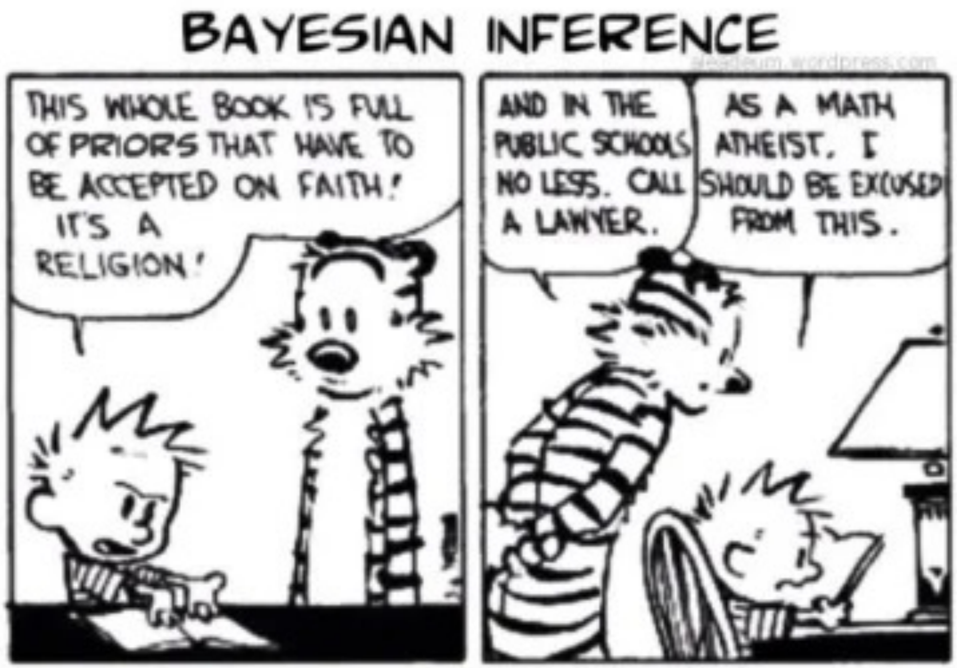
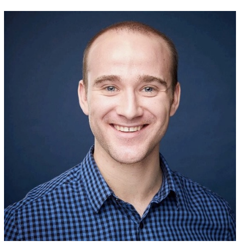
BIOL 501: Bayesian Data Analysis

Peer-Feedback Survey



Evan's Feedback Survey



14 March. 2023 Guest lecture by Evan Sidrow
UBC Stats Department, PhD Candidate,
evan.sidrow@stat.ubc.ca
UBC 2022 Term 2

Assignment #2 Extended until 9pm on March 20

Office Hours Announcements

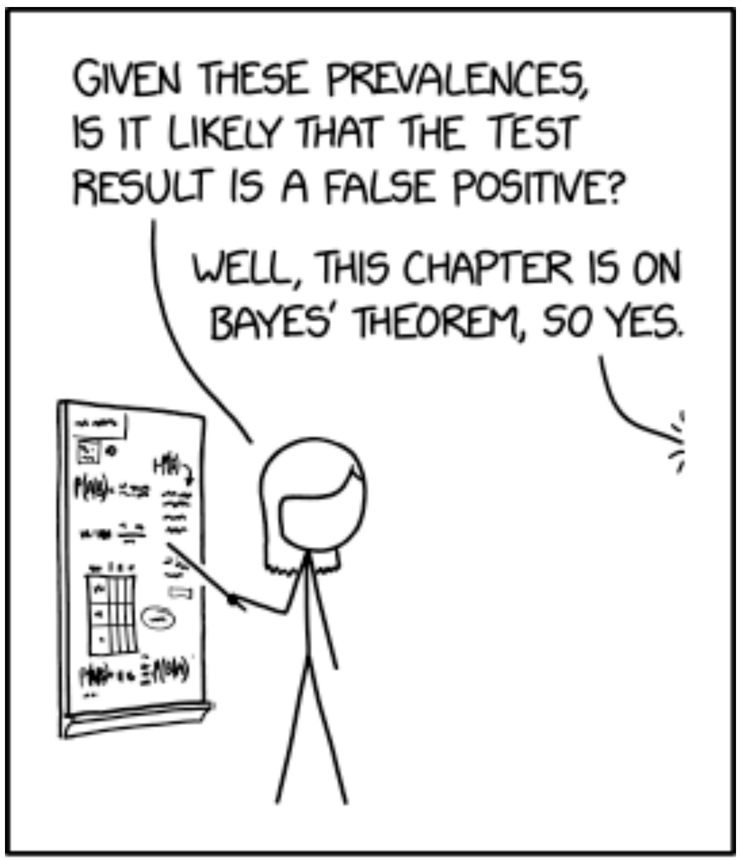
- **Beth this week**
  - No office hours Weds
  - **Thursday: 11-12** (Zoom)
  - **3-4pm** (in person) DMR #101 after class

- **Avery this week—Extra office hours for Assignment #2**
  - Tuesday 3-4pm (DMP #101)
  - Wednesday 1-2pm (BRC #336)
  - Friday 1-2pm (BRC #336)

Assignment #2 Extended until 9pm on March 20

Can also email **either of us** for an appointment. Suggest 3 specific dates/times and we will pick one

Our goal today is to introduce Bayesian inference and make it fun and accessible!



SOMETIMES, IF YOU UNDERSTAND BAYES' THEOREM WELL ENOUGH, YOU DON'T NEED IT.

Only an introduction into Bayesian inference and the tip of the iceberg

Outline

- Introduction to Evan Sidrow
- Why are Bayesian stats cool (and not scary)?
- What is probability?
  - Compare Frequentist vs. Bayesian definitions of prob.
- Bayes' Theorem
  - Prior probability
  - Posterior probability
- Bayesian Parameter Estimation
- Bayesian uncertainty quantification
- Hypothesis testing using Bayes factor
- **Assignment #2 due 18 March (Beth)**
- **Workshop Prep**

Thesis Topic

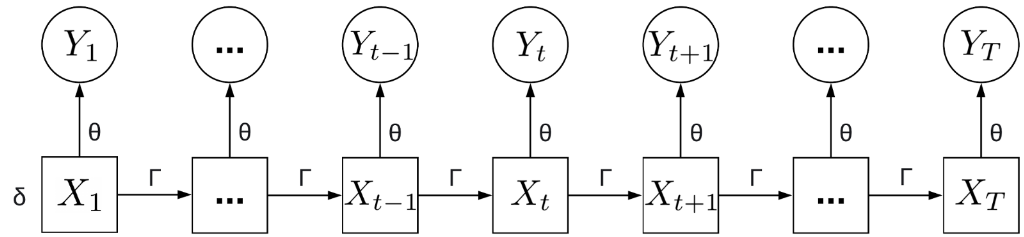
Find differences in behaviour between Northern and Southern Resident Killer Whales using kinematic data

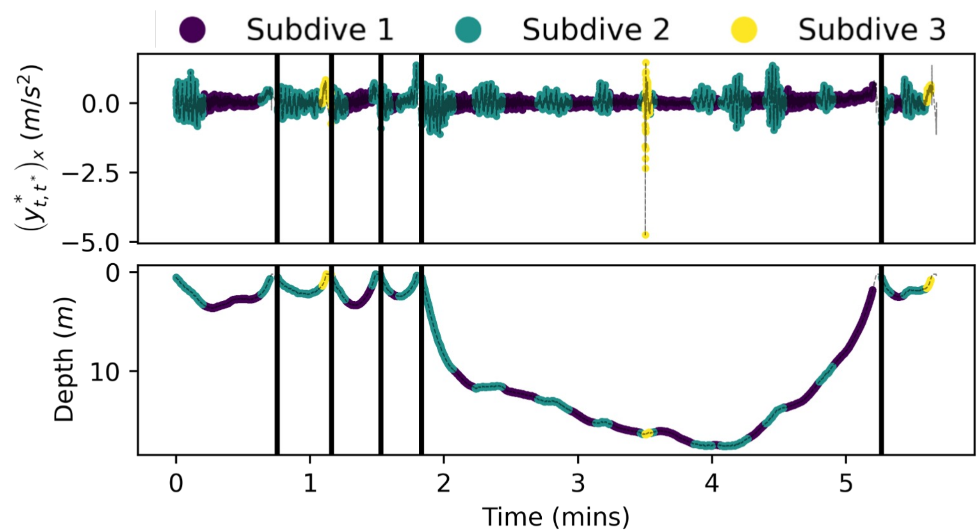
Hidden Markov Models

- **What** is the animal doing?
  - Behavioral state (X_t)

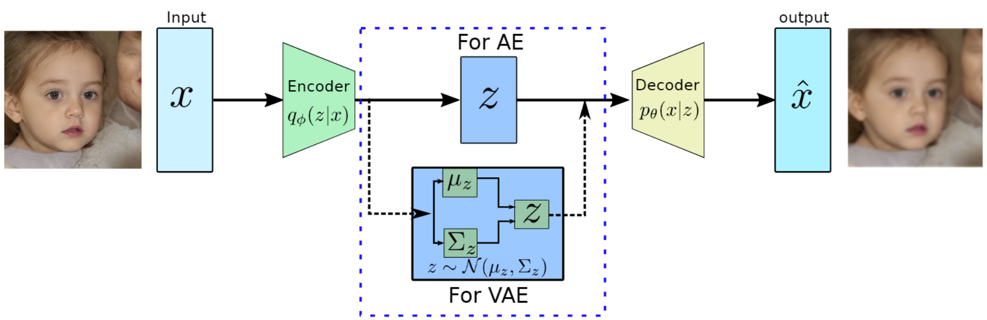
- **When** is it doing it?
  - Probability transition matrix (Γ)

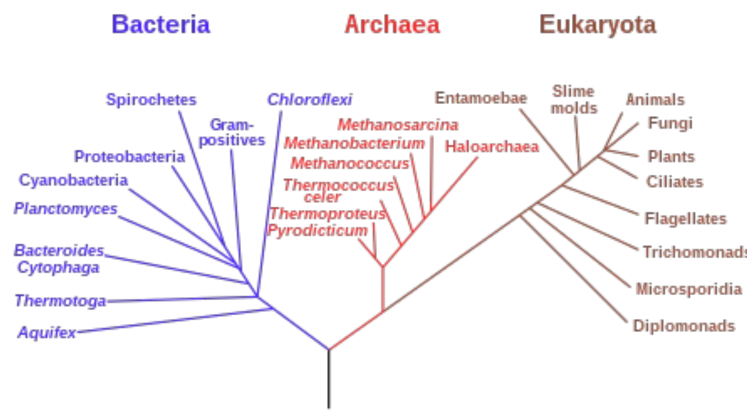
- **How** is it doing it?
  - Emission distributions (θ)

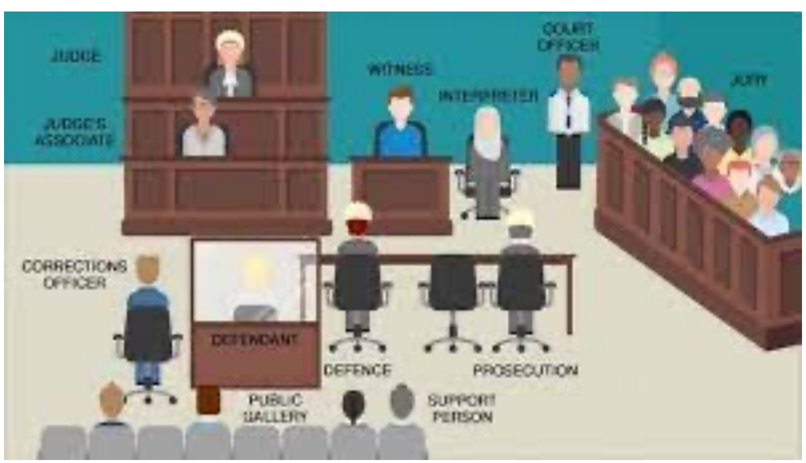


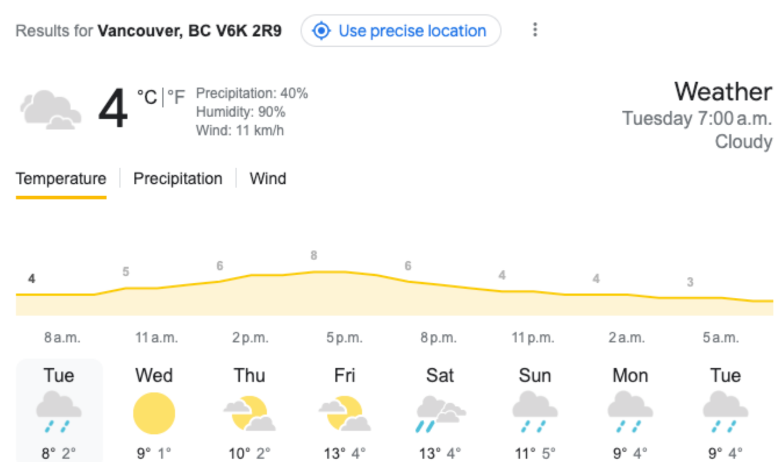


Why is Bayesian really cool?!









Bayesian methods are increasingly used in ecology and evolution

"Ecologists should be aware that Bayesian methods constitute a **radically different way of doing science**. Bayesian statistics is not just another tool to be added into the ecologists' repertoire of statistical methods. Instead, Bayesians categorically reject various tenets of statistics and the scientific method that are currently widely accepted in ecology and other sciences."

- B. Dennis, 1996, *Ecology*

"Ecologists are facultative Bayesians" (M. Mangel, pers. comm. To D. Schluter 2013)

What is probability?

Compare frequentist vs. Bayesian definitions of probability

Probability (Frequentist)

- **Probability:** Proportion of times that event would occur if a random trial is repeated over and over again under the same conditions

- **Probability distribution:** set of all mutually exclusive outcomes of a random trial and their probabilities of occurrence.

Probability (Frequentist) vs likelihood

This is an important slide and common area of confusion

- **Probability** refers to chance that a particular outcome occurs based on the values of parameters in a model.
  - Assumes parameters are trustworthy.

- **Likelihood** refers to **how well a sample provides support for particular values of a parameter** in a model.
  - Determines if we can trust the parameters in a model based on the sample data that we've observed.
  - **Likelihood works backwards from probability**

Determines if we can trust the parameters in a model based on the sample data that we've observed.

Probability (Frequentist)

Probability statements that **makes sense** under a frequentist definition

- If we toss a fair coin, what is the *probability* of 10 heads in a row?
- If we assign treatments randomly to subjects, what is the *probability* that a sample mean difference between treatments will be greater than 1 standard deviation?
- What is the *probability* of a result at least as extreme as that observed if the null hypothesis is true?

In these examples, the source of uncertainty is <u>sampling error</u>.

Probability (Frequentist)

Probability statements that **don't makes sense** under a frequentist definition

- What is the probability that Iran is building nuclear weapons?
- What is the probability that the fish sampled from that newly discovered lake represent two species rather than one?
- What is the probability that polar bears will be extinct in the wild in 40 years?

In these examples there is no random trial, so no sampling error.
The source of uncertainty is lack of information—not sampling error.

Probability (Frequentist)

Why they **<u>don't</u> make sense:**

- What is the probability that Iran is building nuclear weapons?
  [either Iran is or isn't – no random trial here]
- What is the probability that the fish sampled from that newly discovered lake represent two species rather than one?
  [either there is one species or there are two – no random trial]
- What is the probability that polar bears will be extinct in the wild in 40 years?
  [maybe this is from the accumulation of outcomes of random trials?]

In these examples there is no random trial, so no sampling error.
The source of uncertainty is lack of <u>information</u>—not sampling error.

Alternative Definition of Probability (Bayesian)

- *Probability* is a measure of a degree of belief associated with the occurrence of an event.

- A *probability distribution* is a list of all mutually exclusive events and the degree of belief associated with their occurrence.

- Bayesian statistics applies the mathematics of probability to uncertainty measured as subjective degree of belief.

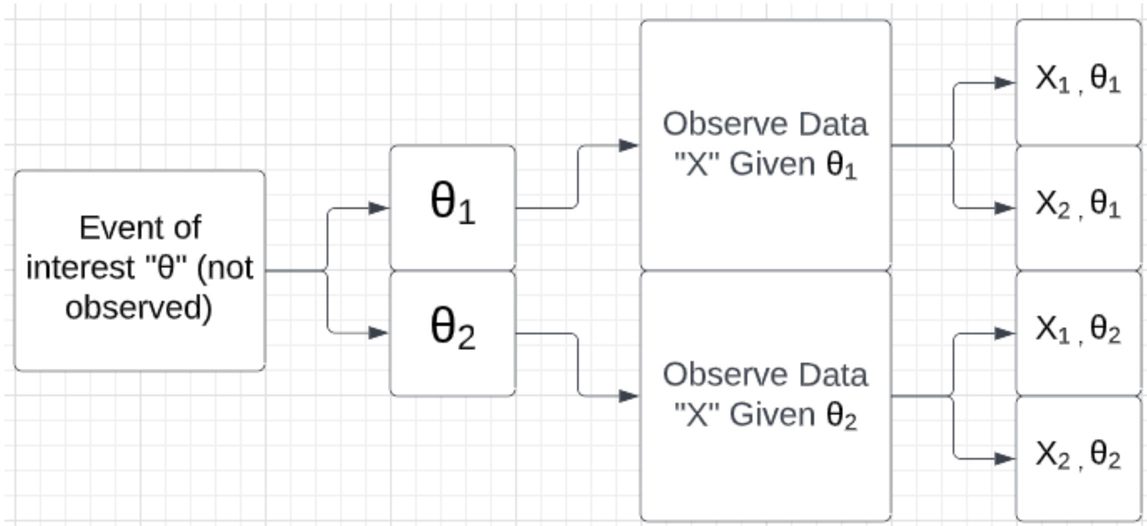
Bayesian data analysis

- Bayesian framework is related to likelihood methods.
- With likelihood, we treat the data as given and vary the parameter to find that value for which the probability of obtaining the data is highest.
- Bayesian methods go one step further, **treating the parameter (or hypothesis) as a random variable and seeking the value having highest posterior probability, given the data.**
- **New:** We need to specify a prior probability distribution for the parameter values.

Bayes' Theorem

Prior and posterior probability

Bayes' Theorem

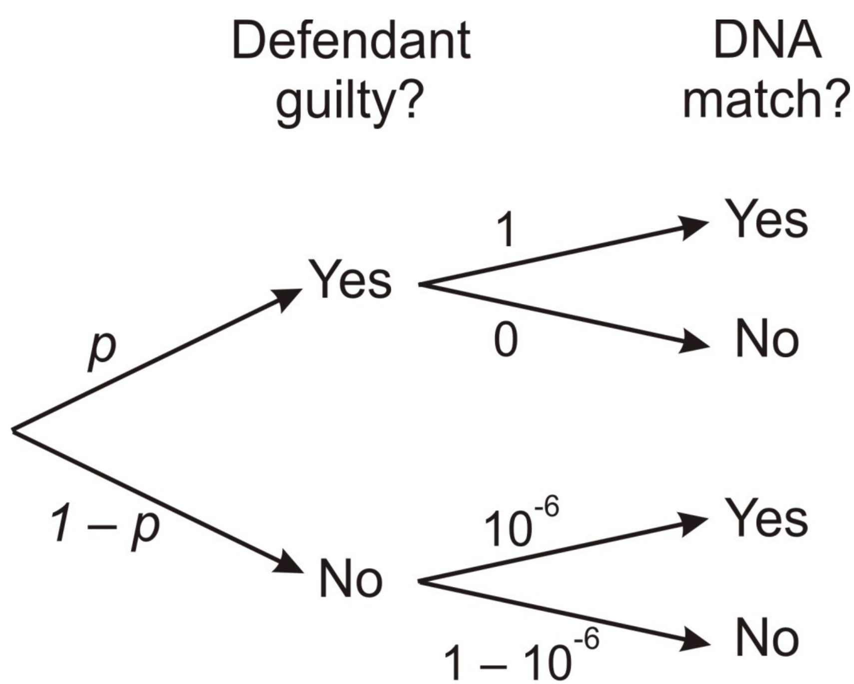


$$\Pr[\theta_1|X_1] = \frac{\Pr[X_1|\theta_1]\,\Pr[\theta_1]}{\Pr[X_1]}$$

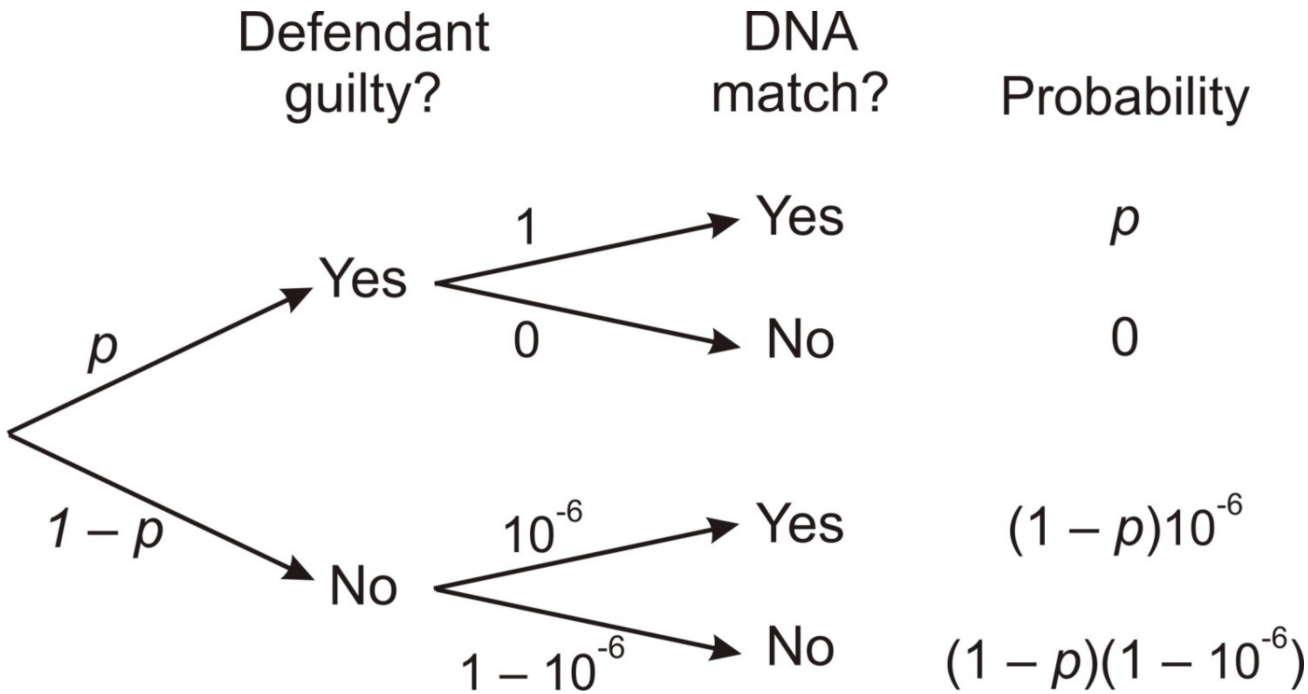
$$\Pr[\theta_1|X_1] = \frac{\Pr[X_1|\theta_1]\,\Pr[\theta_1]}{\Pr[X_1|\theta_1]\,\Pr[\theta_1] + \Pr[X_1|\theta_2]\,\Pr[\theta_2]}$$

Example of Bayes' Theorem for forensic evidence

- Court of law to quantify the evidence for and against the guilt of the defendant based on a match to DNA evidence left at the crime scene.
- What is the probability of guilt given a positive DNA match (assuming no contamination of samples)?



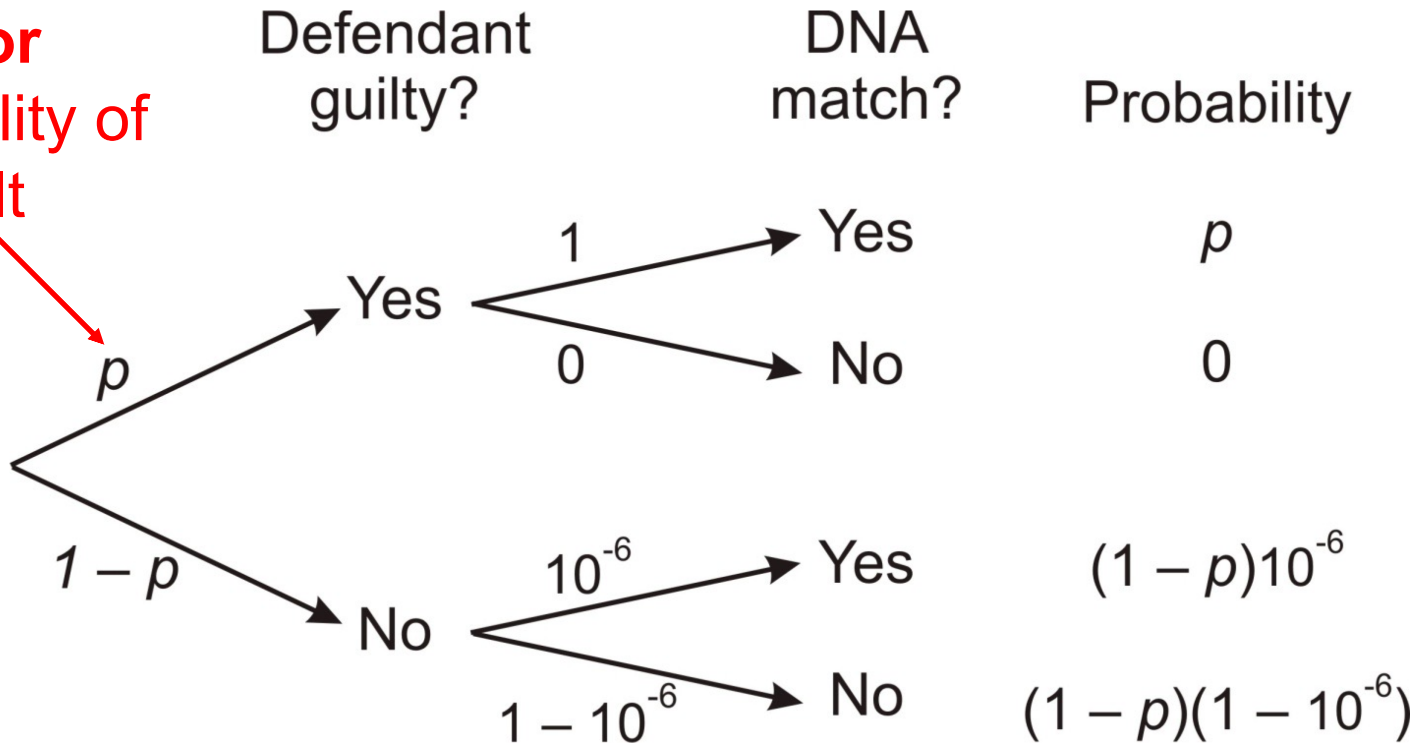
What is the probability of guilt given a positive DNA match?



$$\Pr[\text{guilt} \mid \text{match}] = \frac{1(p)}{1(p) + 10^{-6}(1-p)}$$

Bayesian inference with data

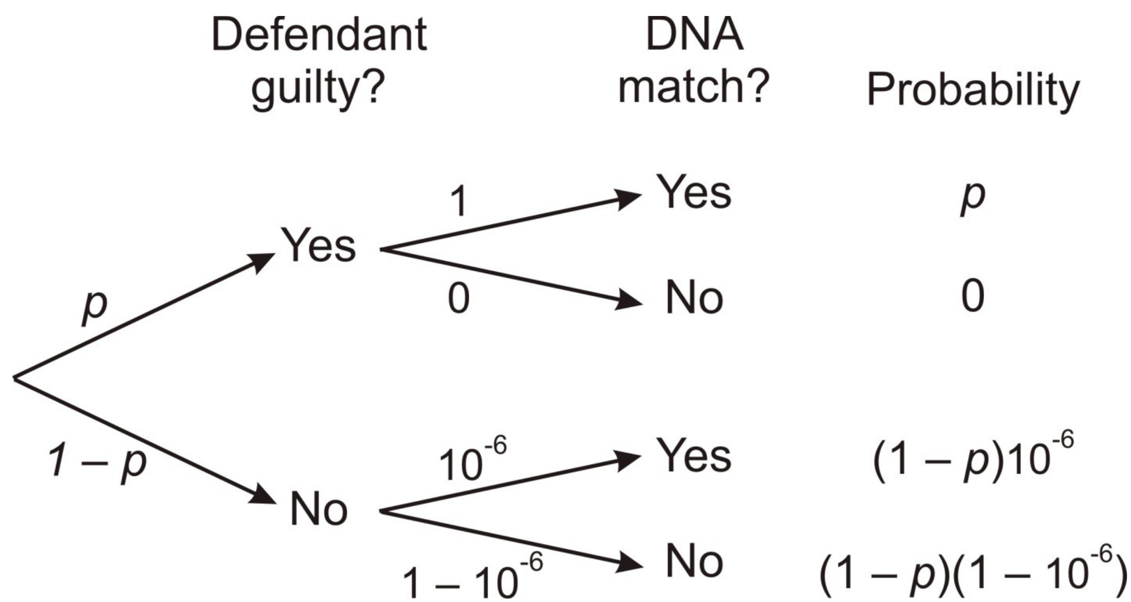
Prior
probability of guilt



Posterior
probability of guilt

$$\Pr[\text{guilt} \mid \text{match}] = \frac{1(p)}{1(p) + 10^{-6}(1 - p)}$$

Bayesian inference in action



$$\Pr[\text{guilt} \mid \text{match}] = \frac{1(p)}{1(p) + 10^{-6}(1-p)}$$

If $p = 10^{-6}$ then $\Pr[\text{guilt} \mid \text{match}] = 0.5$

If $p = 0.5$ then $\Pr[\text{guilt} \mid \text{match}] = 0.999999$

So, is the defendant guilty or innocent?

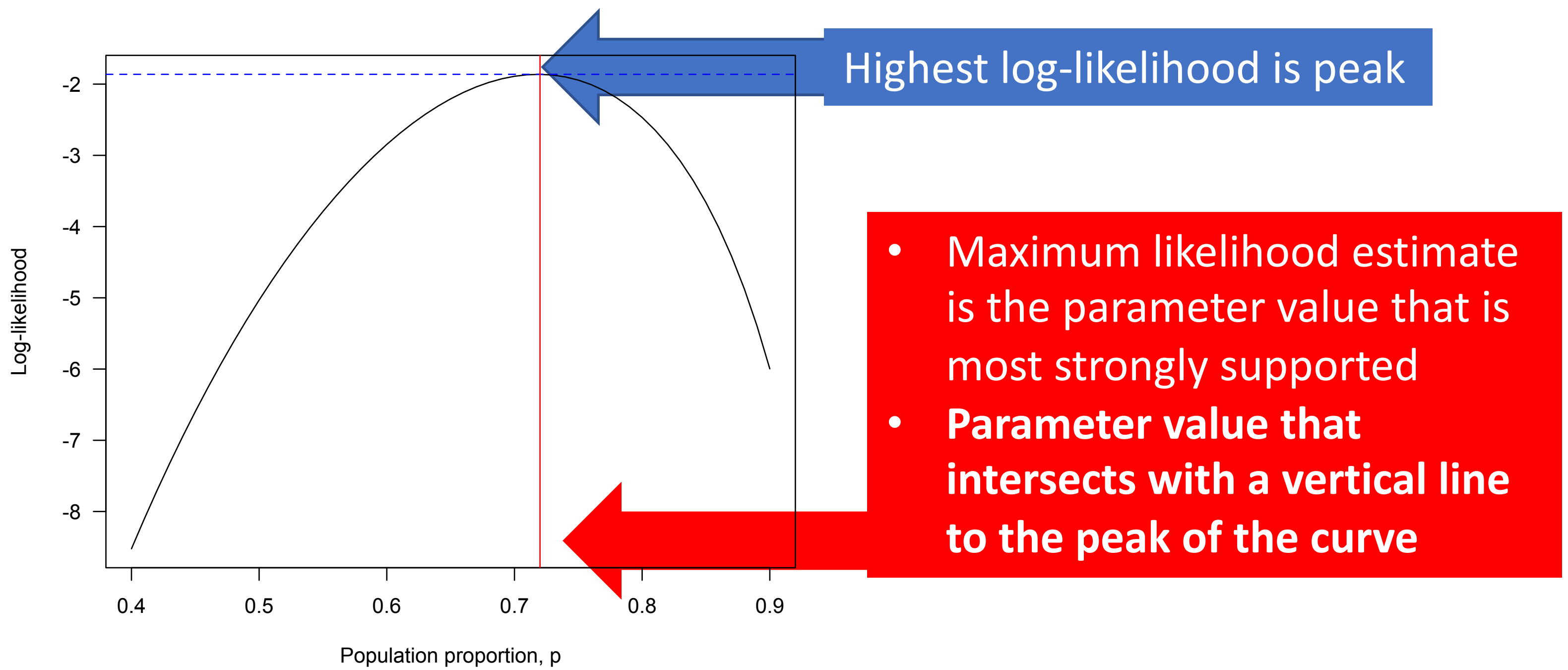
Discussion: Chat with partner 2 min

- Is the defendant guilty or innocent?

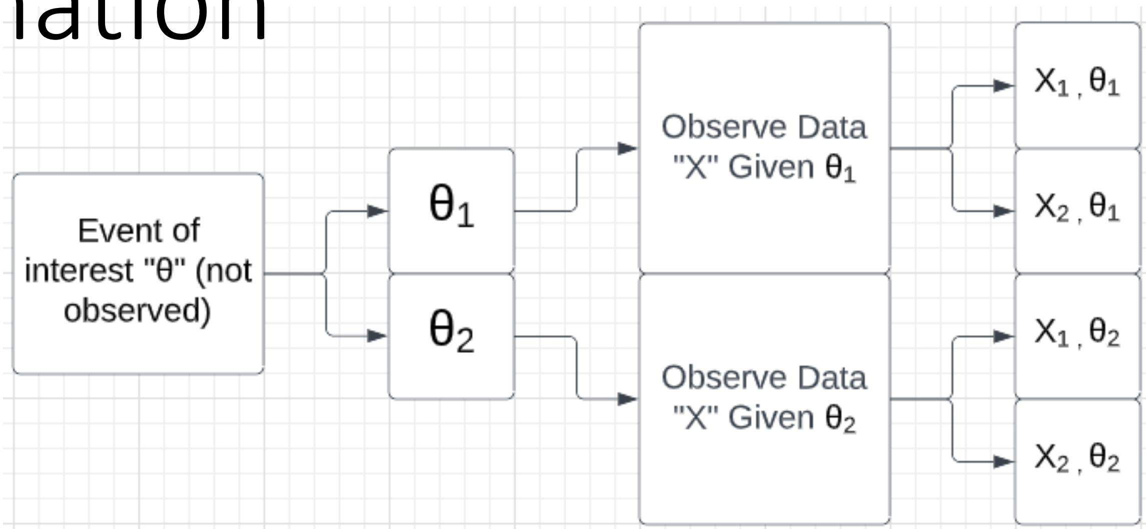
Parameter estimation

Frequentist: Maximum Likelihood Estimate

- **Likelihood ratio** (difference of log-likelihoods) measures relative <u>support</u> for alternative parameter values
- **Maximum likelihood estimate (MLE)** of a parameter is the parameter value having the <u>highest likelihood</u> (and log-likelihood), given the data
  - MLE is the parameter value **most strongly supported** by the data



Bayesian: Maximum a Posteriori Estimation



Likelihood

Prior

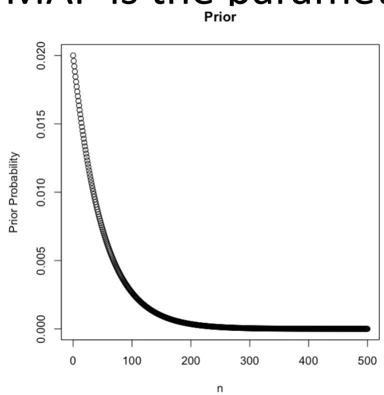
Normalization (Doesn't depend on θ)

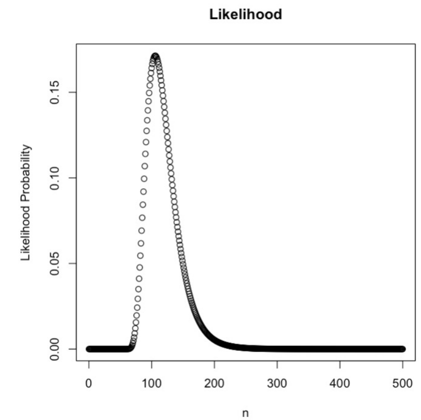
$$\Pr[\theta_1|X_1] = \frac{\boxed{\Pr[X_1|\theta_1]}\,\boxed{\Pr[\theta_1]}}{\boxed{\Pr[X_1]}}$$

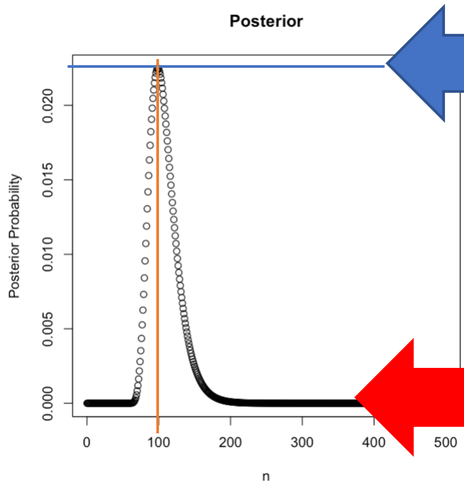
$$\Pr[\theta_1|X_1] = \frac{\Pr[X_1|\theta_1]\,\Pr[\theta_1]}{\Pr[X_1|\theta_1]\,\Pr[\theta_1] + \Pr[X_1|\theta_2]\,\Pr[\theta_2]}$$

Bayesian: Maximum a Posteriori Estimation

- Maximum a Posteriori Estimation **(MAP):**of a parameter is the parameter value having the highest posterior (and log-posterior), given the data
  - MAP is the parameter value **with the highest probability density**







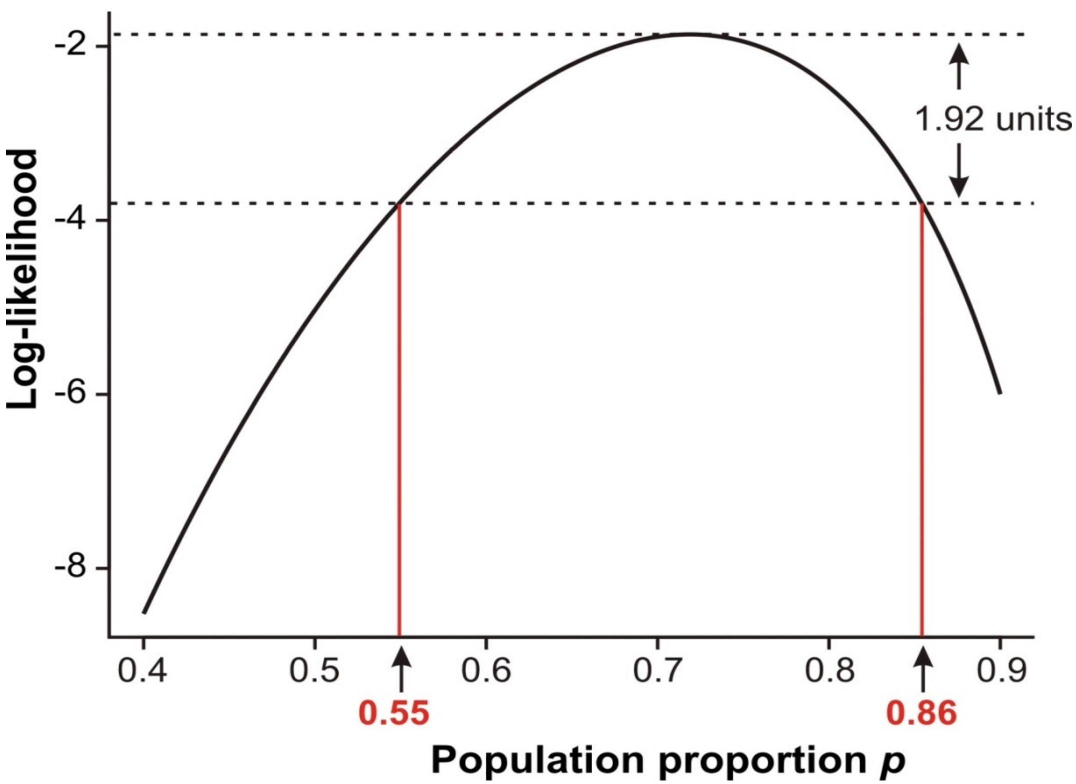
Highest posterior is peak

- Maximum a Posteriori estimate is the parameter value with the highest probability
- **Parameter value that intersects with a vertical line to the peak of the curve**

Uncertainty Quantification

Frequentist: Confidence Intervals

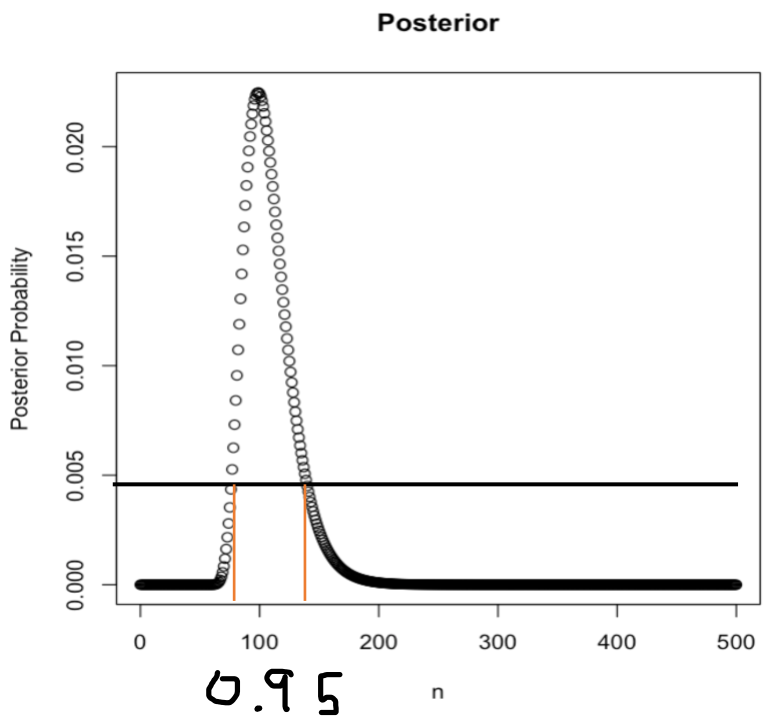
- Approximate 95% confidence interval is obtained with the values corresponding to 1.92 log-likelihood units below the maximum



Definition: random interval whose confidence level represents the long-run proportion of intervals that contain the true parameter

Bayesian: Credible Intervals

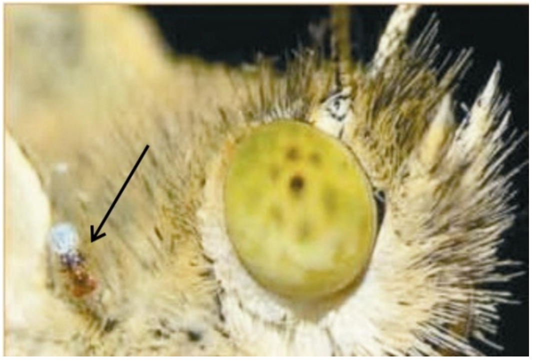
- 95% credible interval is obtained by defining a level and including all parameter values with a posterior higher than that level



Definition: interval where an unobserved parameter value falls within it with a particular probability (often 0.95)

Example: Estimate binomial proportion p with Wasps

Example: Estimate a binomial proportion p with Wasps



- **Data:** The tiny wasp rides on female butterflies. When a butterfly lays her eggs, the **wasp parasitizes the freshly laid butterfly eggs**.
- Can wasps distinguish mated female butterflies from unmated females? (pointless to ride on a female that already laid eggs).
- **Trials:** Wasp presented with 2 female butterflies to ride on (mated vs. unmated butterflies)
- **Observed**: 23 of the 32 wasps chose the mated female
- **Probability** of the wasp riding on a mated caterpillar in each trial was 0.50

What is the proportion of wasps in the population choosing the mated female caterpillar
Success=chose female mated caterpillar
Failure=wasp chose unmated female caterpillar

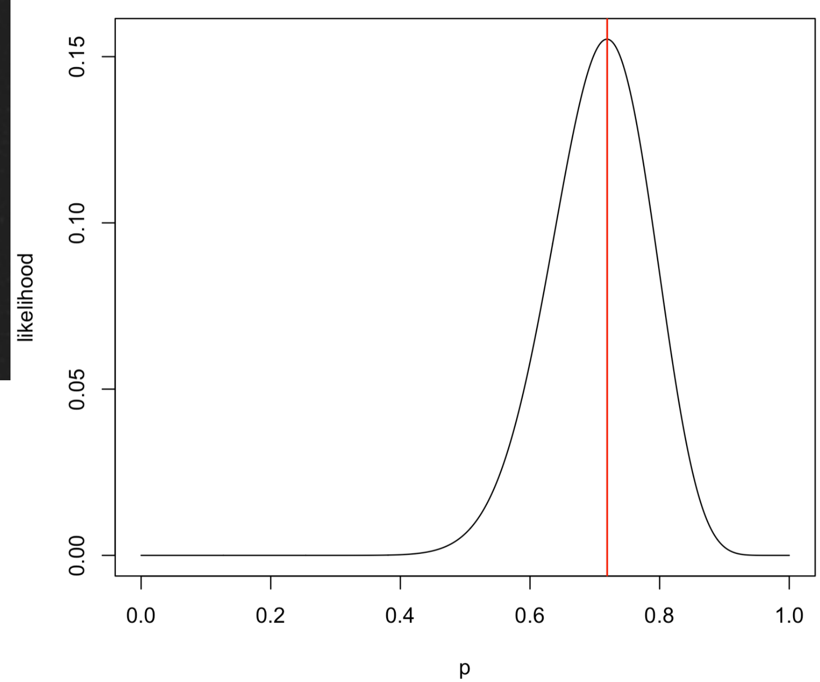
We will use this case study again in GLMS

Fatouros et al. (2005)

Review: Find MLE

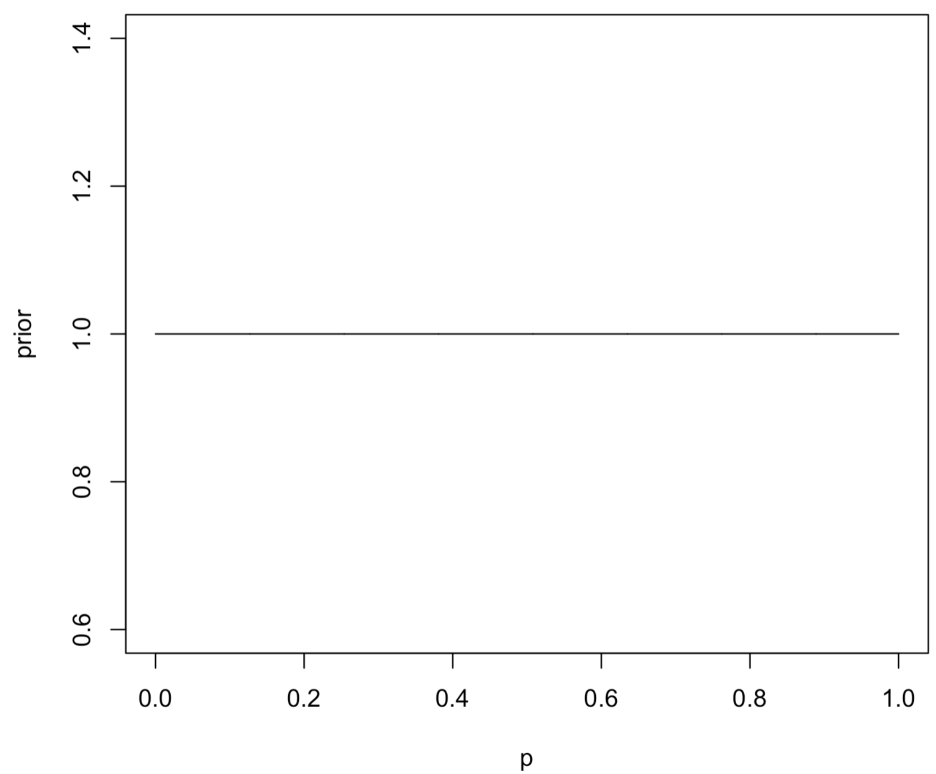
```r
# Part 1: find the MLE for the WASP example

p <- seq(0.0,1.0,by=0.001) # possible values for probability of success
x <- 23 # number of "successes" where wasp chose mated female
n <- 32 # number of random trails

# evaluate the likelihoods
likelihood <- dbinom(x, size=n, p=p)
plot(p,likelihood,type="l")

# find MLE
p.MLE <- p[which.max(likelihood)]
print(max(likelihood))
print(p.MLE)
abline(v=p.MLE,col='red')
```

```
> print(max(likelihood))
[1] 0.1552514
> print(p.MLE)
[1] 0.719
```



Define a *Prior distribution* over p

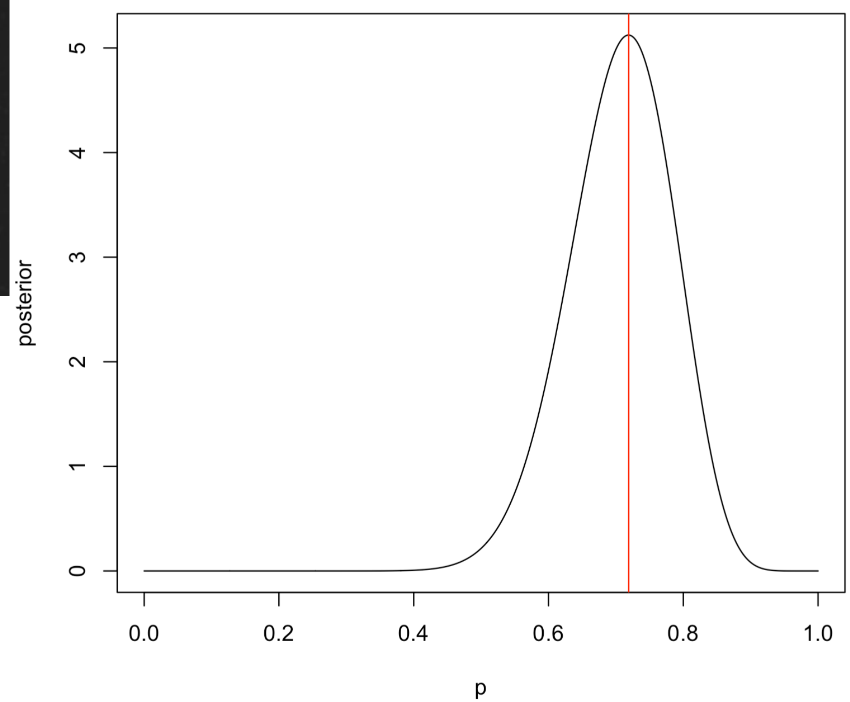
```
# Part 2: add a prior to the WASP example
prior = dbeta(p,1,1)
plot(p,prior,type="l")
```



Find *Posterior distribution* and MAP

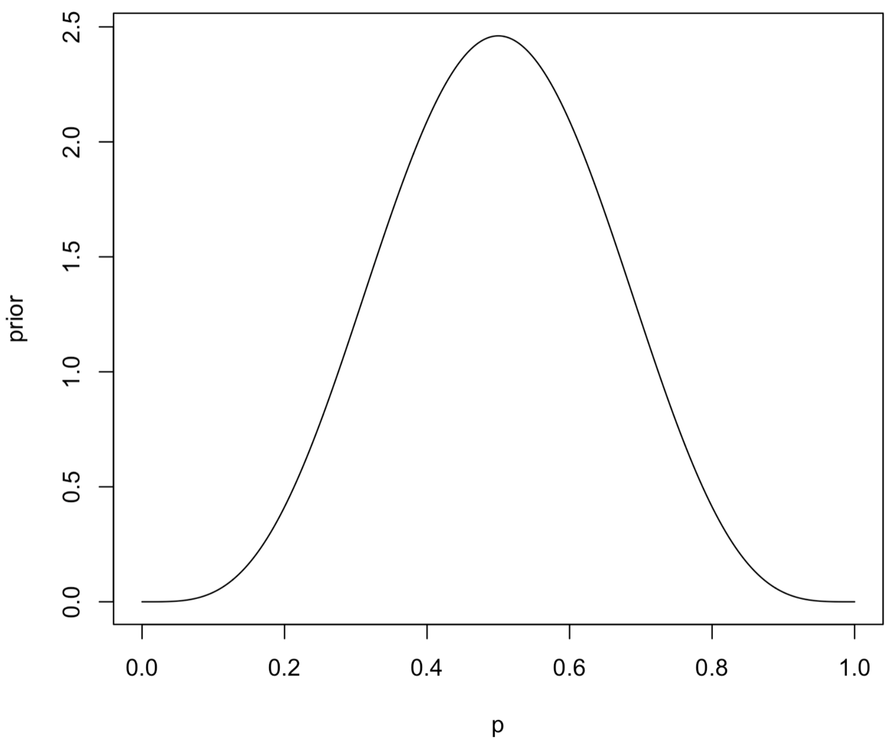
```r
# Part 3: find the posterior distribution
posterior = prior * likelihood

# normalize the posterior
posterior = posterior / (sum(posterior) * 0.001)
plot(p,posterior,type="l")

# get the MAP
p.MAP <- p[which.max(posterior)]
print(max(posterior))
print(p.MAP)
abline(v=p.MAP,col='red')
```

```
> print(max(posterior))
[1] 5.123298
> print(p.MAP)
[1] 0.719
```



Discussion: Chat with partner 2 min

- How would the MAP estimate change with the following prior? Why would we pick the following prior?




```
# Part 4: new prior
prior = dbeta(p,5,5)
plot(p,prior,type="l")
```

Hypothesis testing
(if time)

Bayesian hypothesis testing using the Bayes factor

- Bayes factor $= \dfrac{\Pr[\text{data}|H_A]}{\Pr[\text{data}|H_0]}$
- A Bayes factor greater than 1 indicates that H_A has more support from the data than H_0. What value for the Bayes factor constitutes strong evidence for H_A?

- A Bayes factor of 1 – 3 is considered "anecdotal evidence" for H_A
- A Bayes factor of 3 – 10 is considered "substantial evidence" for H_A
- A Bayes factor of 10 – 30 is considered "strong evidence" for H_A

Bayesian Inference Summary

- Bayesian probability is a different concept than frequentist probability
- Bayes’ Theorem can be used to estimate and test hypotheses using posterior probability
- Requires prior probability
- Influence of prior probability declines with more data
- Interpretation of interval estimates (credible interval) differs from the frequentist definition (confidence interval)
- Bayesian hypothesis testing using the Bayes factor
- Bayesian ideas are becoming used more in ecology and evolution

Assignment #2: Due 20 March 9pm

- **Dataset requirement details on Canvas**
- **Linear, mixed, or generalized linear model in R**. You can choose which type of **linear** model to use, as long as it is appropriate for your dataset.
  - Fixed→Lm, glm, gam
  - Mixed→LME, GLMM, GAMM
- Only 1 response variable
- At least 1 categorical factor
- Include at least 1, and no more than 2, additional explanatory variables

Self-assessment with rubric *prior* to turning it in

How to find data for Assignment #2

- Ask labmates, supervisor, other grad students for data
- Extract data that was used for something else other than a linear model
  - "How to find data for practicing R.docx"/ Assignment #2 on Canvas
- **Not** ok to redo a figure that's already published and analyzed <u>as a linear model</u>, but you can extract data from published paper that was used for something else

- **Don't** use simulated data
- Yes it's ok, if it's your own thesis data, undergrad data

Workshop
Thursday 1-3pm

Workshop Thurs: Bayesian data analysis

- 2 hours student self-led work through example with us here to help
- Workshop link: https://www.zoology.ubc.ca/~bio501/R/workshops/bayes.html
- Revisits the elephant mark-recapture data from Likelihood Workshop
- **Do the elephant exercise first**, not the biodiversity advanced example