Bayesian Statistics:

Countings for Nerds

Contrived Example: COVID-19 Infection Rate

- We are a statistician tracking COVID-19 in Philadelphia county
- We have convinced LabCorp to send out 1,000 tests to random addresses in the county
- LabCorp will process tests as they are received, and release updated results every day for the next week (5 business days)

Simplifying Assumptions

- The tests are perfect!
- Every person who receives a test, completes it and sends it back on the same day
- The infection rate is the same across the entire county

COVID-19 Infection Rate: Warm Up

Day	# Tests	# Positive
1	200	2
2	150	2
3	300	1
4	250	3
5	100	2

1. After day 1, what is the estimated infection rate?

$$\frac{2}{200} = 0.01$$

2. Considering only data from day 3 what is the estimated infection rate?

$$\frac{1}{300} = 0.0\overline{33}$$

3. After all 5 days, what is the estimated infection rate?

$$\frac{10}{1000} = 0.01$$

Agenda

1. Classical statistics review
2. Bayesian approach
3. Binomial Example

Classical Statistics: Goals

1. Point Estimation -- Single best value for an unknown parameter
2. Variability Estimation -- Estimate an interval of likely values for an unknown parameter
3. Hypothesis Testing -- Probability of a specific hypothesis
4. Prediction of a future event or unobserved data

Classical Statistics: Approach

Construct a model to estimate the unknown parameters using available data

Model

Links the observed data to the unknown parameters

Unknown Parameters

- Parameters are **fixed** values
- There is some true value of the parameters, but these values are unknown
- The true parameters are what generated the observed data (data-generating process)

Estimation

- Find the single best estimate of the unknown parameters
- Conditional on the model and observed data

Classical Statistics: Estimation

Find the single best estimate of the unknown parameters, given a model and observed data

Likelihood Principle -- Our model is based on a probability density function $p(y|\theta)$

If we have observations y_1, y_2, …, y_n, then we can write the likelihood of the observed data given this model as:

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

Maximum Likelihood Estimation -- We want to identify the parameter value that makes our observed data as likely as possible

$$\hat{\theta}_{MLE} = argmax_{\theta} p(y|\theta)$$

Classical Statistics: Dealing with Uncertainty

All point estimates are based on $\hat{\theta}_{MLE}$ but we recognize that our observed data was just one of many possible samples.

How much does our point estimate change from sample to sample?

- Leads to the concept of sampling distributions, which are used for constructing confidence intervals (variability estimation) and hypothesis testing

Alternatively:

- Rely on asymptotics (Central Limit Theorem)
- Bootstrap procedures

Classical Statistics: COVID-19 Infection Rate

θ -- Unknown infection rate

$y \sim Binomial(n, \theta)$

Day	# Tests	# Positive
1	200	2
2	150	2
3	300	1
4	250	3
5	100	2

1. If you had to guess a single value for the IR, what would you choose?

$$\hat{\theta}_{MLE} = \frac{\#test_+}{n} = \frac{10}{1000} = 1\%$$

2. If we repeated this experiment many times, what interval would capture the true value 95% of the time?

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = [0.00998, 0.01002]$$

3. What is the probability that the IR is less than 1%

Classical Statistics: Drawbacks

1. Cannot make direct probability statements about the unknown parameters

   A 95% CI does not mean that the probability of the unknown parameter being in the CI is 95%. For any particular sample, the parameter either is or is not (binary) in the interval. There is no randomness involved because unknown parameters are fixed values.

2. Sampling distributions can be difficult to calculate

   a. Asymptotic results do not hold in small samples

   b. Bootstrap procedures also fail in small samples

Bayesian Statistics: Fundamentals

Instead of being fixed values, the unknown parameters are considered random variables, and therefore each one has its own probability distribution

Consequences

1. We can make probability statements about the unknown parameters

   This leads to more direct and natural interpretations than classical p-values

2. Point estimation is less important

   If we can compute an entire distribution, then why focus on a single value?

3. Less reliant on asymptotic results

   If we have small sample sizes, the posterior distribution will have larger variance.

Bayesian Statistics: Estimation

Given some data, how do we estimate the distribution of an unknown parameter?

- We continue to assume a model, which is encapsulated by the **likelihood function**
- Parameters are now random, so we need an assumed **(prior) distribution** for them

If we have these two pieces (likelihood and a prior), we can use Bayes' Rule to calculate the **posterior distribution** of the unknown parameters given the observed data

$$\boxed{p(\theta|y)} = \frac{\boxed{p(y|\theta)} \cdot \boxed{p(\theta)}}{p(y)}$$

$$p(\theta|y) \propto p(y|\theta) \cdot p(\theta)$$

Bayesian Statistics: Drawbacks

1. Additional assumptions are required
    - We have to specify a prior distribution in addition to the likelihood
        - Can be a good thing in cases where prior information exists
        - With large data, the likelihood dominates the prior
        - Danger zone is when there is little data and little prior information
2. Estimating a full posterior distribution is more difficult than MLE
    - Even for relatively simple models, the algebra sucks
    - More complicated models require good computational techniques

Binomial Model: Setup

What proportion of the population has been infected with COVID-19?

Assumptions

1. The test is perfect
2. Population infection rate follows a Binomial distribution
3. Prior beliefs about the population infection rate (theta) follows a Beta distribution

<u>Model</u>

$$y \sim Binomial(n, \theta)$$

$$p(\theta) \sim Beta(\alpha, \beta)$$

Binomial Model: Some Algebra

Bayes’ Rule

$$\boxed{p(\theta|y)} \propto \boxed{p(y|\theta)} \cdot \boxed{p(\theta)}$$

Substitution

$$p(\theta|y) \propto \boxed{\binom{n}{y}\theta^{y}(1-\theta)^{n-y}} \cdot \boxed{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}$$

Cancellation

$$p(\theta|y) \propto \boxed{\theta^{y}}\boxed{(1-\theta)^{n-y}} \cdot \boxed{\theta^{\alpha-1}}\boxed{(1-\theta)^{\beta-1}}$$

Combine like terms

$$p(\theta|y) \propto \theta^{y+\alpha-1} \cdot (1-\theta)^{n-y+\beta-1}$$

Binomial Distribution

$$\boxed{p(y|\theta) = \binom{n}{y}\theta^{y} \cdot (1-\theta)^{n-y}}$$

Beta Distribution

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

Binomial Model: Results

$$p(\theta|y) \sim Beta(y+\alpha, n-y+\beta)$$

Given a Beta prior, the posterior distribution also follows a Beta distribution!

This is called conjugacy and the prior distribution is called a conjugate prior. These priors have a few really nice properties.

1. Conjugate priors result in posteriors that follow standard distributions (beta in this case)
2. The hyperparameters (alpha and beta) of conjugate priors can be interpreted as pseudo-observations.
    a. alpha -- number of prior successes (positive test results)
    b. beta -- number of prior failures (negative test results)

Binomial Model: Prior Choice

How do we choose values for alpha and beta?

$$p(\theta|y) \sim Beta(y + \alpha, n - y + \beta)$$

$$E[\theta|y] = \frac{y + \alpha}{n + \alpha + \beta}$$

Noninformative Prior

- Alpha = 0 and Beta = 0
  - Improper prior distribution, but posterior is proper so long as n, y > 0
  - Posterior mean is equivalent to the maximum likelihood estimate!
- Alpha = 1, Beta = 1
  - Proper prior distribution and proper posterior distribution

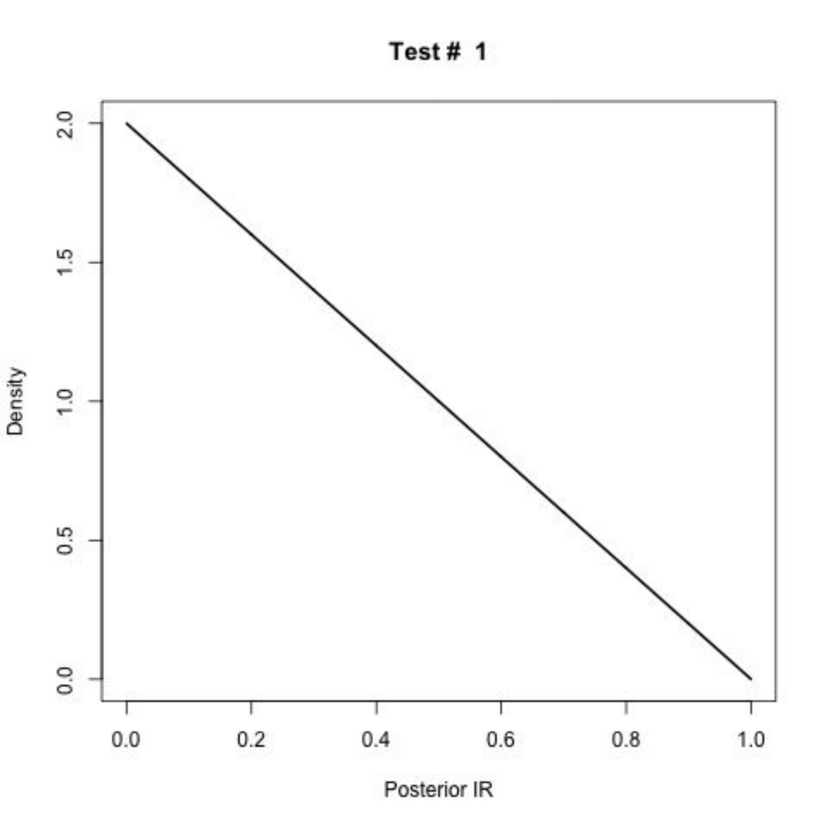
Informative Prior

- Great if we have actual prior information, e.g. prior test results (previous days, similar areas, etc.)
- Set alpha and beta based on that prior information

Binomial Model: Simulation

- Simulate 100 flips of an imbalanced coin (10% chance of heads)

- After each flip, what does the posterior distribution look like?



Bayesian Statistics: COVID-19 Infection Rate

Day	# Tests	# Positive
1	200	2
2	150	2
3	300	1
4	250	3
5	100	2

$$p(\theta|y) \sim Beta(y+\alpha, n-y+\beta)$$

$$E[\theta|y] = \frac{y+\alpha}{n+\alpha+\beta}$$

1. Given a noninformative prior (alpha = beta = 0), what is the posterior mean of the IR after day 1?

$$E[\theta|y] = \frac{2}{200} = .01$$

2. What if we use a flat prior (alpha = beta = 1)?

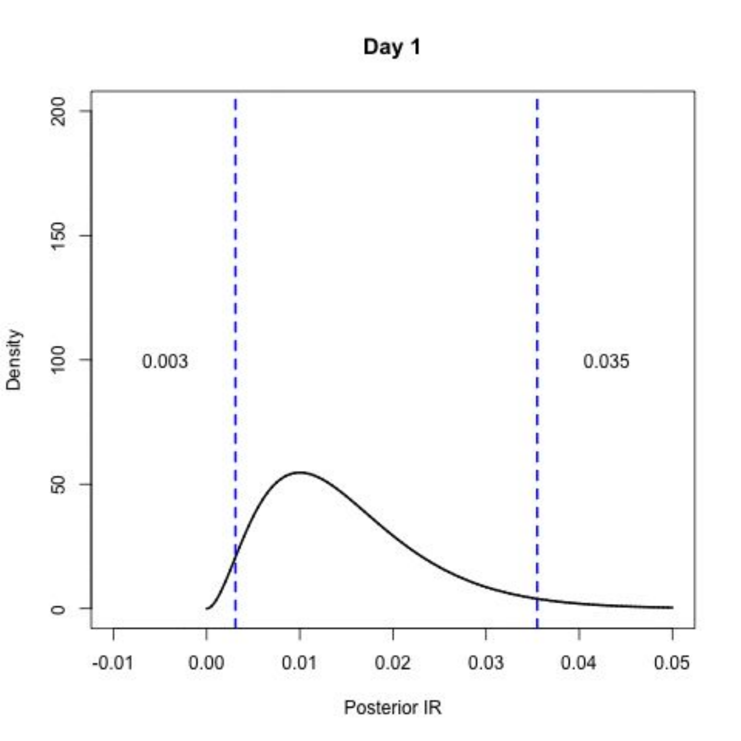
$$E[\theta|y] = \frac{3}{201} = .0149$$

Bayesian Statistics: COVID-19 Infection Rate

1. What is the posterior mean after each day?

2. What interval captures 95% of the posterior probability after each day?

3. What is the probability that the IR is less than 1% after each day?



Concluding Remarks

1. Be careful with p-values and confidence intervals: they are surprisingly convoluted

    It’s probably safest not to say “probability” at all when talking about them

2. No silver bullets!

    Bayesian statistics solves some of the classical problems, but introduces its own

3. Using “subjective” to dismiss the Bayesian approach is a red herring argument

    Most classical procedures have Bayesian analogs (usually with flat priors)

4. We are all (at Blackfynn at least) already Bayesians

    Critical thinking is basically a Bayesian process… Learn, evolve, and adapt

Thank You!

If you want to learn more….

Books

- Statistical Rethinking
- Bayesian Data Analysis
- Course Materials (slack me for access)

Software

- Stan
- PyMC3