Likelihood

Review: Conditional Probability:

P[we see an elephant today| we are in the Serengeti]

P[we see an elephant today| we are in Manhattan]

OR

Back to the coin toss world: if our coin was fair,

i.e. P(heads) = 0.5 or unfair, i.e. P(heads) =0.1

P[12 success in *n* trials|p=0.5]

P[12 success in *n* trials|p=0.1]

Review: Conditional Probability:

The best parameter value is the one that fits the data the best

P[we see an elephant today | we are in the Serengeti]

P[we see an elephant today | we are in Manhattan]

OR

P[12 success in *n* trials | p=0.5]

P[12 success in *n* trials | p=0.1]

parameters(Θ)

Likelihood allows us to take into account the evidence presented….we have learned about updating previously…

Likelihood:

- Conditional probability
- Components of Bayes':

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]}$$

If **A** = parameter set, **Θ**

And **B** = observed data, **y**

Then we can re-write the above formula as:

$$P[\Theta|y] = \frac{P[y|\Theta]P[\Theta]}{P[y]}$$

Likelihood:

- Conditional probability
- Components of Bayes':

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]}$$

Likelihood

Prior

Data

If **A** = parameter set, **Θ**

And **B** = observed data, **y**

Then we can re-write the above formula as:

$$P[\Theta|y] = \frac{P[y|\Theta]P[\Theta]}{P[y]}$$

Likelihood:

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]}$$

Where **A** = parameter set, **Θ** and **B** = observed data, **y**

Then we can re-write the above formula as:

$$P[\Theta|y] = \frac{P[y|\Theta]P[\Theta]}{P[y]}$$

Re-writing hypothesis parameters and data:

$$P[H_i|Data] = \frac{P[H_i \text{ and } Data]}{P[Data]} = \frac{P[Data|H_i]P[H_i]}{P[Data]}$$

You can use this to test ratios of likelihood:

$$\frac{P[H_1|Data]}{P[H_2|Data]} = \frac{P[Data|H_1]\boxed{P[H_1]}}{P[Data|H_2]\boxed{P[H_2]}}$$

Prior

If there are 'n' independent observations (sites on a sequence, for instance):

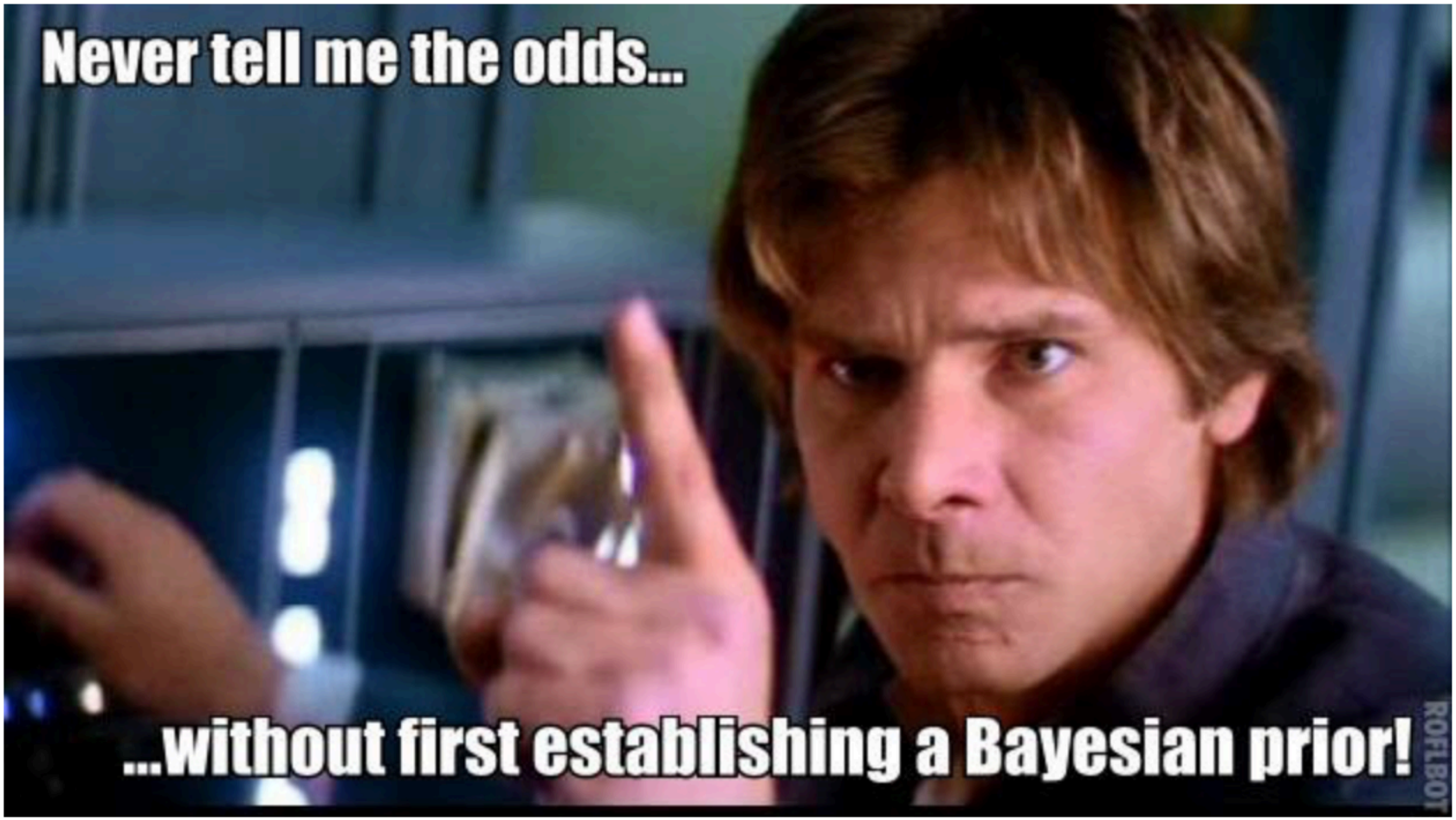
$$\frac{P(\text{Data}\mid H_1)}{P(\text{Data}\mid H_2)} = \frac{P(\text{Data}^{(1)}\mid H_1)\cdot P(\text{Data}^{(2)}\mid H_1)\cdot P(\text{Data}^{(3)}\mid H_1)\cdots P(\text{Data}^{(n)}\mid H_1)}{P(\text{Data}^{(1)}\mid H_2)\cdots P(\text{Data}^{(n)}\mid H_2)}$$

$$\therefore \frac{P[H_1\mid \text{Data}]}{P[H_2\mid \text{Data}]} = \underbrace{\left[\prod_{i=1}^{n} \frac{P(\text{Data}^{(i)}\mid H_1)}{P(\text{Data}^{(i)}\mid H_2)}\right]}_{\text{when there is a LARGE amount of data, the 1st term (likelihood) dominates so prior doesn't influence results}} \cdot \underbrace{\frac{P(H_1)}{P(H_2)}}_{\text{Bayes' Prior}}$$

“Never tell me the odds!”

Star Wars fans:

http://www.countbayesie.com/blog/2015/2/18/hans-solo-and-bayesian-priors



Likelihood:

- Quantifies data support for a **particular value of a parameter (**considers many possible hypotheses, not just one)

- Probability of obtaining observed data if parameter(s) were equal to that specific value
  - Compare to alternate values of the parameter
  - High for values close to the true population parameter and low for values far from it
  - **Maximum likelihood estimate:** best estimate for parameter

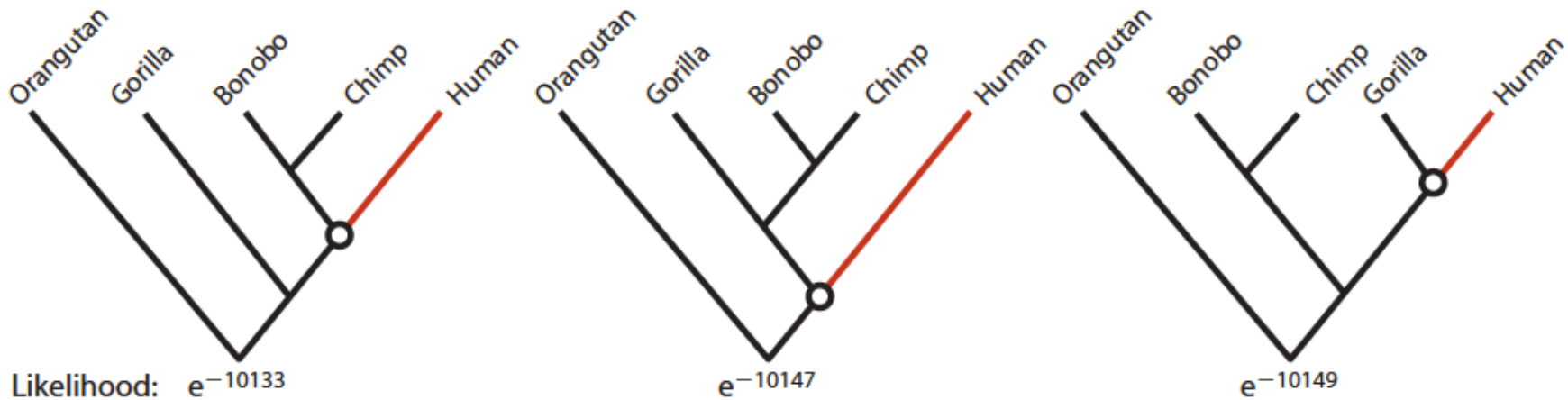
$$L(\theta, \text{data}) = L(\theta \mid \text{data}) = P(\text{data} \mid \theta = \theta_0)$$

Does not assume normally distributed data

Major Applications:

1. **Phylogeny estimation**

L(Tree=i|Sequence Data) = P(Sequence Data|Tree=i)

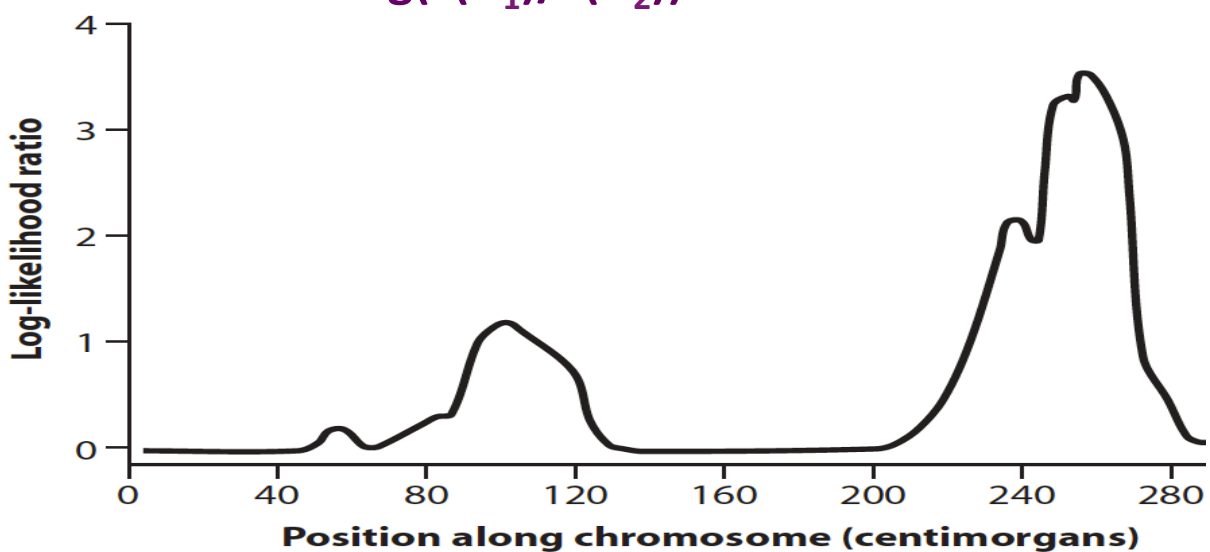


Major Applications:

2. Gene mapping

- "markers" differ between people with disease and those without
- Comparing two hypotheses:
  - H_1: disease gene is present at a particular site
  - H_2: no disease gene is present at that site
  - Review frequency data in a sliding window along the genome
  - LOD score: log of the ratio of these two hypotheses

LOD score = $\log(L(H_1)/L(H_2))$



Major Applications:

2. Gene mapping

- H_1: disease gene is present at a particular site
- H_2: no disease gene is present at that site
- Review frequency data in a sliding window along the genome
- LOD score: log of the ratio of these two hypotheses

LOD score = $\log(L(H_1)/L(H_2))$

This can be thought of as an 'odds ratio' in favour of hypothesis 1 over hypothesis 2

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)\,P(H_1)}{P(D|H_2)\,P(H_2)}$$

Maximum Likelihood Estimation:

- All statistical estimates that we have encountered so far are maximum likelihood estimates
  - Excellent properties such as consistency and precision as data → ∞

- **Requires a probability model**
  - Ex. Mutation models in DNA sequence changes
  - Ex. Binomial distribution for proportions

- Determines the parameter value that yields the largest L.
  - A particular data set supports one hypothesis better than another if the likelihood of that hypothesis is higher than the likelihood of the other hypothesis

- **Very versatile**

Maximum Likelihood Estimation:

Example: 11 coin tosses with a supposedly fair coin

HHTTHTHHTTT

What is the likelihood of this sequence?

Assume coin tosses are **independent**.

Maximum Likelihood Estimation:

Example: 11 coin tosses with a supposedly fair coin

HHTTHTHHTTT

What is the likelihood of this sequence?

Assume coin tosses are **independent**.

$$L(p=0.5|HHTTHTHHTTT) = P(HHTTHTHHTTT|p=0.5)$$
$$= pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p)$$
$$= p^5(1-p)^6$$

Maximum Likelihood Estimation:

Example: 11 coin tosses with a supposedly fair coin

HHTTHTHHTTT

What is the likelihood of this sequence?

Assume coin tosses are **independent**.

$L(p=0.5|HHTTHTHHTTT) = P(HHTTHTHHTTT|p=0.5)$

$= pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p)$

$= p^5(1-p)^6$

To determine Max. likelihood you are going to need to use some simple **calculus**…followed by some **algebra**

Maximum Likelihood Estimation:

Example: 11 coin tosses with a supposedly fair coin

HHTTHTHHTTT

What is the likelihood of this sequence?

Assume coin tosses are **independent**.

$L(p=0.5|HHTTHTHHTTT) = P(HHTTHTHHTTT|p=0.5)$

$= pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p)$

$= p^5(1-p)^6$

---

1. $\frac{dL}{dp} = 5p^4(1-p)^6 - 6p^5(1-p)^5$
2. $\frac{dL}{dp} = 5p^4(1-p)^6 - 6p^5(1-p)^5 = 0$

$= p^4(1-p)^5[5(1-p)-6p]=0$

$$\hat{p} = 5/11$$

Maximum Likelihood Estimation:

Example: 11 coin tosses with a supposedly fair coin

HHTTHTHHTTT

What is the likelihood of this sequence?

Assume coin tosses are **independent**.

$L(p=0.5|HHTTHTHHTTT) = P(HHTTHTHHTTT|p=0.5)$

$= pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p)$

$= p^5(1-p)^6$

$$\ln(L)=5\ln p+6\ln(1-p)$$

$$d(\ln L) = \frac{5}{p} - \frac{6}{(1-p)} = 0$$

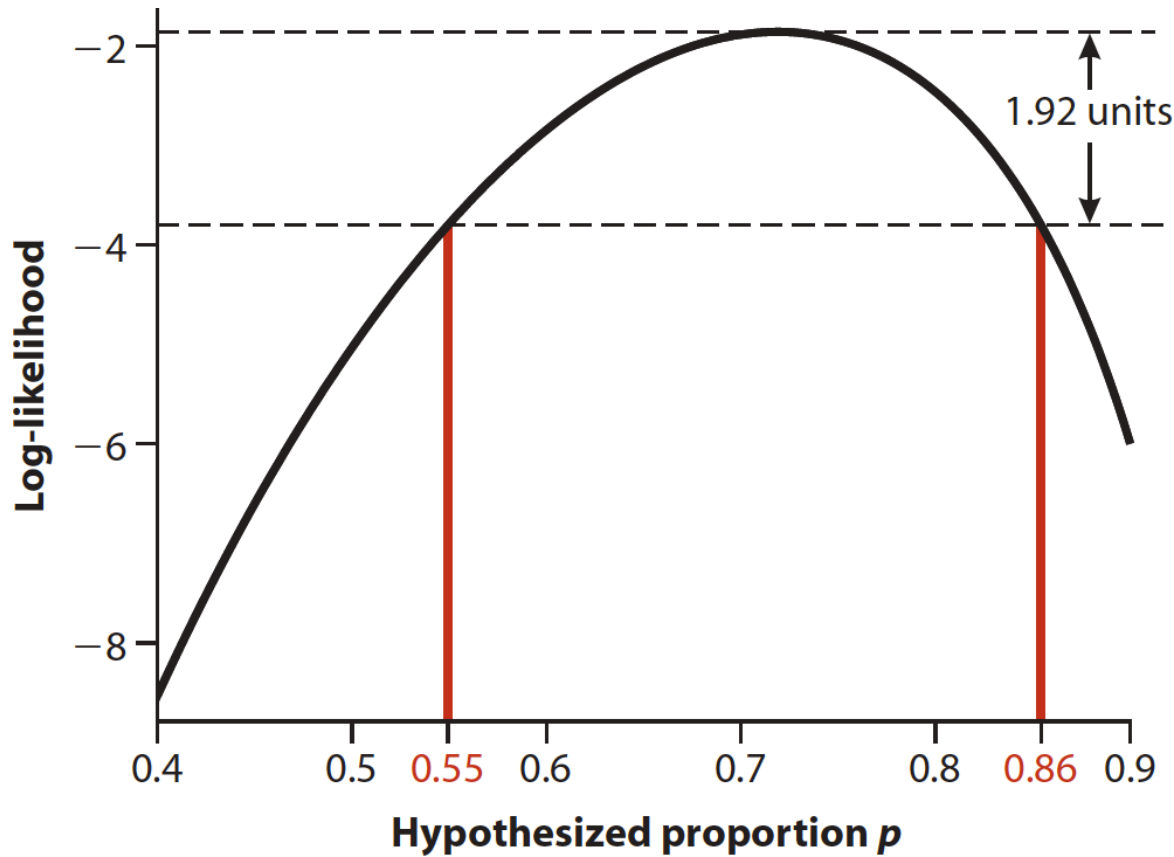
$$\hat{p} = 5/11$$

Maximum Likelihood Estimation:

- **Log likelihood** is computationally easier
  - Turns multiplication into addition

- Log-likelihood curve
  - Allows us to calculate interval estimate: **likelihood-based confidence interval**

- Use $X^2_{1,\alpha}/2$



Maximum Likelihood Estimation:

Example: Counting elephants via dung 'genetic fingerprints'

- originally identify 27 elephants
- sample 74, 15 of which were already sampled in original group of 27 ('resampled')

1. Probability model
   - Random sampling
   - Population is constant
   - Recapture number is 15 out of 74 in our sample
   - *what we want to know: likelihood of true **N** is the probability of obtaining this proportion of recaptures for different **N** values*
   - *This is the difficult step*
2. Likelihood
3. Maximum Likelihood estimate

Maximum Likelihood Estimation:

Example: Counting elephants via dung 'genetic fingerprints'

- originally identify 27 elephants
- sample 74, 15 of which were already sampled in original group of 27 ('resampled')

1. Probability model
   - Recapture number is 15 out of 74 in our sample
   - *what we want to know to determine the true **N** is the probability of obtaining this proportion of recaptures for different **N** values*
   - *This is the difficult step; in this case already done for us!*

$$P[Y \mid N] = L[N \mid Y] = \frac{\binom{n_1}{Y}\binom{N-n_1}{n_2-Y}}{\binom{N}{n_2}}$$

2. Likelihood
3. Maximum Likelihood estimate

Maximum Likelihood Estimation:

Example: Counting elephants via dung 'genetic fingerprints'

- originally identify 27 elephants
- sample 74, 15 of which were already sampled in original group of 27 ('resampled')

1. Probability model
   - Recapture number is 15 out of 74 in our sample
   - *what we want to know to determine the true **N** is the probability of obtaining this proportion of recaptures for different **N** values*
2. (log) Likelihood

$$\ln L[N \mid Y] = \ln[\binom{n_1}{Y}] + \ln[\binom{N-n_1}{n_2-Y}] - \ln[\binom{N}{n_2}]$$

$$\ln L[N \mid Y] = \ln[\binom{27}{15}] + \ln[\binom{N-27}{74-27}] - \ln[\binom{N}{74}]$$

3. Maximum Likelihood estimate

Maximum Likelihood Estimation:

Example: Counting elephants via dung 'genetic fingerprints'

- originally identify 27 elephants
- sample 74, 15 of which were already sampled in original group of 27 ('resampled')

1. Probability model
   - Recapture number is 15 out of 74 in our sample
   - *what we want to know to determine the true **N** is the probability of obtaining this proportion of recaptures for different **N** values*
2. Likelihood

$$\ln L[N \mid Y] = \ln[\binom{27}{15}] + \ln[\binom{N-27}{74-27}] - \ln[\binom{N}{74}]$$

3. Maximum Likelihood estimate

$$\hat{N} = 133$$