Statistical Probability and Models

C. Ryan Campbell

Duke University

c.ryan.campbell@duke.edu

30 Nov 2017

Project Update

- Tuesday - Group AlanAustinRaymond & ChrisJenniferNayib
- Thursday - Everyone Else
- Present what you have, should be more polished than the rough draft
- 20 minutes per group presentation
- 5min intro, 5min per member

Rough Guidelines

- Intro group's dataset, everyone in your group should participate - 5 mins
- Dicuss your question and results - 5 mins

  Suggested order:
- Question/Hypothesis - what did you ask and what did you expect
- Methods - how you asked this question?
- Results - statistical results and plots
- Conclusion - was your hypothesis correct?
- Future Directions/Improvements - what could be done or changed?

Overview

1. Probability
   - And versus Or
2. Models
   - Maximum Likelihood
   - Bayes Rule
3. ML versus Bayes Rule: An Example

This Week's Goals

- Understand p-values & adjustments
- Learn about Maximum Likelihood
- Learn about Bayes Theorem & Bayesian Probability

Probability

- "Odds" that some event occurs
- Bounded from 0 to 1
- Usually expressed as a fraction or percent
- Often using the notation: Pr(event) or P(event)

Or

- Probabilities of multiple events can be combined
- "Or" condition
- Probability either thing happens: A or B
- when A and B are independent and mutually exclusive:
- $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$

Or

- Probabilities of multiple events can be combined
- "Or" condition
- when A and B are independent and
  <u>not exclusive</u>:
- Pr(A or B) = Pr(A) + Pr(B) - Pr(A & B)

And

- Probabilities of multiple events both occuring can be combined
- "And" condition
- Probability both things happen, A & B
- When A & B are independent:
- Pr(A & B) = Pr(A) * Pr(B)
- "And" is commutative:
- Pr(B & A) = Pr(A & B)

Conditional Probabilities

- Probabilities of event A given event B
- Probability of A if we know B has occured
- When B happens, how likely is it that A happens
- Numerator = Pr(A & B)
- Denominator = Pr(B)

$$\frac{\Pr(A \,\&\, B)}{\Pr(B)}$$

Models

- What are models?
- Take in parameters, predict data
- If a die is fair, how many times in 10 rolls should we see a 2?
- BIO EXAMPLE

Models

- What is an example of a model?
- Mathematical equations that predict experimental results
- Dice are weighted equally, each side should have a 1/6 chance of coming up
- BIO EXAMPLE

Models

- Why are models useful?
- Allow us to predict results, test those predictions, and put mathematical estimates to natural processes
- Dice example - good for teaching, winning at craps
- BIO EXAMPLE -

Parameter Estimation

- Maximum Likelihood
- Bayesian Estimation

Maximum Likelihood

- Parameters for which the data are most likely are correct
- Vary the parameters and check how likely the data are
- "Plug & Chug"

Conditional Probabilities

- Probabilities of data given parameters/model
- With these parameters, how likely is this data?
- Numerator = Pr(params & data)
- Denominator = Pr(data)

$$\frac{\text{Pr(params \& data)}}{\text{Pr(data)}}$$

Maximum Likelihood

- Likelihoods aren't meaningful alone
- Only used relative to other likelihoods from the same model
- Greatest likelihood is accepted as the parameter's value

Maximum Likelihood: Toy Example

- I flipped a coin 10 times
- 6 Heads & 4 Tails
- Model - Binary probability distribution & Pr(head) = 1 - Pr(tails)
- What is the true probability of it landing heads?

Maximum Likelihood: Toy Example

- Binary Probabilty Distribution

$$\Pr(k \text{ heads and } n\text{-}k \text{ tails}) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

- Which are data and which are parameters (of n, k & p)?

Maximum Likelihood: Toy Example

- What is the probability of seeing this data (6H 4T):
- If the coin is 50-50?
- If the coin is 75-25?

Maximum Likelihood: Toy Example

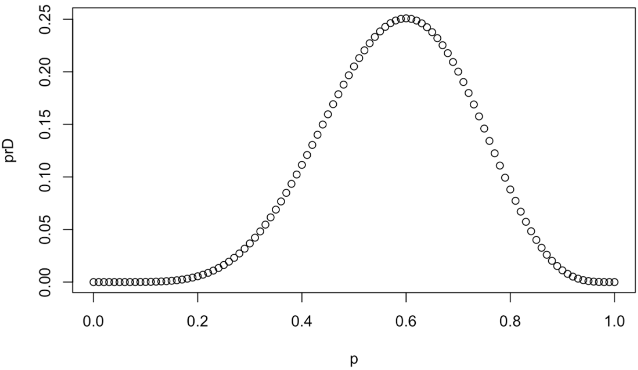
```
n <- 10
k <- 6
p <- .5

prD <- (factorial(n) / (factorial(k) * factorial(n - k))) * p^k
* (1 - p)^(n - k)
```

- What is the probability of the coin being 50-50?
- What about 75-25?
- Can you plot the results of many values of p?
  hint: use `seq()`

Maximum Likelihood: Toy Example

```
p <- seq(0,1,.01)
prD <- (factorial(n) / (factorial(k) * factorial(n - k))) * p^k
* (1 - p)^(n - k)
plot(p,prD)
```



Maximum Likelihood

- Strictly following the Maximum Likelihood to estimate the parameter:
- What is the true probability of the coin landing on heads?
- Does this seem accurate?
- Why is this a misleading example?

Maximum Likelihood: Bio Example

- Normally used for complicated parameters and datasets
- Estimating mutation rates or branch lengths on a phylogeny
- What models would be useful for RNAseq?

Maximum Likelihood: Bio Example

- What models would be useful for RNAseq?
- Inferring isoform expression levels:

  (given a proprotion of reads mapping to exons, how many of each isoform are present)
- Gene pathway interactions:

  (given a set of expression across the genes in a gene pathway, are they being activated together?)

Conditional Probabilities

- Probabilities of data given parameters/model
- With these parameters, how likely is this data?
- Numerator = Pr(params & data)
- Denominator = Pr(data)

$$\frac{\text{Pr(params \& data)}}{\text{Pr(data)}}$$

Deriving Bayes Rule

Definitions Conditional probability:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} \tag{1}$$

Commutative property of probability sets:

$$Pr(A \cap B) = Pr(B \cap A) \tag{2}$$

Beginning with the probability of B given A:

$$Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)} \tag{3}$$

$$Pr(B|A) * Pr(A) = Pr(B \cap A) \tag{4}$$

Substituting from Equation 2:

$$Pr(B|A) * Pr(A) = Pr(A \cap B) \tag{5}$$

Finally, substituting into Equation 1 give Bayes' Theorem:

$$Pr(A|B) = \frac{Pr(B|A) * Pr(A)}{Pr(B)} \tag{6}$$

Bayes Rule

$$Pr(H_1|D) = \frac{Pr(D|H_1) * Pr(H_1)}{Pr(D)}$$

- Likelihood
- Prior
- Posterior

Bayes Rule

- Likelihood - probability of the data given the hypothesis
- Prior - probability of the hypothesis
- Posterior - probability of the hypothesis given the data (what we want!)

Bayes Rule

$$Pr(H_1|D) = \frac{Pr(D|H_1) * Pr(H_1)}{Pr(D)}$$

- Likelihood
- Prior
- Posterior

Bayesian: Toy Example

- I flipped a coin 10 times
- 6 Heads & 4 Tails
- Model - Binary probability distribution & Pr(head) = 1 - Pr(tails)
- What is the true probability of it landing heads?

Bayesian: Toy Example

- Binary Probabilty Distribution

$$\Pr(k \text{ heads and } n\text{-}k \text{ tails}) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

- Which are data and which are parameters (of n, k & p)?

Bayesian: Toy Example

- What is the prior probability of the parameters?
- The coin is 50-50?
- The coin is 75-25?

Bayesian: Toy Example

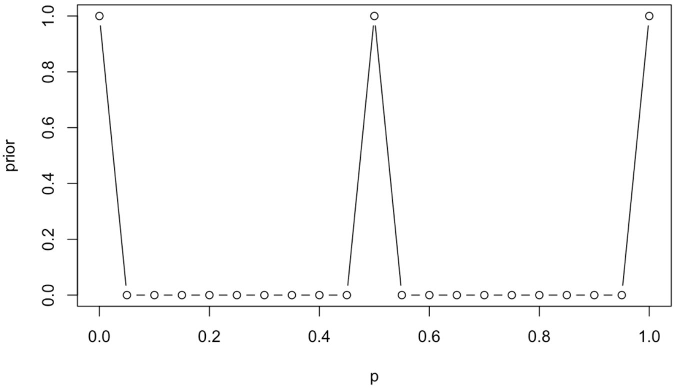
```
n <- 10
k <- 6
p <- .5

prD <- (factorial(n) / (factorial(k) * factorial(n - k))) * p^k
* (1 - p)^(n - k)
```

- How do we add the prior in?

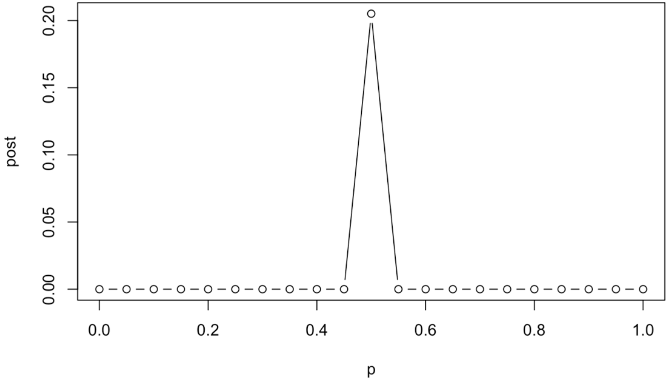
Bayesian: Toy Example

- Prior Distribution



Bayesian: Toy Example

- Posterior Distribution



Bayesian: Toy Example

$$Pr(H_1|D) = \frac{Pr(D|H_1) * Pr(H_1)}{Pr(D)}$$

- Likelihood
- Prior
- Posterior

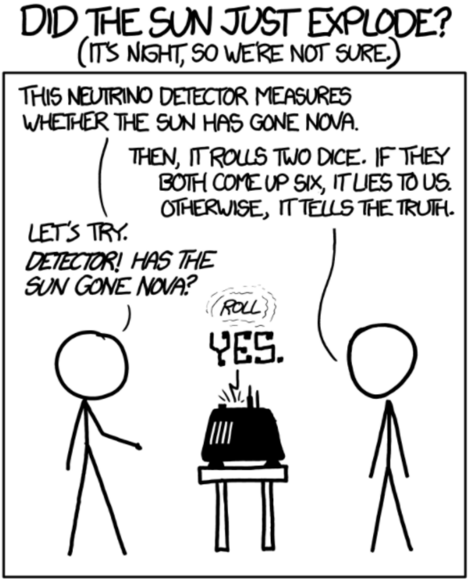
Bayesian Parameter Estimation

- Following the Bayesian Estimation:
- What is the true probability of the coin landing on heads?
- Does this seem accurate?

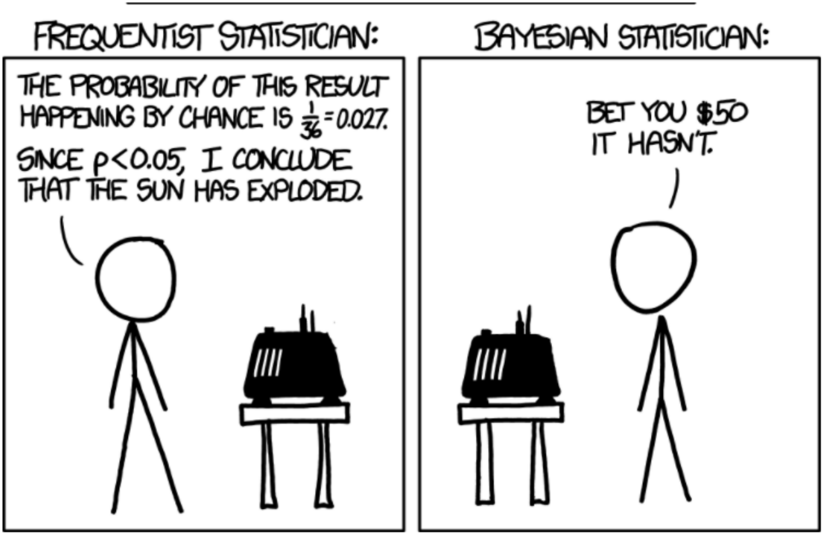
Bayes vs. ML

- What is the main difference?
- What problems might come along with a prior?
- What are biologically relevant problems that this method solves?

More XKCD


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.

More XKCD


FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

The End