From Hypothesis Testing to Bayesian Inference

Boris Babic,
Assistant Professor of Decision Sciences

Plan for Today

- We're going to jump right into a simple problem regarding an unknown proportion of interest: The case of the curious coin.
- Reason about the problem using hypotheses testing procedures.
- Motivate the Bayesian approach, and construct a Bayesian answer to this problem.
- In the classes to follow, we'll slow down substantially and develop the Bayesian approach with care.
- My hope is that while you may not follow every detail today, you will be curious enough to learn more!

  –

- In general, all slides and code notebooks will be posted to the course website before class. I will use Jupyter notebooks.
- For today, the notebook can be found here.
- Reading: Hoff, A First Course in Bayesian Statistics (pgs. 13-35).

Problem

A Curious Coin

You have come across a curious coin. It seems (you suspect) bent in a way that biases it toward landing on heads. You will give this coin to your trusty RA, and ask them to perform an experiment (i.e., toss it repeatedly) in order to help you decide whether the coin is biased.

Hypothesis Testing

A seemingly natural place to start would be to construct a hypothesis test regarding the coin's bias.

Definition

A hypothesis is a statement about a population parameter $\theta \in \Omega$.

In our coin case, θ is the unknown bias of the coin, and $\Omega = [0, 1]$.

Two statements about θ

- Null hypothesis: $H_0 : \theta \in \Omega_0$
- Alternative hypothesis: $H_1 : \theta \in \Omega_0^c$
- $\Omega = \Omega_0 \cup \Omega_0^c$

Hypothesis Testing

Hypothesis testing procedure

A rule that specifies

- For which sample points H_0 will be accepted as true (the subset of the sample space for which H_0 will be accepted is called the acceptance region).
- For which sample points H_0 is rejected and H_1 is accepted as true (the subset of the sample space for which H_0 is rejected is called the rejection region or critical region).

Rejection region

Rejection region (R) on a hypothesis is usually defined through a test statistic $W(X)$. For example,

$$R_1 = \{\boldsymbol{x} : W(X) > c, \boldsymbol{x} \in \mathcal{X}\}$$

$$R_2 = \{\boldsymbol{x} : W(X) \leq c, \boldsymbol{x} \in \mathcal{X}\}$$

Problem

Back to the coin. We might start as follows.

$$X_1, X_2, ..., X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta) \text{ where } n = \text{ (say) } 12.$$

$$H_0 : \theta \leq 0.5$$
$$H_1 : \theta > 0.5$$

Now we propose a hypothesis test.

Test

Reject H_0 if and only if all successes are observed. That is,

$$R = \{\boldsymbol{x} : \boldsymbol{x} = (1,1,1,1,1,1,1,1,1,1,1,1)\}$$
$$= \{\boldsymbol{x} : \sum_{i=1}^{12} x_i = 12\}$$

where R is the rejection region and $W(X) = \sum_{i=1}^{n} X_i$.

Properties of Proposed Test

As a statistician, there are several properties of this test that you would like to investigate. By way of review, we will do this now.

1. Compute the power function.
2. What is the maximum probability of making a Type I error?
3. What is the probability of making a Type II error if $\theta = 2/3$?

Notice that we do not have any data yet. Questions 1-3 can be answered without the data!

Question 1: Power Function

Power function

The power function of a hypothesis test with rejection region R is a function of θ and is defined as

$$\beta(\theta) = \Pr(\boldsymbol{X} \in R|\theta) = \Pr(\text{reject} H_0|\theta)$$

If $\theta \in \Omega_0^c$ (alternative is true), the probability of rejecting H_0 is called the power of the test for this particular value of θ.

Question 1 asks us to compute the power function. For our test,

$$\begin{aligned}\beta(\theta) &= \Pr(\text{reject} H_0|\theta)\\ &= \Pr(\boldsymbol{X} \in R|\theta)\\ &= \Pr(\sum X_i = 12|\theta)\end{aligned}$$

Since $\sum X_i \sim \text{Binomial}(12, \theta)$, $\beta(\theta) = \theta^{12}$.

Questions 2 and 3: Type I and II Errors

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct Decision	Type I error
	H_1	Type II error	Correct Decision

Type I error

If $\theta \in \Omega_0$ (null hypothesis is true), the probability of making a type I error is

$$\Pr(\boldsymbol{X} \in R|\theta)$$

Type II error

If $\theta \in \Omega_0^c$ (alternative hypothesis is true), the probability of making a type II error is

$$\Pr(\boldsymbol{X} \notin R|\theta) = 1 - \Pr(\boldsymbol{X} \in R|\theta)$$

- Power $= \beta(\theta)$ if $\theta \in \Omega_0^c$.
- Probability of Type I error $= \beta(\theta)$ if $\theta \in \Omega_0$.
- Probability of Type II error $1 - \beta(\theta)$ if $\theta \in \Omega_0^c$.
- Ideal test: $\beta(\theta) = 0$ for all $\theta \in \Omega_0$ (no Type I error) and $\beta(\theta) = 1$ for all $\theta \in \Omega_0^c$ (no Type II error).

Questions 2 and 3: Type I and II Errors

Question 2 asks us to compute the maximum probability of making a Type I error.

- When $\theta \in \Omega_0$, the power function $\beta(\theta)$ is Type I error.
- For us, $\Omega_0 = [0, 0.5]$. Thus,

$$\max_{\theta \in \Omega_0} \beta(\theta) = \max_{\theta \in [0,0.5]} \theta^{12} = 0.5^{12} = 0.0002.$$

- This is to be expected. Type I error is where we falsely reject the *null* hypothesis.
- Since our test will only reject it if we observe all successes (heads), we are extremely unlikely to falsely reject the hypothesis that the coin's bias is less than 0.5.

Questions 2 and 3: Type I and II Errors

Question 3 asks us to compute the probability of making a Type II error if $\theta = 2/3$.

- When $\theta \in \Omega_0^c$, Type II error is given by $1 - \beta(\theta)$.
- We need only consider the case where $\theta = 2/3 \in \Omega_0^c$. Thus,

$$1 - \beta(\theta)|_{\theta=\frac{2}{3}} = 1 - \theta^{12}|_{\theta=\frac{2}{3}} = 1 - (2/3)^{12} = 0.99.$$

- This is again to be expected. Type II error is where we falsely reject the alternative hypothesis. The alternative hypothesis is that $\theta > 0.5$, but we are willing to reject it only in the case where we observe *all* successes. If θ were indeed equal to $2/3$ this would be quite unlikely – 0.01 unlikely – to occur.

Finally Some Data!

A Curious Coin

Your RA reports having tossed the coin 12 times, with the following results:

$$H, T, H, H, H, H, H, T, H, H, H, T$$

$$9H, 3T$$

- Under our proposed test, we would not reject $H_0 : \theta \leq 0.5$.
- Our rejection region was simple: reject only if we observe all heads.
- Perhaps this was a bad test after all.
- In Frequentist statistics, we search for tests that are reliable in the sense that we can get guarantees about what would happen if we performed the test many times (eg. Neyman-Pearson Lemma, Karlin-Rubin Theorem).
- For example, our test had a very low maximum Type I error rate.
- We could search for a better test, but is there anything else we can try?
- I know how to compute something called a p-value. Let me through!

A Slightly Different Approach

- Earlier, our strategy was to reject the null hypothesis if the test statistic lies in the rejection region.
- The p-value is defined as the probability of our result, or a more extreme result, under the null hypothesis.
- For Fisher, it was a rough and ready way to get a sense of the weight of evidence. It has since become a rule for evaluating whether one has a statistically significant result.
- So now, we will take a slightly different approach and identify a significance level ex-ante, compute the p-value, and reject the null hypothesis if the p-value is less than the level of significance.
- Conventionally, the significance level is 0.05. This gets (or used to get) a star in the leading journals.
- So let's try this.
- And let's make things as simple as possible and test whether we can reject the null hypothesis that the coin is fair.

A Slightly Different Approach

$$X_1, X_2, ..., X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta) \text{ where } n = 12.$$

$$H_0 : \theta = 0.5$$

Recall that from our experiment,

$$H, T, H, H, H, H, H, T, H, H, H, T$$

$$9H, 3T$$

Need to compute the probability of observing our result, or a more extreme result, under $H_0 : \theta = 0.5$.

A Slightly Different Approach

Let $Y = \sum X_i$. Then

$$Y \sim \text{Binomial}(n, \theta)$$

$$\begin{aligned} P(Y \geq 9 | \theta = 0.5, n = 12) &= \sum_{y=9}^{12} \binom{12}{y_i} \left(\frac{1}{2}\right)^{y_i} \left(\frac{1}{2}\right)^{12-y_i} \\ &= \left[\binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0}\right] \left(\frac{1}{2}\right)^{12} \\ &= \frac{299}{4096} \approx 0.07 \end{aligned}$$

The result is not statistically significant. Again we cannot reject the hypothesis that the coin is fair.

An Unexpected Mixup

- After you report the results back to your RA, you learn there was a mix up!
- You thought you told the RA to toss the coin 12 times.
- But your RA actually tossed it until observing 3 tails.
- As it happens, it took 12 tosses to get 3 tails.
- Should this matter? The evidence is what it is isn't it?
- But now n is random and Y is fixed.
- Thus, a result more extreme than $(9, 3)$ is no longer $(10, 2)$, $(11, 1)$, and $(12, 0)$. Rather, it is $(10, 3)$, $(11, 3)$, $(12, 3)$, and so on.
- We have to re-calculate the p-value. Thoughts on how to do this?

$$\Pr(N = n|\theta, r) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}$$

where r is the number of tails.

One More Try

$$
\begin{aligned}
P(N \geq 12|\theta = 0.5, r = 3) &= \sum_{n=12}^{\infty} \binom{n_i - 1}{r - 1} .5^r .5^{n_i - r} \\
&= \sum_{n=12}^{\infty} \binom{n_i - 1}{2} .5^{n_i} \\
&= 1 - \sum_{n=1}^{11} \binom{n_i - 1}{2} .5^{n_i} \\
&\approx 0.03
\end{aligned}
$$

Now the result *is* statistically significant!

What if the RA stopped tossing the coin so that they can get a coffee?

Or to watch Riverdale on Netflix?

Sometimes there are ethical reasons to stop collecting data
(HIV antiretroviral drug example)

Bayesian Statistics

If you test positive for a certain disease...

You may want to ask

- Do I have the disease or not?
- What is the chance that I have the disease?

Possible answers

- Frequentist: I do not know. You're asking the wrong question. Whether you have the disease or not is not a random variable. It is a fixed value. Therefore, the question does not make sense.
- Bayesian: The chance that you have the disease is ... % (How?)

Bayesian Statistics

Differences from frequentist statistics

- On the Bayesian approach, the parameter θ is considered as a random quantity.
- We describe our uncertainty about θ by a probability distribution, referred to as the prior distribution.
- A sample is taken from a population indexed by θ, and the prior is then updated, using Bayes' Rule, to get a posterior distribution for θ given the sample.
- Inferences are then made from the posterior distribution.

Bayes' Rule

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)}$$

Bayesian Framework

- Prior distribution for θ:
$$\theta \sim \pi(\theta)$$
- Sample distribution (or likelihood) of $\boldsymbol{X}$ given θ:
$$\boldsymbol{X}|\theta \sim f(\boldsymbol{x}|\theta)$$
- Joint distribution of $\boldsymbol{X}$ and θ:
$$f(x,\theta) = f(\boldsymbol{x}|\theta)\pi(\theta)$$
- Marginal distribution of $\boldsymbol{X}$:
$$m(\boldsymbol{x}) = \int_{\theta\in\Omega} f(\boldsymbol{x},\theta)d\theta = \int_{\theta\in\Omega} f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$$
- Posterior distribution of θ (conditional distribution of θ given $\boldsymbol{X}$):
$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x},\theta)}{m(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})} \propto f(\boldsymbol{x}|\theta)\pi(\theta) \qquad \text{(Bayes' Rule)}$$

Bayesian Approach to the Coin Problem

- In our problem, we know that the likelihood of $\boldsymbol{X}$ given θ is Bernoulli in θ:

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \{\theta^{x_i}(1-\theta)^{1-x_i}\}$$

- Now we need to identify a prior distribution for θ.
- A flexible prior distribution for the unknown parameter of a Bernoulli process is a beta distribution with parameters α and β:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

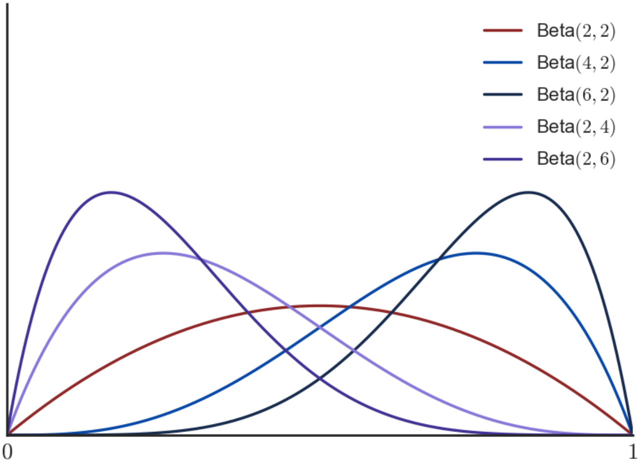
where $\Gamma(x)$ is the complete Gamma function, $\int_0^\infty t^{x-1}e^{-t}dt$ and for positive integers n, $\Gamma(n) = (n-1)!$.

- Note that

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Note also that $\Gamma(\alpha+\beta)/[\Gamma(\alpha)\Gamma(\beta)]$ is not a function of θ. It is the normalizing constant for this distribution. Often we can ignore it and renormalize after updating.

Bayesian Approach to the Coin Problem



Bayesian Approach to the Coin Problem

- The posterior distribution for θ is

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto f_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\pi(\theta)\\ &= \prod_{i=1}^{n}\{\theta^{x_i}(1-\theta)^{1-x_i}\}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &\propto \prod_{i=1}^{n}\{\theta^{x_i}(1-\theta)^{1-x_i}\}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &= \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}\end{aligned}$$

- Using the fact from the previous slide, we know that

$$\begin{aligned}&\int_0^1 \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}d\theta\\ &= \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma\left(\left[\sum_{i=1}^n x_i+\alpha\right]+\left[n-\sum_{i=1}^n x_i+\beta\right]\right)}\end{aligned}$$

Bayesian Approach to the Coin Problem

- Let $\alpha^* = \sum_{i=1}^n x_i + \alpha$. Let $\beta^* = n - \sum_{i=1}^n x_i + \beta$. Then our posterior distribution for θ is

$$\pi(\theta|\boldsymbol{x}) = \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}$$

- In other words, the posterior distribution is still of the beta form, except that our new α corresponds to the initial α plus the number of successes/heads and the new β corresponds to the initial β plus the number of failures/tails.
- This lends itself to a natural interpretation: The initial α value corresponds to the number of pseudo tosses that came up heads, whereas the initial β value corresponds to the number of pseudo tosses that came up tails.
- Bayesian updating is easily accomplished by adding the pseudo heads to the observed heads and pseudo tails to observed tails.

Bayesian Approach to the Coin Problem

Conjugate family

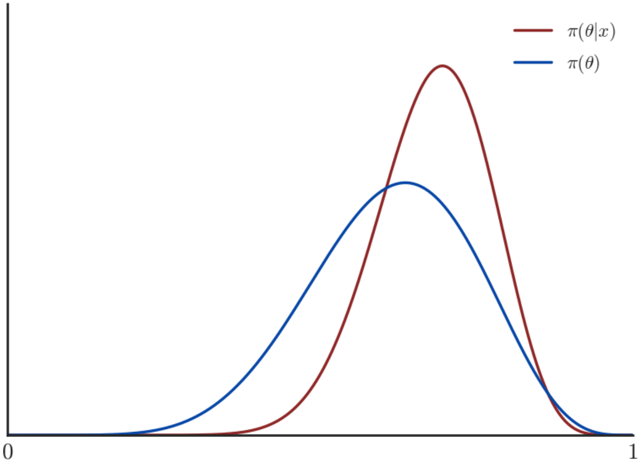
Let $\mathcal{F}$ denote the class of distributions for $f(x|\theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$ if the posterior distribution is the class Π for all $f \in \mathcal{F}$ and all priors $\pi \in \Pi$, and all $x \in \mathcal{X}$.

- What we have seen so far is that the beta distribution is conjugate to the Bernoulli process. This is sometimes called the "beta-binomial" family.
- In the classes to follow, we will look at other commonly used conjugate families of distributions.

Bayesian Approach to the Coin Problem

- Now we can tackle our problem.
- The coin looked bent in a way that made it biased toward heads.
- What are reasonable values for α and β?
- Suppose $\alpha = 8$ and $\beta = 5$.
- The posterior distribution is beta with $\alpha = 9 + 8 = 17$ and $\beta = 3 + 5 = 8$.

Posterior Distribution



Inferences on θ

- Any statements that we wish to make about θ can be easily computed from the posterior distribution.
- The posterior distribution describes all our beliefs about θ after viewing the data.
- For example, we way want to make a point estimate using the posterior mean.

  This is given by $\alpha/(\alpha+\beta)$.

  Before seeing the data, this was $8/(8+5) \approx 0.61$.

  After seeing the data, this is $17/(17+8) \approx 0.68$.

  Note that the sample mean is 0.75. The data has nudged our prior toward a stronger belief in the coin's bias toward heads.
- We may also want the mode, which is the value we think most likely. This is $(\alpha-1)/(\alpha+\beta-2) = (17-1)/(17+8-2) \approx 0.69$.

Bayesian Hypothesis Tests

- Recall that what we really wanted to know was a simple question: is the coin biased toward heads?
- We used frequentist procedures to set up a hypothesis test which helps us evaluate this. Now we can answer it directly:

$$\begin{aligned}\Pr(\theta > 0.5) &= \int_{0.5}^{1} \pi(\theta|\boldsymbol{x})d\theta \\ &= 1 - CDF(\theta|\boldsymbol{x})|_{\theta=0.5} \\ &= 1 - 0.03 \\ &= 0.97\end{aligned}$$

- R code: `1 - pbeta(0.5, 17, 8)`
- We are 97% confident that the coin is biased toward heads.
- We now have an answer to a one-sided hypothesis test:

$$H_0 : \theta \leq 0.5 \qquad\qquad H_1 : \theta > 0.5$$

- But instead of accepting/rejecting the null hypothesis, we make probabilistic statements from the posterior distribution.

Bayesian Two-Sided Hypothesis Tests

- But what what if we want to know whether the coin is fair or not? That is,

$$H_0 : \theta = 0.5 \qquad\qquad H_1 : \theta \neq 0.5$$

- On the picture developed so far, we cannot do this.
- The probability that θ takes on any specific value is 0. Thus the posterior probability for any such H_0 will be 0.
- We will see how to make binary decisions in the Bayesian framework once we introduce the notions of loss and Bayes risk.

Credible Intervals

- However, we can calculate a $(1-\alpha)100\%$ credible interval for θ. For example, a 95% credible interval for θ is,

$$\Pr(a < \theta < b) = \int_a^b \pi(\theta|\boldsymbol{x})d\theta = 0.95$$

- In our case, $a = 0.49$ and $b = 0.84$.
- R code: `qbeta(c(0.025,0.975),17,8)`
- We can also compute the probability that θ is in any desired region of the posterior distribution. This gives us a probabilistic statement about a small region around a point null hypothesis. For example:

$$\begin{aligned}\Pr(0.4 < \theta < 0.6) &= \int_{0.4}^{0.6} \pi(\theta|\boldsymbol{x})d\theta \\ &= CDF(\theta|\boldsymbol{x})|_{\theta=0.6} - CDF(\theta|\boldsymbol{x})|_{\theta=0.4} \\ &= 0.19\end{aligned}$$

- R code: `pbeta(0.6, 17, 8) - pbeta(0.4, 17, 8)`.
- We are about 20% confident that θ is between 0.4 and 0.6.

Takeaway

- We can give a direct answer to most questions of interest regarding θ.

- Our inference is not sensitive to the reason our RA stopped experimenting.
- However, it is sensitive to the choice of prior.
- In the classes to follow, we will talk more about how to identify reasonable priors, how to construct sensible models, and how to evaluate and refine them with use.
- See you next class!