y* can also take values 1, 2, or 3.

You try the ice cream and observe *Y=3*, it was really good.

The probability of observing *Y=3* depends on the value of Θ.

	P(Y=1|Θ)	P(Y=2|Θ)	P(Y=3|Θ)
Θ=1	0.6	0.3	0.1
Θ=2	0.25	0.5	0.25
Θ=3	0.2	0.2	0.6

Simple example

You want to calculate the probability that Θ=3, given your observation

Let's use Bayes theorem:

$P(\Theta=3 \mid Y=3)$

$= \{P(Y=3 \mid \Theta=3)\ P(\Theta=3)\} / P(Y=3)$

$= \{P(Y=3 \mid \Theta=3)\ P(\Theta=3)\} / \{P(Y=3|\Theta=1)\ P(\Theta=1) + P(Y=3|\Theta=2)\ P(\Theta=2) + P(Y=3|\Theta=3)\ P(\Theta=3)\}$

$= (0.6 * 0.33) / (0.1 * 0.33 + 0.25*0.33 + 0.6 *0.33)$

$= 0.63$

Interpretation: There is a 63% chance that Θ=3, given the observed data.

Notice that this value is between your prior (0.33) and likelihood (0.70)

A note on randomness

Before you observe the data, Y is a random variable

After you observe the data, then you have a realization of the random variable, y

This is fixed now and no longer random

Random here means having a distribution which Reetam will discuss more fully

For frequentists, Θ is fixed. Not random, does not have a distribution

For Bayesians, Θ is random. It does have a distribution which changes after you observe data.

More on Random Variables

This example had a bi-variate distribution.

You can get the individual distributions of ϴ and Y from this table

ϴ (and Y) take 3 unique values, and you distribute the total probability (i.e., 1) among those 3 values.

But what if was ϴ had 1000 unique outcomes? What if ϴ was continuous?

	P(Y=1|ϴ)	P(Y=2|ϴ)	P(Y=3|ϴ)
ϴ=1	0.6	0.3	0.1
ϴ=2	0.25	0.5	0.25
ϴ=3	0.2	0.2	0.6

Frequencies to distributions

Toss a coin **once**. What is the probability distribution of the **number of Heads**?

Let Y be the outcomes; Y = {0, 1}

Let θ be the probability of getting a Head. If you assume a fair coin, $\theta = 0.5$

$P[Y = 1 \mid \theta] = \theta$ and $P[Y = 0 \mid \theta] = 1 - \theta$

$P[Y = y \mid \theta] = \theta^{y}.(1 - \theta)^{1-y}$, for $y = 0, 1$ - this is called a **Bernoulli distribution**.

What if you toss a coin 100 times? What is the space of outcomes?

https://shiny.rit.albany.edu/stat/binomial/

Bernoulli to Binomial

Let Y be the number of Heads when you toss a coin 100 times.

$P[Y = y] = {}^{n}C_{y}\ \theta^{y} (1 - \theta)^{n-y}, y = 0, 1 ,..., 100.$

This is called a **Binomial distribution**.

Important things to consider before we take the next steps:

1. This outcome space is still discrete and finite
2. We can observe the underlying experiment/mechanism
3. The coin tosses are independent of each other and identical, i.e.,

$P[Y_1 = y_1, Y_2 = y_2] = P[Y_1 = y_1].P[Y_2 = y_2]$

The uses of a distribution

Since distributions tend to have a functional form, we can compute quantities of interest analytically instead of having to compute them by hand from a histogram.

E.g., if $Y \sim \text{Binomial}(n, \theta)$

Mean of Y = $n.\theta$ (This is called the **expectation of Y** and denoted as $E[Y|\theta]$)

Variance of Y = $n.\theta.(1 - \theta)$

We can also get individual probabilities like $P[Y = 15 \mid \theta]$, or $P[10 < Y < 20 \mid \theta]$

The **support** of Y is 0, 1, ..., n. The support of θ is (0,1)

Uncertainty

For a scientific question, is it always possible to:

1. Know the exact underlying mechanism/experiment?
2. Observe the mechanism, its outcome, or both?
3. Observe it with certainty?

The Poisson distribution

We'll develop an example based on this post from Brookhaven - https://snews.bnl.gov/popsci/poisson.html

How many meteors will hit the Earth's atmosphere per day?

Assumptions:

1. The fact that one event happens does not change the probability that another event will happen later (independent and identical events)
2. We don't observe the underlying mechanism (exactly, at least)
3. No practical upper limit for how many events we can have, i.e., Y = 0, 1, 2,......

The Poisson distribution

Let Y = number of meteor hits in an hour.

Let θ = rate of meteor hits.

$P[Y = y|\theta] = e^{-\theta}.\theta^{y} / y!$

$E[Y|\theta] = V[Y|\theta] = \theta$

If we have data, we can estimate θ

If we have data **and** prior information on θ, we could estimate a **distribution** of θ

A frequentist analysis

A satellite has been able to observe every single meteor hit for the last 24 hours, and has aggregated the number of hits per hour.

$(y_1, .., y_{24}) = (10, 5, 6, 7, 11, 5, 10, 11, 8, 8, 3, 5, 5, 8, 6, 9, 7, 8, 14, 6, 9, 11, 5, 8)$

$E[Y|\theta] = 7.71$, $V[Y|\theta] = 6.73$. **What is the rate of meteor hits**?

Frequentist estimate using R:

```
Parameters:
       estimate Std. Error
lambda 7.708333  0.5667279
```

What if we had some prior information about θ?

The Gamma distribution

Let θ = time between two events. For example:

- The time until your phone will die
- Time until the next meteor will hit

We can model it as a **Gamma** distribution

$\theta > 0$, and is continuous!

Can be (mathematically, not practically) infinite.

The Gamma distribution

$f(\theta \mid a, b) = k.e^{-\theta b}\, \theta^{a-1}$

$a > 0$, $b > 0$, k is a proportionality constant such that f(.) integrates to 1

$E[\theta|a,b] = a/b$, $V[\theta|a,b] = a/b^2$

Time between events, on an average, is a/b

So b/a events happen every time period (b is in fact, known as a rate as well)

What is $P[\theta = c|a,b]$ for any $c > 0$?

We can only talk in terms of inequalities for continuous random variables

Back to our example! - https://snews.bnl.gov/popsci/poisson.html

Bayes Theorem

How can we do this belief updating in a principled way? **Bayes theorem**

Likelihood of seeing the data (same as in frequentist)

Prior distribution (reflects your current information about Θ)

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

Posterior distribution of Θ

(reflects information about Θ after observing the data; what we want)

Bayesian analysis

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

Since θ is continuous, we will use f(.) instead of P[.] throughout

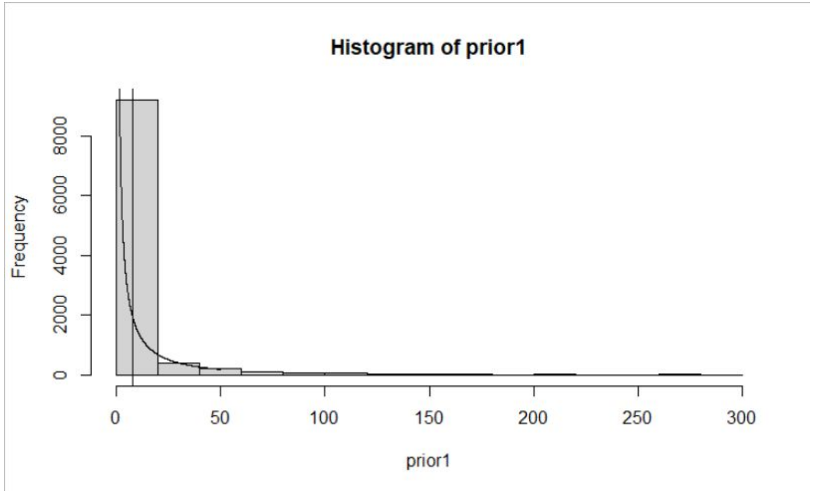
Y|θ ~ Poisson distribution, and θ~Gamma distribution

Note that P[Y] is basically a proportionality constant, can be ignored going forward

What should be choose for a and b?

If we choose a = 0.1, b = 0.1/5.8, we will have a prior whose mean is close to what is suggested by previous experiments

The prior distribution




```
> meteor
 [1] 10  5  6  7 11  5 10 11  8  8  3  5  5  8
6  9  7  8 14  6  9 11  5  8
> a = 0.1; b = 0.1/5.8
> prior1 = rgamma(10000,a,b)
> mean(prior1)
[1] 5.818053
> var(prior1)
[1] 346.2627
> h1 = hist(prior1)
> xx = seq(0.00001,50,by=0.01)
> yy = dgamma(xx,a,b)*diff(h1$mids[1:2]) *10000
> lines(xx,yy,'l')
> abline(v=mean(meteor))
```

Bayes theorem math

$$f(y_1, \ldots, y_{24}|\theta) = \prod_{i=1}^{24} \frac{exp\{-\theta\}\theta^{y_i}}{y_i!}$$

$$\propto \exp\{-24\theta\}\theta^{\sum_i y_i}$$

$$f(\theta|a, b) \propto \exp\{-\theta b\}\theta^{a-1}$$

$$\implies f(\theta|\mathrm{y}, a, b) = \frac{f(\mathrm{y}|\theta)f(\theta|a, b)}{f(y)}$$

$$\propto f(y|\theta)f(\theta|a, b)$$

$$\propto \exp\{-(b + 24)\theta\}\theta^{(a+\sum_i y_i)-1}$$

$$= \exp\{-b^*\theta\}\theta^{a^*-1}$$

That looks like a Gamma distribution?

The posterior of θ will follows a Gamma distribution with parameters

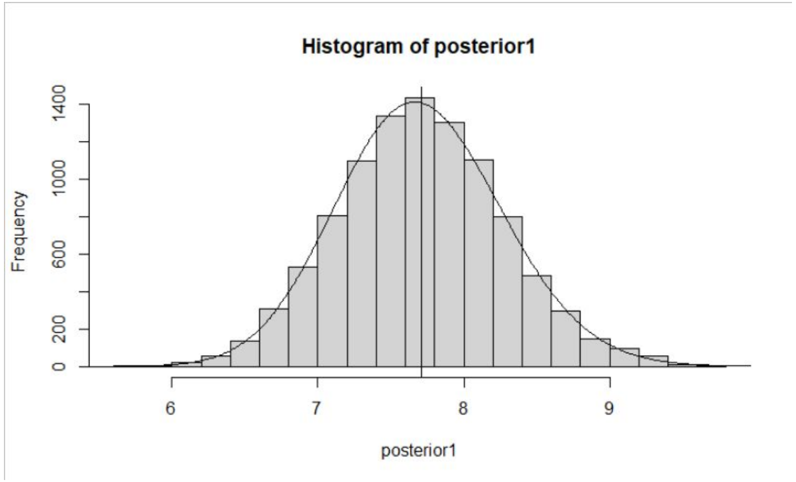
$a^* = a + \Sigma y_i = 185.58$

$b^* = b + 24 = 24.1$

This property, where the posterior has the same distributional form as the prior, is called **conjugacy**.

https://en.wikipedia.org/wiki/Conjugate_prior

The posterior distribution




```
> a_star = a + sum(meteor)
> b_star = b + length(meteor)
> posterior1 =
rgamma(10000,a_star,b_star)
> h1 = hist(posterior1)
> mean(posterior1)
[1] 7.709147
> var(posterior1)
[1] 0.3190985
> sd(posterior1)
[1] 0.5558348
```

The posterior of θ follows a Gamma distribution with parameters a*, b*

Centered around 7.71 (similar to the frequentist case)

Frequentist inference

Based on the sample of observations, the mean is around 7.71, SE = 0.57

Because it's a sample, there is noise, and this isn't completely accurate

Maybe if you had a much larger sample, you can 'identify' θ more accurately

But it's still just a fixed value - there is fuzziness around it but you don't know what that fuzziness looks like mathematically

What is $P[5 < \theta < 10 \mid Y]$?

If you have new data, you either pool it, or start over

Bayesian inference

$\theta|Y$ follows a Gamma distribution with mean 7.1 and variance 0.56

We can make probability statements!

```
> pgamma(10,a_star,b_star) - pgamma(5,a_star,b_star)
[1] 0.9999036
> pgamma(7.8,a_star,b_star) - pgamma(5.8,a_star,b_star)
[1] 0.5744631
> 1 - pgamma(10,a_star,b_star)
[1] 9.638848e-05
```

If we get new data, we can treat this as the prior and update our beliefs about θ

Physical processes are rarely deterministic, so this is intuitive

Doing this in JAGS

Maybe you don't want to do all the math

The math is often complicated for

1. More complicated models
2. When the prior is not conjugate (this is often - conjugacy is convenient but not necessarily the best option)

Software include base R, JAGS, STAN, Python, Julia

Other relevant resources

Distributions:

1. Normal: https://en.wikipedia.org/wiki/Normal_distribution
2. Beta: https://en.wikipedia.org/wiki/Beta_distribution (Special case: Uniform)
3. Exponential: https://en.wikipedia.org/wiki/Exponential_distribution (special case of Gamma)

List of conjugate priors: https://en.wikipedia.org/wiki/Conjugate_prior

Textbook: https://www.bayesianmodelsforastrophysicaldata.com/