Bayesian Statistics

Dustin Leatherman

September 27, 2020

Contents

1 Intro (2020/09/10)

1.1 Motivating Example

There are two students: Student A and Student B, along with an instructor. A secretly writes down a number (1,..,10) then mentally calls heads or tails.

1. The instructor flips a coin
2. If heads, A honestly tells B if the number is even or odd.
3. If A guesses H/T correctly, A tells B if their number is even or odd. Otherwise, they lie.
4. B will guess if the number is odd or even

   Let θ be the probability that B correctly guesses even or odd.

   The class (and myself) initially agreed without much discussion that 0.5 is the obvious answer. Upon further thinking on this, the probabilities breakdown in such way:

   (2): 0.5 (3): 0.5 (4): ?

   The initial logic is that its a 50/50 chance since there are two choices but there is an X-factor here with number 4. A few questions worth asking:

   1. Does B know the rules upfront? As in, are they

   aware that A may or may not lie?

1. Does B see the result of the coin flip?
2. Is this done virtually or in person?

If the answer is no for 1 and 2, then 0.5 is a logical choice because they'd be guessing without much foreknowledge.

If the answer is yes for 1 and 2, then B is in on the "game" and can make a more educated guess. If A or the professor has a "tell", then that could provide information. Reading body language may also provide some information to B on the veracity of A's claim.

I would argue that θ would be > 0.5 *if* A and B know each other well enough. Which is really a great example of Bayesian vs Frequentist view points.

1.2 Frequentist Approach

Quantifies uncertainty in terms of repeating the process that generated the data many times.

1.2.1 Properties

- The parameters θ are fixed, unknown, and a constant.
- The sample (data) Y are random
- All prob. statements would be made about the randomness in the data.
-

A statistic $\hat{\theta} = Y/n$ is a statistic and is an estimator of the population proportion θ

The distribution of $\hat{\theta}$ from repeated sampling is the *sample distribution*.

1.2.2 Things a Frequentist would never say

- $P(\theta > 0) = 0.6$ because θ is not a random variable
- The distribution of θ is Normal(4.2,1.2)
- The probability that the true proportion is in the interval (0.4, 0.5) is 0.95.
- The probability that the null hypothesis is true is 0.03.

1.3 Bayesian Approach

Expresses uncertainty about θ using probability distributions. θ is still fixed and unknown.

Distribution *before* observing the data is the **prior distribution**. e.g. $P(\theta > 0.5) = 0.6$. This is subjective since people may have different priors.

Hopefully, Uncertainty about θ is reduced after observing the data.

Bayesian Interpretations differ from *Frequentist* Interpretations.

Uncertainty distribution of θ after observing the data is the **posterior distribution**.

Bayes Theorem for updating the prior

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)} \tag{1}$$

Described in words: Posterior $\propto$ Likelihood $\times$ Prior

> $f(\theta|Y)$ is the posterior distribution.
>
> Given that I have seen some data, what am I seeing now?

A key difference between Bayesian and frequentist statistics is that all inference is conditional on the single data set we observed (Y).

1.3.1 Likelihood Function

Distribution of the observed data given the parameters. This is the Same function used in a maximum likelihood analysis.

> When prior information is weak, Bayesian and Maximum Likelihood Estimates are similar.

1.3.2 Priors

Say we observed Y = 60 successes in n = 100 trials and $\theta \in [0, 1]$ is the true probability of success.

Want to select a prior that has a domain of [0, 1]

If there is no relevant prior information, we might use $\theta \sim Uni(0, 1)$. This is called an *uninformative prior*. aka a "best guess".

1. Beta

Beta distributions are a common prior for parameters between 0 and 1.

If $\theta \sim Beta(a, b)$, then the posterior is

$$\theta|Y \sim Beta(Y + a, n - Y + b)$$

$Beta(1, 1) == Uni(0, 1)$

2. Gamma Popular distribution for σ (population standard deviation)

1.3.3 Posteriors

The likelihood function $Y|\theta \sim Bin(n, \theta)$
The Uniform prior is $\theta \sim Uni(0, 1)$
The posterior is then $\theta|Y \sim Beta(Y + 1, n - Y + 1)$

1.3.4 Advantages

- Bayesian concepts (posterior probability of the null hypothesis) are arguably easier to interpret than the frequentist ideas (p-value.)
- Can incorporate scientific knowledge via the prior.
  - Even a Small amount of prior information can add stability.
- Excellent at quantifying uncertainty in complex problems.
- Provides a framework to incorporate data/information from multiple sources.

1.3.5 Disadvantages

- Less common/familiar
- Picking a prior is subjective (though there are objective priors)
- Procedures with frequentist properties are desirable.
- Computing can be slow for hard problems
- Non parametric methods are challenging

1.4 Review

Only the interesting parts are placed here. See the rest of this repo for deeper dives on other concepts.

1.4.1 Probability

Objective (associated with Frequentist)

- $P(X = x)$ is a mathematical statement
- If we repeatedly sampled X, the value that the proportion of draws equal to x converges is defined as $P(X = x)$

Subjective (associated with Bayesian)

- $P(X = x)$ represents an individual's degree of belief
- Often quantified as the amount an individual would be willing to wager that X will be x.

A Bayesian Analysis uses both of these concepts.

1.4.2 Uncertainty

Aleatoric (def: indeterminate) uncertainty (likelihood)

- Uncontrollable randomness in the experiment

Epimestic (def: involving knowledge) uncertainty (prior/posterior)

- Uncertainty about a quantity that could be theoretically

A Bayesian Analysis uses both of these concepts

1.4.3 Probability vs Statistics

The common sense, I like the way this is phrased.

Probability is the forward problems

- We assume we know how the data are being generated and computer the probability of events.

  For example, what is the probability of flipping 5 straight heads if the coins are fair?

Statistics is the inverse problem

- We use data to learn about the data-generating mechanism

  For example, if we flipped five straight heads, can we conclude the coin is biased?

2 Probability & Introduction to Bayes (2020/09/17)

if x and y are independent, then the following is true

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

Cannot use $f(x,y)$ as PMF because $\sum_1^Y f(x,y) = f(x) \neq 1$. Need to scale by marginal probability in order to sum to 1 and thus be a proper PMF/PDF.

	1	2	3	4	5	Total [p(y)]
US	.0972	.0903	.0694	.0069	.0069	.2708
Not US	.3194	.1319	.1389	.1181	.0208	.7292
Total [p(x)]	.4167	.2222	.2083	.1250	.0278	1

show that x and y are dependent
$P(x = 1) = 0.4167$
$P(y = 1) = 0.2708$
$P(x = 1) \times P(y = 1) = 0.4167(0.2708) = 0.1128$
$P(x = 1, y = 1) = 0.0972 \neq 0.1128$ so dependent!

2.1 Calculating the Posterior Analytically

2.1.1 Using an Arbitrary PDF

1. Find Joint Probability (f(x,y))

$$
\begin{aligned}
P(x>7,y>40) = & \int_7^{10}\int_{40}^{50} 0.26exp(-|x-7|-|y-40|)\ dx\ dy \\
=& 0.26\int_7^{10}\int_{40}^{50} exp(-x+7-y+40)\ dx\ dy \text{ (Since only interested in positive values)} \\
=& 0.26\int_7^{10}\int_{40}^{50} exp(-(x-7)\ exp(-(y-40))\ dx\ dy \\
=& 0.26\int_7^{10}\int_{0}^{10} exp(-(x-7)\ exp(-u)\ dx\ du \\
=& 0.26\int_7^{10}\int_{0}^{10} exp(-(x-7)\ [-exp(-u)]_0^{10}\ dx\ du \\
=& 0.26(1-e^{-10})\ \int_7^{10} exp(-(x-7)dx \\
=& 0.26(1-e^{-10})(1-e^{-3}) \approx 0.247
\end{aligned} \tag{2}
$$

1. Find Marginal Probability over the Data $f_X(x)$

$$
\begin{aligned}
f_X(x) =& int_{20}^{50} 0.26 + e^{-|x-7|-|y-40|}dy \\
=& 0.26e^{-|x-7|}\int_{20}^{50} e^{-|y-40|}dy \\
=& 0.26e^{-|x-7|}[\int_{20}^{40} e^{-(40-y)}dy + \int_{40}^{50} e^{-(y-40)}dy] \\
=& 0.26e^{-|x-7|}[\int_{20}^{0} -e^{-u}du + \int_{0}^{10} e^{-u}du] \\
=& 0.26e^{-|x-7|}[1-e^{-20}+1-e^{-10} \approx 2] \\
=& 0.52e^{-|x-7|}\ \forall\ x \leq x \leq 10
\end{aligned} \tag{3}
$$

1. Calculate Conditional Probability

$$
f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{2}e^{-|y-40|}
$$

If integrating over an absolute value, break up the integral into two integrals: the first over the negative domain of the integration, the second over the positive domain.

2.1.2 Using Normal Distribution

1. Find Marginal Probability

$$
\begin{aligned}
f(x) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\ exp(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)})\ dy \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-x^2/2(1-\rho^2)}\ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} exp(-\frac{y^2-2\rho xy}{2(1-\rho^2)})\ dy \text{ (Move x's out of integral. Arrange term} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-x^2/2(1-\rho^2)}\ \int_{-\infty}^{\infty} \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi(1-\rho^2)}} exp(-\frac{y^2-2\rho xy+\rho^2x^2-(\rho x)^2}{2(1-\rho^2)})\ dy \\
&= \frac{1\ \sqrt{1-\rho^2}}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-x^2/2(1-\rho^2)}\ e^{\frac{\rho x^2}{2(1-\rho^2)}}\ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} exp(-\frac{(y-\rho x)^2}{2(1-\rho^2)})\ dy,\ N(\rho x, 1-\rho^2) \\
&= \frac{1}{\sqrt{2\pi}} e^{-0.5\ \frac{x^2-\rho^2x^2}{1-\rho^2}} \\
&= \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}, X \sim N(0,1)
\end{aligned} \tag{4}
$$

1. Assume Joint Normal PDF

2. Find Conditional probability

$$
\begin{aligned}
f(y|x) =& \frac{f(x,y)}{f_X(x)} \\
=& \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} exp(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)})}{\frac{1}{\sqrt{2\pi}}\ exp(-\frac{x^2}{2})} \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)} + \frac{x^2}{2}) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{x^2+y^2-2\rho xy}{1-\rho^2} - x^2]) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{x^2+y^2-2\rho xy-(1-\rho^2)x^2}{1-\rho^2}]) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{y^2-2\rho xy-\rho^2x^2}{1-\rho^2}]) \\
=& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\ exp(-\frac{1}{2}[\frac{(y-\rho x)^2}{1-\rho^2}]),\ y|x \sim N(\rho x, 1-\rho^2)
\end{aligned} \tag{5}
$$

2.2 Bayes Theorem

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

How do you know you are using Bayes Rule?
Given $P(y|\theta)$, want to find $P(\theta|y)$

- Bayesians quantify uncertainty about fixed but unknown parameters by treating

them as random variables.

- This requires that we set a prior distribution $\pi(\theta)$ to summarize uncertainty before observing the data.
- The distribution of the observed data given the model parameters is the *likelihood function*, $f(Y|\theta)$
  - The likelihood function is the most important piece of a Bayesian Analysis because it links the data and the parameters.

2.3 Bayesian Learning

The posterior distribution $P(\theta|Y)$ summarizes uncertainty about the parameters given the prior and data.

Reduction in uncertainty from prior to posterior represents **Bayesian Learning**

Bayes Theorem (again):

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

$m(Y) = \int F(Y|\theta)\pi(\theta)d\theta$: marginal distribution of the data and can usually be ignored.

2.4 Subjectivity

Choosing Likelihood function and a prior distribution are subjective.

If readers disagree with assumptions, findings will be rejected so assumptions must be justified theoretically and empirically.

3 Summarizing a Posterior Distribution (2020/09/24)

3.1 SIR Model

Susceptible-Infected-Recovered

At time t , $S_t + I_t + R_t = N$ where N is the population.

States evolved according to the following differential equations

$$\begin{aligned} \frac{dS_t}{dt} =& -\beta\frac{S-tI_t}{N} \\ \frac{dI_t}{dt} =& \beta\frac{S_tI_t}{N} - \Gamma I_t \end{aligned} \tag{6}$$

β: Controls rate of new infections

Γ: Controls recovery rate

We will use a discrete approx to these curves with hourly time steps.

So? $dt = \frac{1}{24}$

Goal: Fit SIR Model for given values of β and Γ

3.2 Summarize a univariate Posterior with Beta-Binomial

Posterior = Likelihood × Prior

Say there is a parameter θ

Likelihood: $Y|\theta \sim Bin(N, \theta)$

Prior: $\theta \sim Uni(0,1) \equiv Beta(1,1)$

Posterior: $\theta|Y \sim Beta(Y + a, N - y + b)$

> Peak of the Posterior is the MLE of the Likelihood function when using an uninformative prior.

3.3 MAP Estimator

Posterior Mode is call the max a posterioiri (MAP) estimator

$$\hat{\theta} = \underset{\theta}{argmax}\ P(\theta|y) = \underset{\theta}{argmax}\ log[f(Y|\theta)] + log[\pi(\theta)] \tag{7}$$

if prior is uniform, MAP is MLE assuming $Y|\theta \sim Bin(\theta, n)$.

3.4 Uncertainty Measures

Posterior Std. Dev. is one measure of uncertainty

- If approx Gaussian, can use empirical rule
- Analogous but fundamentally different than std error.
  - Std err is the standard deviation of $\hat{\theta}$'s sampling distribution

Do not call them call them **confidence** intervals. Called **Credible** Intervals in Bayesian Statistics.

Interval (l, u) is $100(1-\alpha)\%$ posterior credible interval if $P(l < \theta < u|Y) = 1 - \alpha$

Interpretation: "Given the data and the prior, the probability that θ is between l and u is 0.95."

> Confidence interval interpretation:
>
> With 95% Confidence, θ is between l and u.
>
> A Bayesian Posterior is a distribution for $\theta|Y$ whereas the sampling distribution is for $\hat{\theta}$. While their expected values are both μ, the sampling distribution is not a distribution of θ hence why

"Confidence" is used when in the interpretation. The Bayesian Posterior is a distribution of θ so the posterior can be used for the interpretation.

3.4.1 Credible Sets

Not unique.

Let q_τ be the τ quantile of the posterior of the posterior such that $P(\theta < q_\tau | Y) = \tau$. Then $(q_{00}, q_{0.95})$, $(q_{0.01}, q_{0.96})$, etc. are all valid 95% credible sets.

Equal-Tailed intervals: $(q_{\alpha/2}, q_{1-\frac{\alpha}{2}})$

Highest posterior density interval searches for the smallest interval that contains the proper probability

3.5 Hypothesis Tests

Conducted by computing posterior prob of each hypothesis.

$$P(\theta < 0.5|Y) = \int_0^{0.5} P(\theta|Y) d\theta$$

analogous but different than a p-value.

p-value: Assuming the null hypothesis is true, the probability we got X or a value more extreme is Y.

Bayesian Hypothesis Test: Given the prior and the data, the probability the null hypothesis is true is Y.

3.6 Monte Carlo Sampling

A useful tool for summarizing a posterior.

In MC sampling, we draw S samples from the posterior;

$$\theta', ..., \theta^{(s)} \sim P(\theta|Y)$$

and use these samples to approx the posterior.

3.6.1 Transformations

MC sampling facilitates studying the **transformations** of parameters.

For example, the odds corresponding to θ are $\gamma = \frac{\theta}{1-\theta}$

$$\gamma^{(1)} = \frac{\theta^{(1)}}{1 - \theta^{(1)}}, ..., \gamma^{(S)} = \frac{\theta^{(S)}}{1 - \theta^{(S)}} \tag{8}$$

How to approximate the posterior mean and variance of γ?
Transform the odds and use the draws to approximate θ's posterior!

3.7 Summarizing Multivariate Posteriors

Univariate posteriors captured by a simple plot. Not easy or impossible to do with multivariate posteriors.

Let $\theta = (\theta_1, ..., \theta_p)$.

Ideally, we reduced to the univariate marginal posteriors. Then the same ideas for univariate models apply

$$P(\theta_1|Y) = \int ... \int P(\theta_1, ..., \theta_p|Y) d\theta_2, ..., d\theta_p$$

Can use Monte Carlo sampling to estimate these integrals.

Need to confirm the above statement

3.8 Bayesian One Sample t-test

Likelihood: $Y_i|\mu, \sigma \sim N(\mu, \sigma^2)$ idep over $i = 1, ..., n$ Priors: $\mu \sim N(\mu_0, \sigma_0^2)$ independent of $\sigma^2 \sim InvGamma(a, b)$

Typically we are interested in marginal posterior because it accounts for uncertainty about σ^2

Marginal Posterior: $f(\mu|Y) = \int_0^\infty P(\mu, \sigma^2|Y) d\sigma^2,\ Y = (Y_1, ..., Y_n)$

if σ is known, the posterior of $\mu|Y$ is Gaussian and 95% Credible Interval is $E(\mu|Y) \pm Z_{0.975} SD(\mu|Y)$

if σ is unknown, the marginal (over σ^2) posterior of μ is t with $\nu = n+2a$ degrees of freedom.

$$E(\mu|Y) \pm t_{0.975}\ SD(\mu|Y)$$

$SD(\mu|Y)$: Standard Deviation

Can summarize results best in a table with Posterior Mean, Posterior SD, and 95% Credible Set.

3.9 Frequentist vs Bayesian Analysis of a Normal Mean

Frequentist

Estimate of the μ is $\bar{Y}$ If σ is known, the 95% C.I. is: $\bar{Y} \pm z_{0.975} \frac{\sigma}{\sqrt{n}}$
If σ is unknown, the 95% C.I. is: $\bar{Y} \pm t_{0.975,n-1} \frac{s}{\sqrt{n}}$
where t is the quantile of a t-distribution.

Bayesian

Estimate of μ is its marginal posterior mean.

Interval estimate is 95% Credible Interval.

If σ is known, Posterior of $\mu|Y$ is Gaussian

$E(\mu|Y) \pm Z_{0.975} SD(\mu|Y)$

If σ is unknown, the marginal (over σ^2) posterior of μ is t with $\nu = n+2a$ degrees of freedom.

$E(\mu|Y) \pm t_{0.975,\nu}\ SD(\mu|Y)$

3.10 Multiple Parameters in Multivariate Posteriors

Want to compute $P(\theta_2 > \theta_1|Y_1, Y_2)$.

Monte Carlo sampling of the posteriors a key tool!

Model is:

$$\begin{aligned} Y_1|\theta_1 &\sim Bin(N, \theta_1) \\ Y_2|\theta_2 &\sim Bin(N, \theta_2) \\ \theta_1, \theta_2 &\sim Beta(1, 1) \end{aligned} \tag{9}$$

Marginal Posteriors both independent of each other.

- $\theta_1|Y_1, Y_2 \sim Beta(Y_1 + 1, N - Y_1 + 1)$
- $\theta_2|Y_1, Y_2 \sim Beta(Y_2 + 1, N - Y_2 + 1)$

N <- 10; Y1 <- 5; Y2 <- 8;

S <- 10000

theta1 <- rbeta(S, Y1 + 1, N - Y1 + 1) theta2 <- rbeta(S, Y2 + 1, N - Y2 + 1)

(Y1 + 1) / (N + 2)

mean(theta1)

mean(theta2 > theta1)

3.11 Types of Uncertainty

Sampling

Parametric: Uncertainty about my guesses of the distribution of the parameter

3.11.1 Resolving Uncertainty

1. Plugin approach

   If $\hat{\theta}$ is an estimate, thus $Y^* \sim f(Y|\hat{\theta})$

   For example, Let $\hat{\theta} = \frac{2}{10}$. Predict $P(Y > 0) = 1 - (1 - 0.2)^{10}$.

   If $\hat{\theta}$ has small uncertainty, this is fine. Otherwise, this underestimated uncertainty in Y^*

2. Posterior Predictive Distribution (PPD)

   For the sake of prediction, the parameters aren't of interest as the parameters are vehicles by which the data inform about the predictive model.

   PPD averages over their posterior uncertainty which *accounts* for parametric uncertainty.

   $$f(Y^*|Y) = \int f(Y^*|\theta)p(\theta|Y)\ d\theta$$

   Input = `data Output` = prediction distribution

   > Given I've observed a certain amount of data Y, what is the distribution of the predictor values?

   Monte Carlo sampling approximates the PPD.

   (a) Example
   Let $\theta^{(1)}, ..., \theta^{(S)}$ be samples from the posterior.
   Let $Y^{*(s)} \sim f(Y|\theta^{(s)})$ where $Y^{*(s)}$ are samples from the PPD for each $\theta^{(s)}$
   Posterior Predictive Mean $\approx$ sample mean of the $Y^{*(s)}$
   $P(Y^* > 0) \approx$ sample proportion of non-zero $Y^{*(s)}$
   Y < -2; n <- 10;
   A <- Y + 1; B <- N - Y + 1
   1-dbinom(0,10,0.2)
   theta <- rbeta(100000,A,B) Ystar <- rbinom(100000,10,theta) mean(Ystar>0)