

UNIVERSITY *of* WASHINGTON

Data Science UW

Methods for Data Analysis



Intro to Bayesian models
Steve Elston



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?

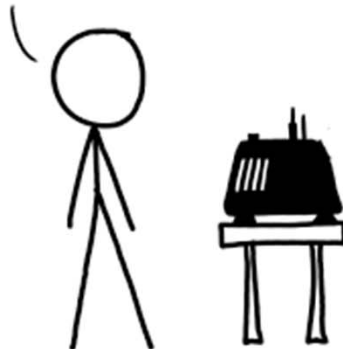
ROLL
YES.



FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.

SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.



W

Review

- > Other regularization methods
 - Ridge
 - Lasso
- > Overview of logistic regression
- > Time series



Time Series Summary

- > Time series data are everywhere!
- > A stochastic process is 'stationary' if there is **no trend** and **constant variance**.
- > Time series values can only be sampled once
- > Time series have serial dependency
- > Decompose time series into trend, seasonal and remainder (noise) components
- > ARIMA process:
 - $AR(p)$ – AR process of order p , for dependency in values
 - $I(d)$ – d th order difference operator, removes trend
 - $MA(q)$ – MA process of order q , dependency in noise



Time Series Summary

- > ARIMA(p,q,q,P,D,Q) process to model seasonal component
- > White noise is an ARIMA(0,0,0) process
- > Random walk is not stationary, but difference series is
- > Use R forecast package to make life easy
- > With variable volatility use ARCH or GARCH models – Beyond the scope of course





Topics



- > Bayesian Statistics
 - Bayesian Inference
 - MCMC distributions



Where do Bayesian Models Fit With Other Methods?

Bayesian models are in the class of modern stats methods

- > Jackknife
- > Bootstrap
- > Cross validation
- > Modern Bayesian models using Markov Chain Monte Carlo (MCMC) methods
- > All are computationally intensive



Bayesian statistics have a long history

- > *An Essay towards solving a Problem in the Doctrine of Chances*, Thomas Bayes, 1763
- > *Essai philosophique sur les probabilités*, Pierre-Simon Laplace, 1814



Introduction to Bayesian Statistics

- > Most of the statistics we done use assumed parameters and limiting distributions. This is called 'Frequentist Statistics'.
- > Main difference between Bayesian and Frequentist statistics
 - Bayesian view of the world includes updating/changing beliefs new observations
 - Bayesian view takes prior beliefs into account.
- > Example: If we've lost our keys, we either
 - (1) Search our house from top to bottom.
 - (2) Search our house starting at the areas we have previously lost our keys before (laundry basket, desk, coat pockets,...), then we move onto more and more less likely places.



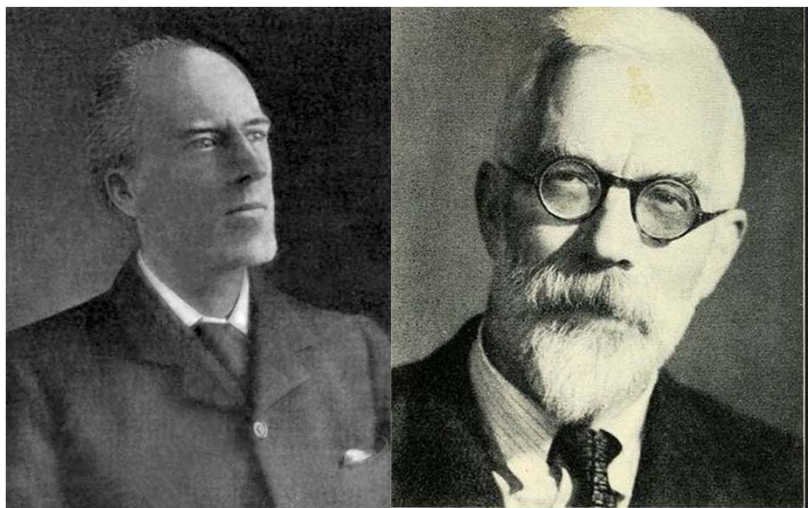
Bayesian Statistics Have Not Always Been Accepted

Many 20th century frequentists considered Bayesian models heretical

- > Bayesian methods use priors to quantify what we know about parameters.
- > Frequentists do not quantify anything about the parameters, using p-values and confidence intervals to express the unknowns about parameters.
- > And yet:
 - Bayesian models were used to improve artillery accuracy in both world wars
 - Bayesian models used by Alan Turing to break codes



Frequentist vs. Bayesian Views



Frequentist	Bayesian
Goal is a point estimate and confidence interval	Goal is posterior distribution
Start from observations	Start from prior distribution
Re-compute model given new observations	Update belief (posterior) given new observations
Examples: Mean estimate, t-test, ANOVA	Examples: posterior distribution of mean, overlap in highest density interval (HDI)

When Should You Use Bayesian Models?

- > Using a specific way to solve some problems is not a lifetime commitment.
- > In fact, the common belief is that some problems are better handled by Frequentist methods and some with Bayesian methods.



Bayes Law

> Remember the rule for conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

> And

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

> Solving for $P(A \cap B)$

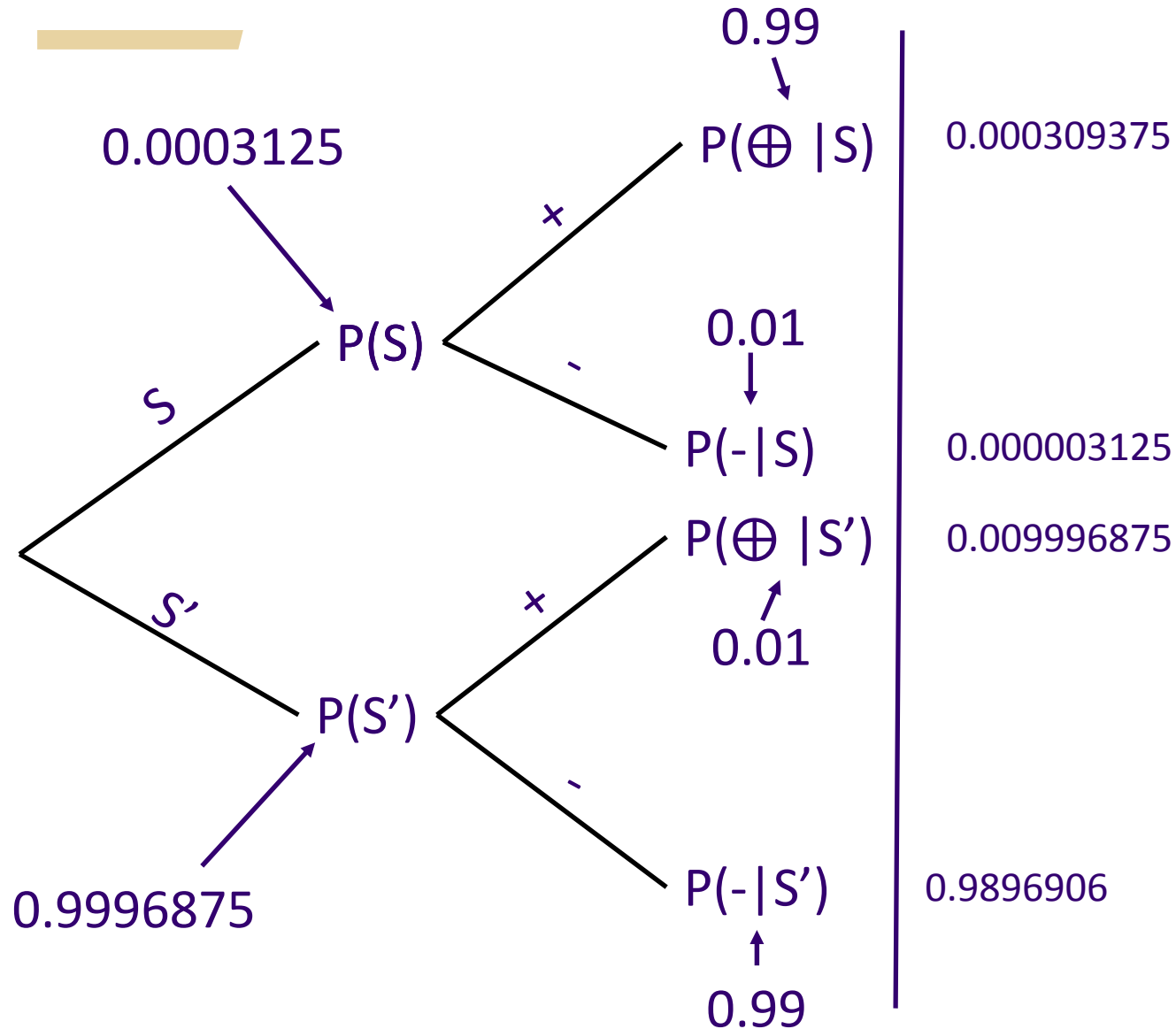
$$P(B)P(A|B) = P(A)P(B|A)$$

> Or

$$P(A|B) = P(B|A) \frac{P(A)}{P(B)}$$



Recall Conditional Probability Tree For a 99% Accurate Medical Test



W

Bayes Law

$$P(A|B) = P(B|A) \frac{P(A)}{P(B)}$$

> Applications:

- Disease Testing: A = Have Disease, B = Tested Positive


$$P(\text{Test} + | \text{Disease}) \neq P(\text{Disease} | \text{Test} +)$$

$$P(\text{Disease} | \text{Test} +) = P(\text{Test} + | \text{Disease}) \frac{P(\text{Disease})}{P(\text{Test} +)}$$

↑
High Probability,
usually the reported
accuracy of test.

↑
If the disease is rare,
the P(disease) will
be very small.

> Example:

$$P(\text{Disease} | \text{Test} +) = (0.99) \frac{0.0003125}{0.010306} = 0.031$$


Another example

Marginal probabilities of eye and hair color

	P(Hair Color Eye Color)				
Eye Color	Black	Brunette	Red	Blond	Marginal (eye color)
Brown	0.11	0.2	0.04	0.01	0.36
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.11	0.22	0.99



Another example

What is the probability hair color given Blue eyes?

	P(Hair Color Blue Eyes)				
Eye Color	Black	Brunette	Red	Blond	Marginal (eye color)
Blue	0.08	0.39	0.08	0.44	1

W

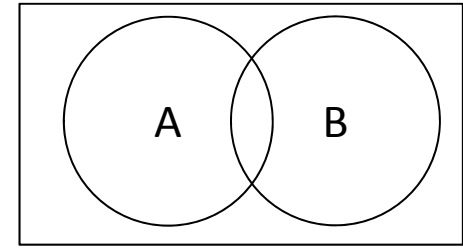
Why are Bayesian Models Useful Now?

Combination of new algorithms and cheap computers

- > Statistical sampling - Ulam, von Neuman; 1946, 1947
- > MCMC - Metropolis et al. (1953) Journal of Chemical Physics
- > Hastings (1970) ; Monte Carlo sampling methods using Markov chains and their application
- > Geman and Geman (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images
- > Duane, Kennedy, Pendleton, and Roweth (1987) Hamiltonian MCMC
- > Gelfand and Smith (1990) Sampling-based approaches to calculating marginal densities.



Revisiting Conditional Probability



Fun Fact #1

$$P(B) = P(B \cap A) + P(B \cap \text{not } A)$$

Fun Fact #2

$$P(B|E) = \frac{P(B \cap E)}{P(E)} \quad \text{OR} \quad P(B|E)P(E) = P(B \cap E)$$

Combining these results in:

$$P(B) = P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)$$

W

Another way to write Bayes Law:

$$P(A|B) = P(B|A) \frac{P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

If $P(B) = P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)$

Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)}$$

Note: Usually $P(B)$ is hard to estimate.

W

A Simpler Way to Write Bayes Law:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\text{not } A)P(\text{not } A)}$$

Or $P(A|B) = k \cdot P(B|A)P(A)$

Or $P(A|B) \propto P(B|A)P(A)$

Posterior Distribution

The Likelihood

Prior Distribution

W

Interpretation with Modeling

$$P(A|B) \propto P(B|A)P(A)$$

Posterior \propto Likelihood * Prior

- > As this applies to parameters in a model (partial slopes, intercept, error distributions, lasso constant,...) and the observed data:

$$P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters})$$

- > Given prior assumption about the behavior of the parameters (the prior), produce a model which tells us the probability of observing our data, to compute new probability of our parameters.



Interpretation with Modeling

$$P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters})$$

- > Identify data relevant to the research question. E.g.: what are the measurement scales of the data?
- > Define a descriptive model for the data. E.g.: pick a linear model formula.
- > Specify a prior distribution of the parameters. E.g. We think the error in the linear model is Normally distributed as $N(0, \sigma^2)$.
- > Use the Bayesian inference formula (above) to re-assess parameter probabilities.
- > Update if more data is observed.



How do we choose a prior distribution?

- > Recall $P(A|B) \propto P(B|A)P(A)$
- > If we want posterior to be same family as the likelihood, we need a **conjugate prior** distribution

Likelihood	Prior
Binomial	Beta
Bernoulli	Bet
Poisson	Gamma
Categorical	Dirichlet
Normal	Normal, Inverse Gamma



How to choose a prior distribution

Use the information you have

- > Prior observations
- > Domain knowledge
- > **Watch out:** A uniform prior is informative
 - Limits on range of values
 - What is the conjugate distribution?



An overused example, but for good reason.

- > Tasked with identifying where on a target archery board the bullseye is. But we can only see the back of the target and where the arrows puncture through as a marksman fires at it.

Back of target: What we see.



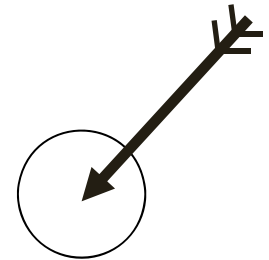
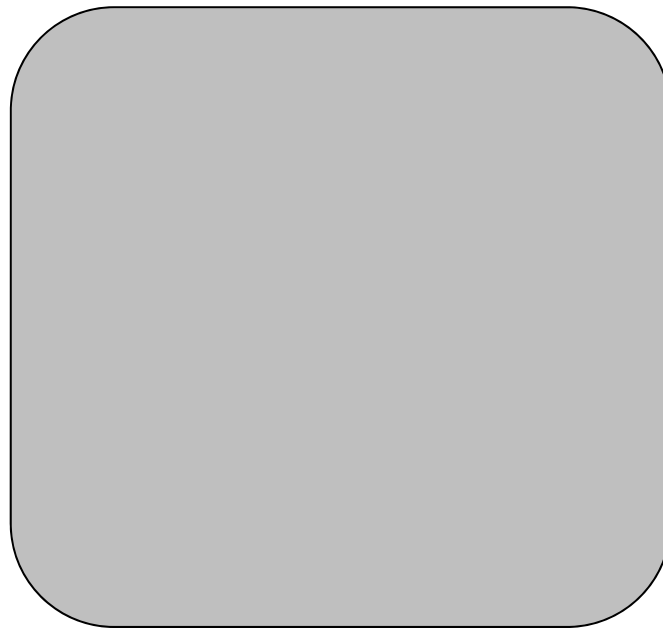
Front of target?



W

Archery Example

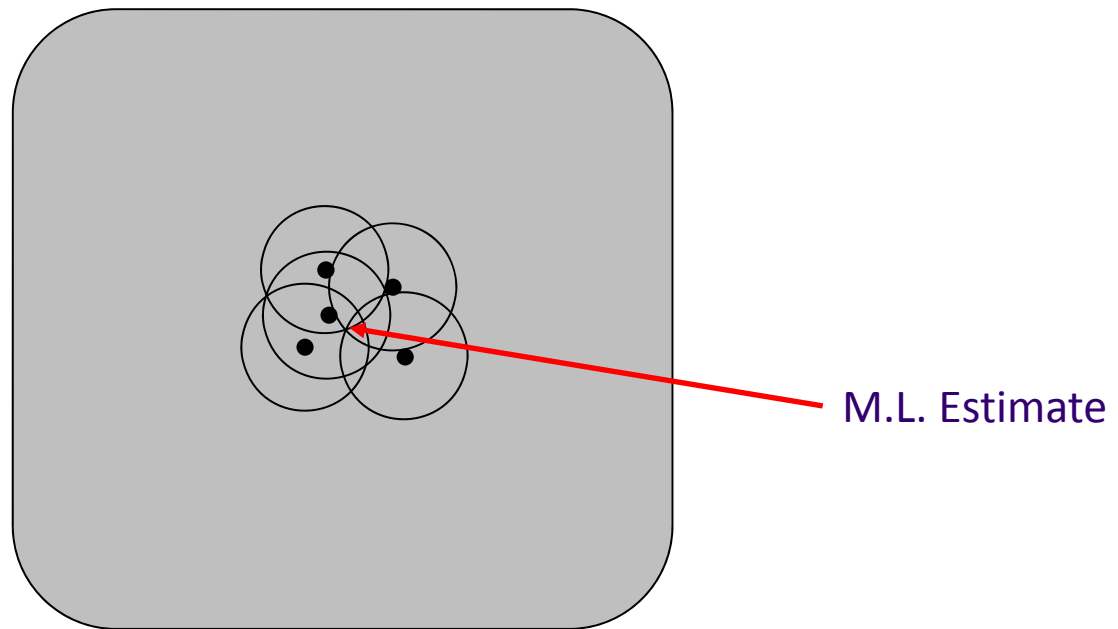
- > We are told that a marksman is firing at it and they are always within 10 centimeters of the target 95% of the time.



W

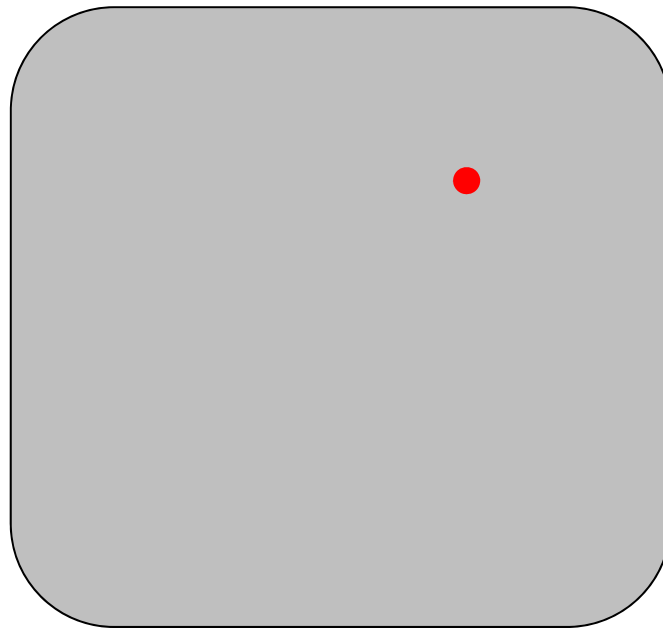
Archery Example

- > Here are the archer's first 5 shots with a 10 cm radius around it.
- > A frequentist observes the maximum likelihood point as the best guess.



Archery Example

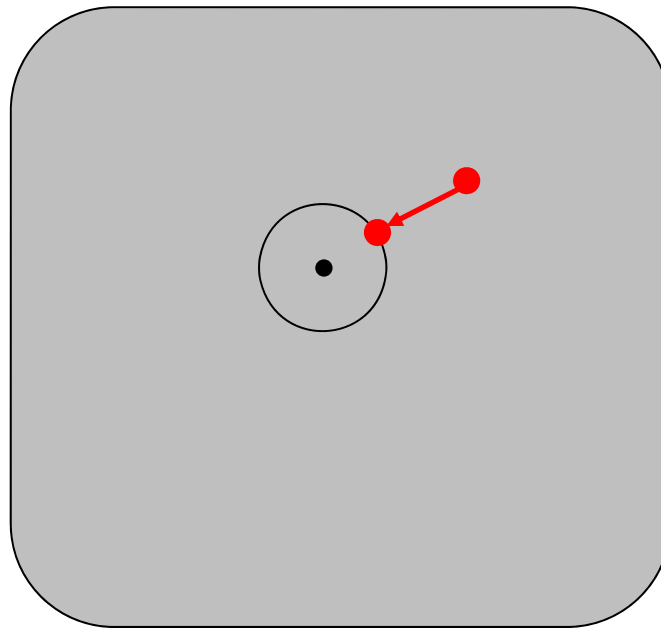
- > A Bayesian approach is to create a 'Prior', or a previous belief of where the target is. (Red point)



W

Archery Example

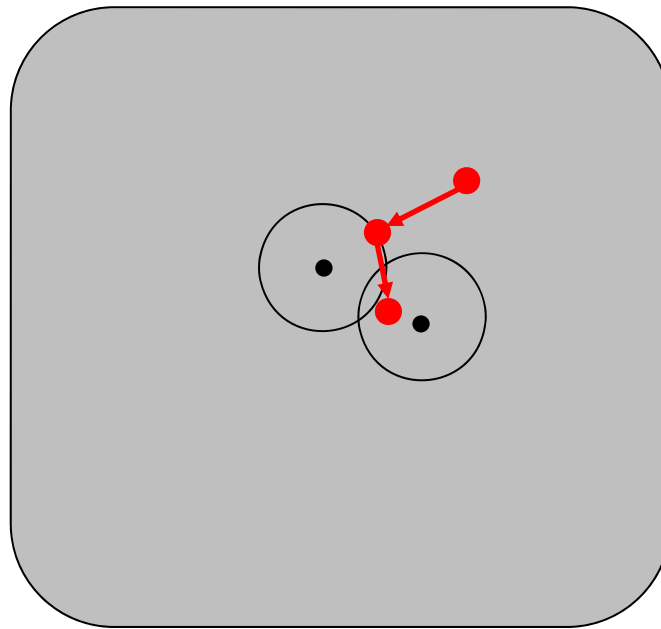
- > As the next arrow fires, we update our prior via the posterior distribution.
- > We iterate on this until our final target is chosen.



W

Archery Example

- > As the next arrow fires, we update our prior via the posterior distribution.
- > We iterate on this until our final target is chosen.



- > This procedure is called Bayesian inference.
- > How is this used in the real world? (Remember lost keys ex.)

W

Frequentist Estimation of Heads in a Coin Flip

- > We will flip a coin N times. We count the number of heads and want to estimate the $p(H)$. E.g. if it is a fair coin, we would expect (with enough trials) that we would estimate $p(H) = 0.5$.
- > Frequentist probability:
 - Most likely (maximum likelihood) answer would be:

$$p(H) = \frac{n(H)}{N}$$



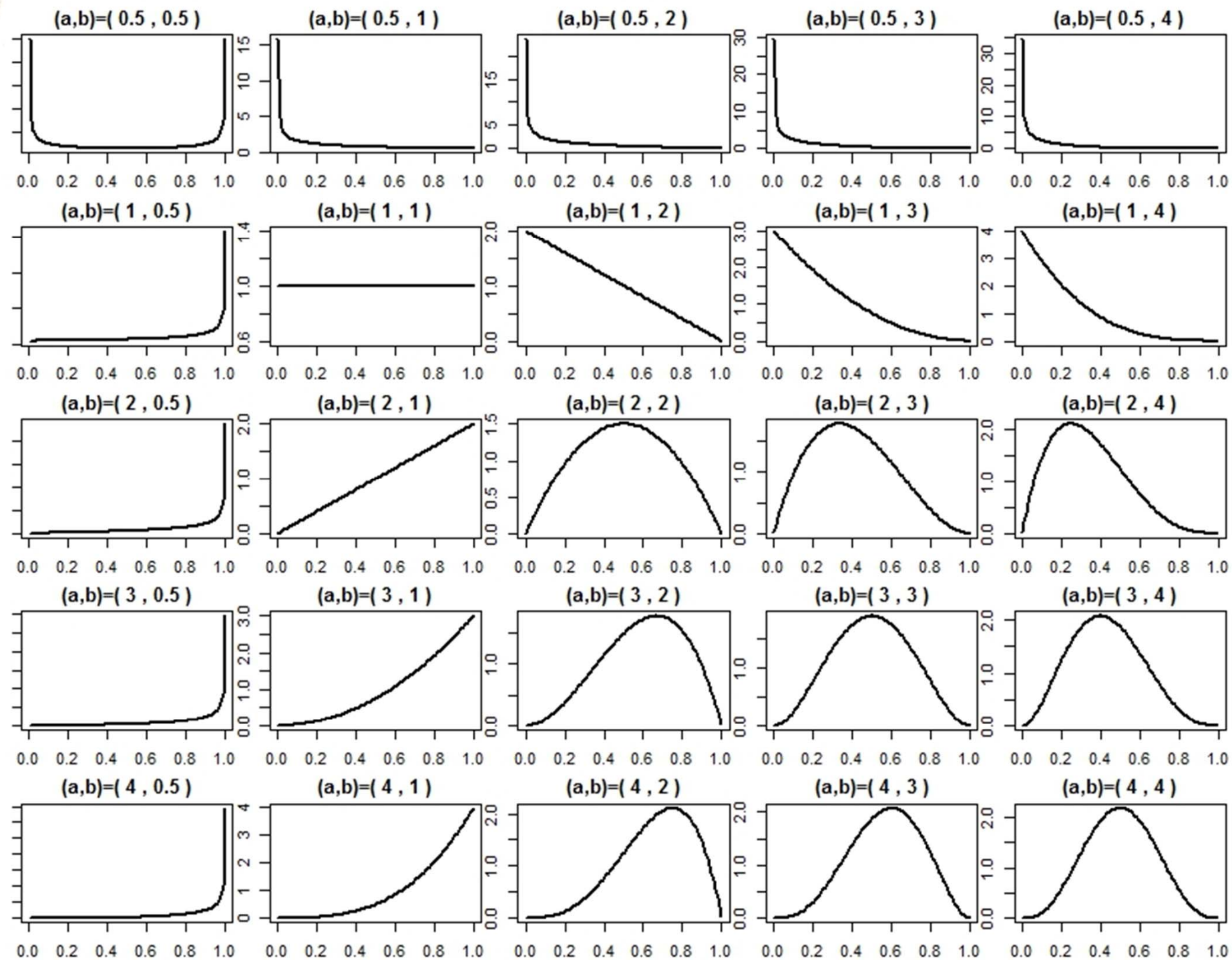
Bayesian Estimation of Heads in a Coin Flip

- > We will need to define a Prior probability for the estimation of $p(H)$.
- > The best choice in this case is a Beta distribution:

$$f(x) = x^{a-1}(1-x)^{b-1} \cdot (\textit{normalizing constant})$$

- > a, b are constants that define the distribution (similar to how the mean and variance define different Normals).
- > X is only defined between 0 and 1.





Bayesian Estimation of a Coin Flip Probability

$$P(\text{parameters}|\text{data}) = P(\text{data}|\text{parameters}) \frac{P(\text{parameters})}{P(\text{data})}$$

$$f(x) = x^{a-1}(1-x)^{b-1}. \text{ (normalizing constant)}$$

> After we choose a prior, we compute the posterior:

$$\text{Posterior} = \text{Likelihood} \frac{\text{Prior}}{P(\text{data})}$$

> Always a problem estimating the $P(\text{data})$. So...

$$\text{Posterior} = \text{Likelihood} \frac{\text{Prior}}{P(\text{data}|\text{all parameters})}$$

$$\text{Posterior} = \text{Likelihood} \frac{\text{Prior}}{\sum P(\text{data}|\theta)}$$



Bayesian Estimation of Multiple Parameters

- > We only had one parameter to estimate for the coin flip example, $p(H)$.
- > We created a grid to check (`seq(0.01,0.99,length=100)`) and used this to calculate the $p(\text{data})$, by checking all the values.
- > What if we had several parameters? If we had 6 parameters with a length 100 grid... $= 100^6 = 1,000,000,000,000 = 1 \text{ trillion points to check}$.
- > Maybe we don't have to sample everything, just enough points to understand and estimate the distribution of how $p(\text{data})$ behaves under the 6 parameters?

W

Markov Chain Monte Carlo

What is a Markov process?

- > A Markov process makes a transition from one state to other states with probability Π
 - Π only depends on the current state
 - Transition to one or more other states
 - Can 'transition' to current state
 - Π is a matrix of dim $N \times N$ for N possible states
- > A Markov process is a random walk



Markov Chain Monte Carlo

Markov chain is a sequence of Markov transition processes:

$$P[X_{t+1} = x | X_t = x_t, \dots, X_0 = x_0] = P[X_{t+1} = y | X_t = x_t]$$

‘Memoryless’ process

And

$\Pi =$

$P_{1,1}$	$P_{1,2}$...	$P_{1,N}$
$P_{2,1}$	$P_{2,2}$
...
$P_{N,1}$			$P_{N,N}$

W

Introducing the Metropolis (Hastings) Algorithm

- > The Metropolis algorithm is a specific MCMC algorithm.
- > Algorithm:
 - 1. Pick a starting point in your parameter space and evaluate it according to your model. (find $p(\text{data})$).
 - 2. Choose a nearby point randomly and evaluate this point.
 - > If the $p(\text{data})$ of the new point is greater than your previous points, accept new point and move there.
 - > If the $p(\text{data})$ of the new point is less than your previous point, only accept with probability according to the ratio: $p(\text{data new}) / p(\text{data old})$.
 - 3. Repeat # 2 many times.



Introducing the Metropolis (Hastings) Algorithm

- > M-H algorithm eventually converges to the underlying distribution.
- > We only have to visit N points, not 1 Trillion points.
- > There is high serial correlation in M-H chain, which slows convergence
- > Need to 'tune' the probability distribution used to find the next point
 - E.g. if we use Normal distribution need to pick σ .
 - If σ is too small chain will only search the space slowly.
 - If σ is too big, higher serial correlation



Gibbs Sampling

Improved version of M-H algorithm

- > Uses systematic sampling of the parameter space
- > Example: round-robin
 - With N dimensions
 - Sample 1, 2, ..., N and then start over again
 - Transition still based on $p(\text{data})$
- > Reduces serial correlation and improves convergence



Remember Bayes Law: $P(A|B) = P(B|A) \frac{P(A)}{P(B)}$

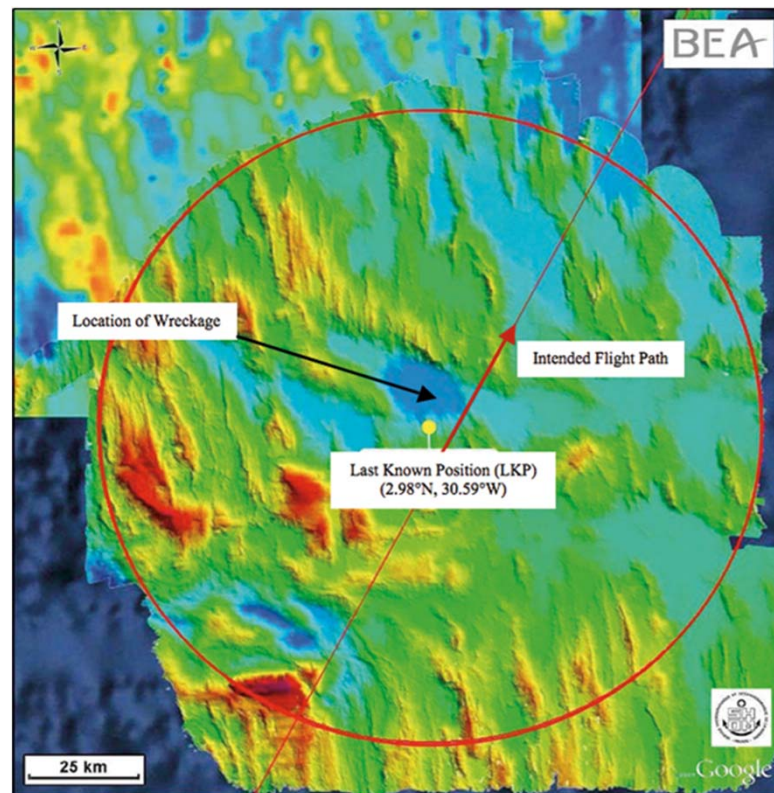
- > Tests are not the event. We have a disease test, which is different than the event of actually having the disease.
- > Tests are flawed. Tests have false positives and false negatives.
- > Tests return test probabilities, not the event probabilities.
- > False positives skew results.
 - E.g. If fraud is rare, then the likelihood of a positive result of fraud is probably due to a false positive

W

Reading assignment: Bayesian Inference Successes

$$P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters})$$

- > Bayesian inference used to successfully find lost planes. E.g. Air France 447
- > <https://www.informs.org/ORMS-Today/Public-Articles/August-Volume-38-Number-4/In-Search-of-Air-France-Flight-447>



W

Assignment

> Probability of texting.

- You are asked to compute the probability that the driver of a car is texting at a specific intersection.
- Nationally the chance that a driver is texting is:
 - > $p = 0.5$, at $x = 0.1$
 - > $P = 0.75$ at $x = 0.3$
- You observe cars at a location three times and note the number of texting drivers:
 - 2 texting out of 20 drivers
 - 4 texting out of 20 drivers
 - 1 texting out of 20 drivers



Assignment continued

> Given these data

- Compute the Beta prior
- Plot the prior, likelihood and posterior three times as you update your belief based on collecting more data
- Simulate the final posterior distribution and do the following:
 - > Plot the posterior with the 90% HDI shown
 - > Report the upper and lower limits of the 90% HDI
 - > Of the next hundred drivers what are the number of texting drivers in the 90% HDI?
 - > Are the drivers in this area better or worse than the national figures indicate?

> Turn in:

- A clear report, **stating your conclusions up-front** in and supporting the conclusions with evidence
- R script to professional standards

