AM207

https://am207.info/

When?

Monday 12pm - 1.15pm, Lecture. Compulsory to attend. NW B101.

Wednesday 12pm - 1.15pm, Lecture. Compulsory to attend. NW B101.

Fridays 12pm - 1.15pm Lab. Compulsory to attend. Pierce 301.

Who

Instructor:

Rahul Dave

TFs:

- Patrick Ohiomoba
- Srivatsan Srinivasan
- Zongren Zou

This is a course on:

Stochastic Optimization or Derivatives

Stochastic Expectations or Integration

in the service of

Modeling and Inference

Why take this course?

- learn how to think in principled ways of modeling..why..not just how..
- ..using bayesian statistics which is far more natural, and which has applications in almost every field
- understand deeply how and why machine leaning works
- learn generative models so that you can understand NNs, GANs better

- learn how to regularize models
- deal with data computationally large/small and statistically small/large
- learn how to optimize objective functions such as loss functions using Stochastic Gradient Descent and Simulated annealing
- Perform sampling and MCMC to solve a variety of problems
- Learn how and when to use parametric and non-parametric stochastic processes

What sorts of problems?

- machine learning hyperparameter optimization
- generalize A/B testing using Bandits (eg see https://support.google.com/analytics/answer/2844870?hl=en)
- generative modeling of images (see https://blog.openai.com/generative-models/)
- many problems in psychology, ecology, phylogenetics, public policy, etc

$$\mathbf{A}_t = \begin{array}{l} \\ \text{not yet entered} \\ \text{dead} \\ \text{offsite} \\ \text{onsite} \end{array} \begin{array}{c} \begin{array}{cccc} \text{not yet entered} & \text{dead} & \text{offsite} & \text{onsite} \end{array} \\ \begin{pmatrix} 1-\psi_t & 0 & 0 & 0 \\ 0 & 1 & 1-\phi_t & 1-\phi_t \\ \psi_t(1-\lambda_t) & 0 & \phi_t\gamma_t' & \phi_t\gamma_t'' \\ \psi_t\lambda_t & 0 & \phi_t(1-\gamma_t') & \phi_t(1-\gamma_t'') \end{pmatrix} \end{array}$$

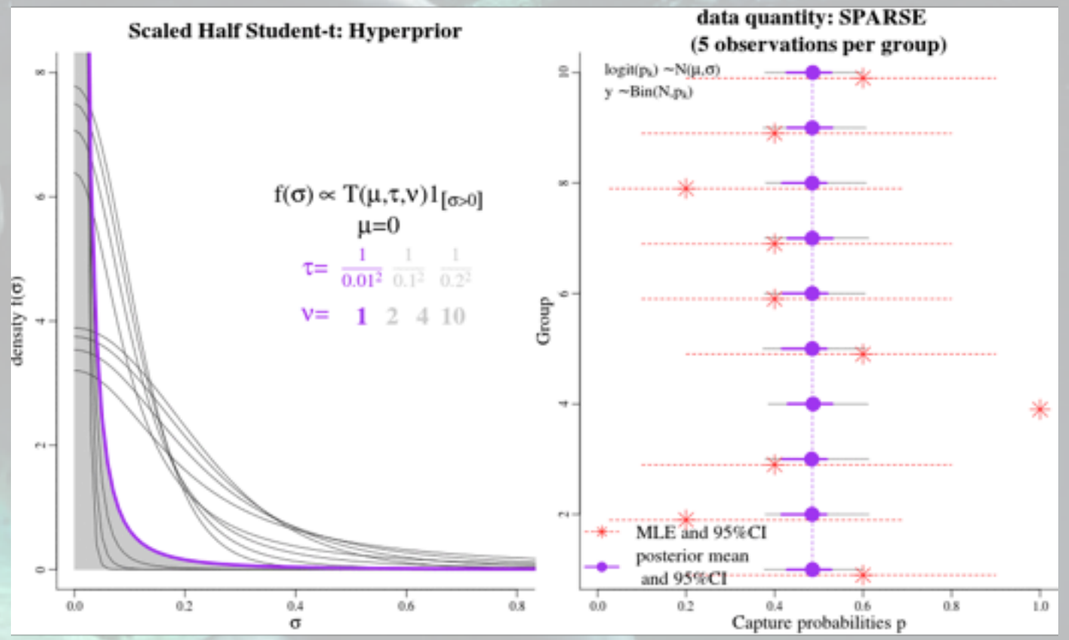
$$\mathbf{B}_{t,s} = \begin{array}{l} \\ \text{observed} \\ \text{unobserved} \end{array} \begin{array}{c} \begin{array}{cccc} \text{not yet entered} & \text{dead} & \text{offsite} & \text{onsite} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & p_{t,s} \\ 1 & 1 & 1 & 1-p_{t,s} \end{pmatrix} \end{array} \quad (1)$$

The most general model is represented as:

$$\begin{aligned} \text{initialize: } z_{0,i} &= 1 \text{ for } i = 1, \ldots, m \\ p(z_{t,i}|z_{t-1,i}, \mathbf{A}_t) &= \text{Cat}(\mathbf{A}_t[\cdot, z_{t-1,i}]) \text{ for } i = 1, \ldots, m; \\ & \quad t = 1, \ldots, T \\ p(y_{t,s,i}|z_{t,i}, \mathbf{B}_{t,s}) &= \text{Cat}(\mathbf{B}_{t,s}[\cdot, z_{t,i}]) \text{ for } i = 1, \ldots, m; \\ & \quad s_t = 1, \ldots, S_t;\ t = 1, \ldots, T \end{aligned}$$

$$\pi(\{\mathbf{A}\}_t^T, \{\mathbf{B}\}_s^{S_t}|\mathbf{Y}, \Lambda) \propto \left(\prod_i^m \left(\prod_{t=1}^T \left(\prod_{s_t=1}^{S_T} p(y_{t,s,i}|z_{t,i}, \mathbf{B}_{t,s})\right)\right.\right.$$
$$\left.\left. p(z_{t,i}|z_{t-1,i}, \mathbf{A}_t)\right)\right)\pi(\Lambda) \quad (2)$$

Bayesian Hierarchical Mark-Recapture Models (see https://www.frontiersin.org/articles/10.3389/fmars.2016.00025/full)



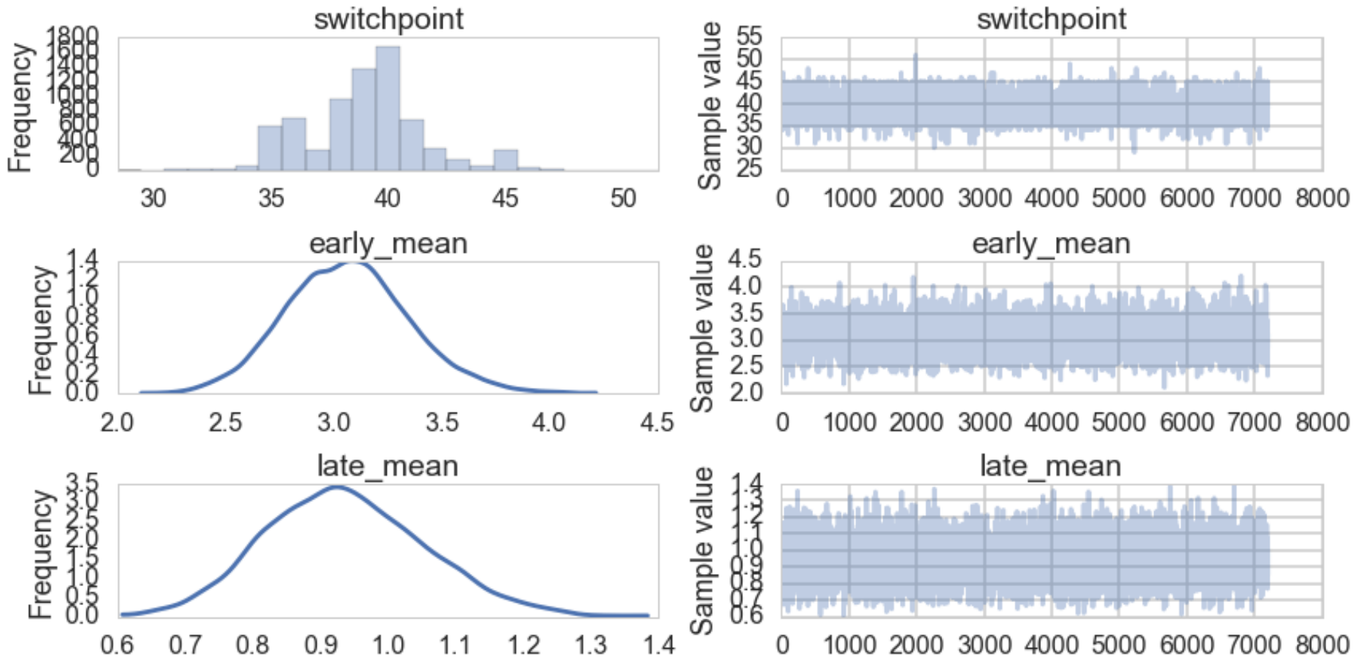
Why not?

- this is a hard course. you will have to work hard. especially on your own
- there is a lot of homework
- you do not have the requisite background
- you are a statistics expert

Modules

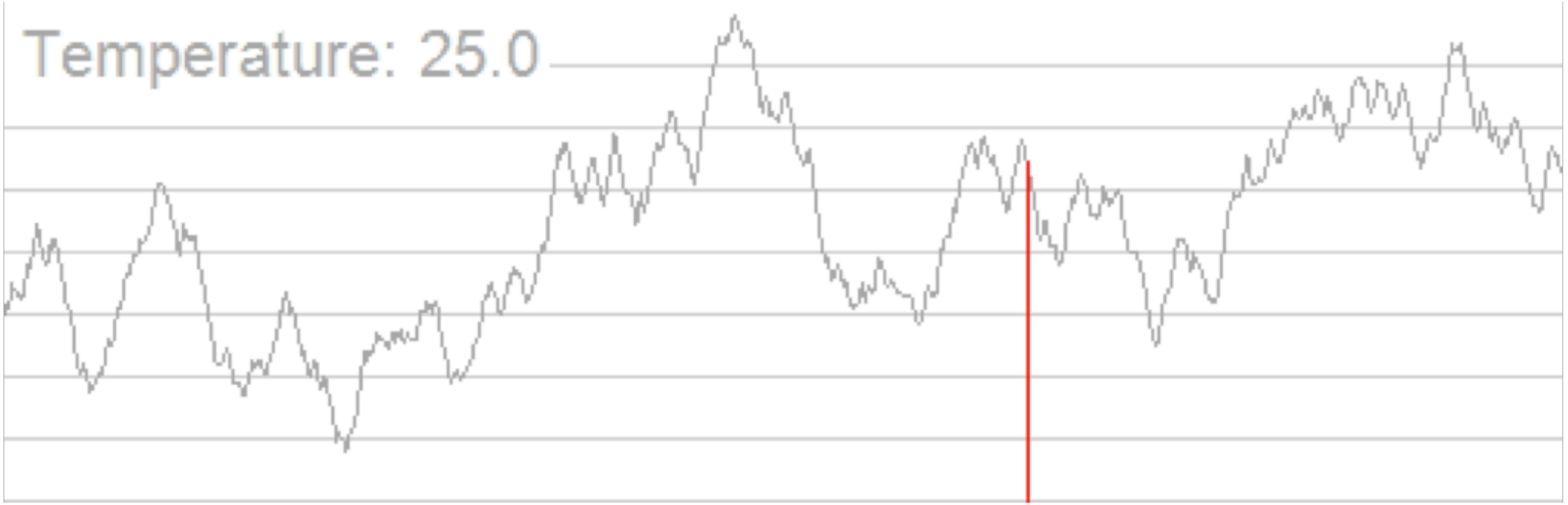
- stats review and sampling
- optimization and machine learning; stochastic optimization
- Bayesian concepts and density estimation
- MCMC and other algorithms to obtain posteriors
- Bayesian regression and glms
- Model checking, comparison, and selection
- Variational Bayes
- EXTRA: Time dependent, non-iid models, Non-parametric Bayes, or Autoencoders

sampling

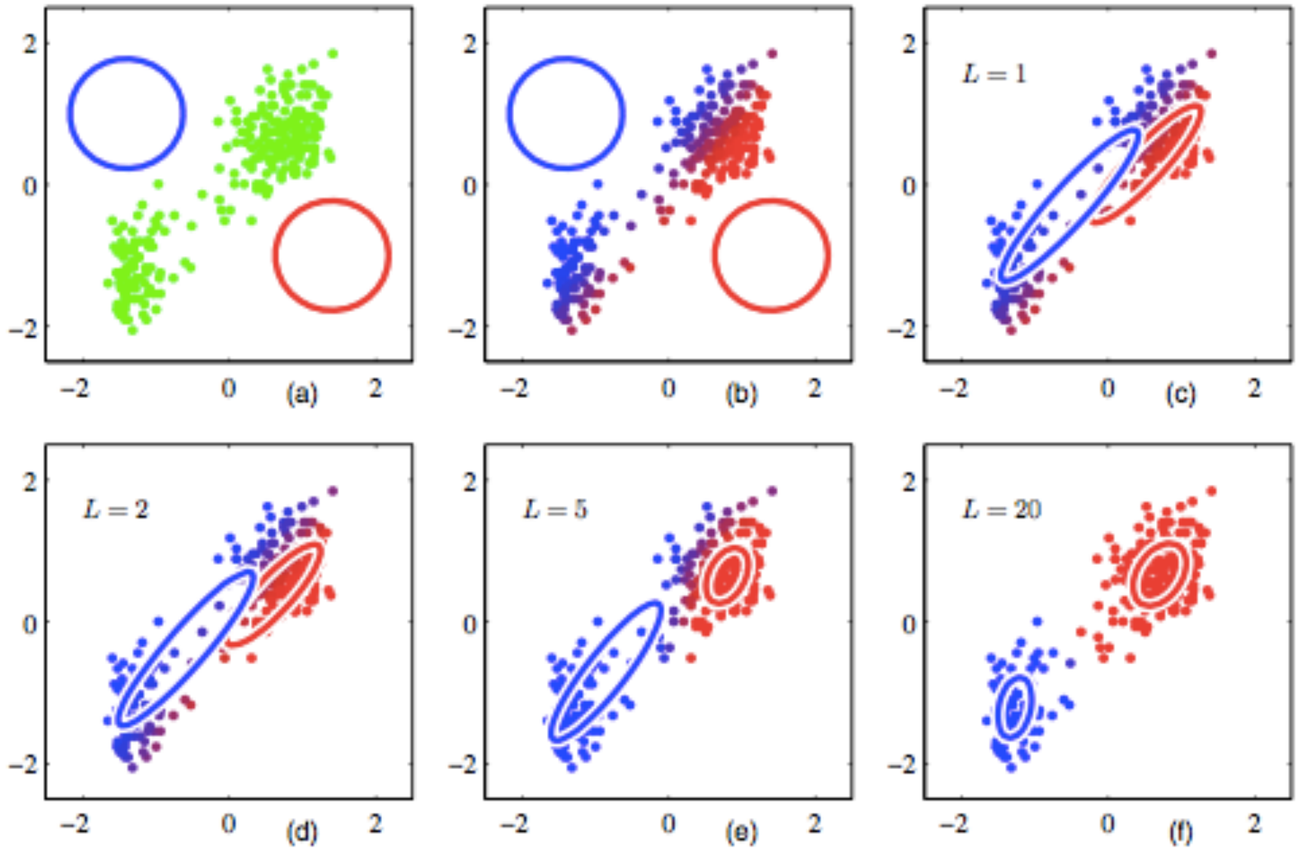


Coal Disasters Switchpoint Model

optimization



(Simulated Annealing, Wikipedia)



Learns a generative model! Unsupervised learning! (EM, Bishop)

Differentiation vs Integration

- optimize a loss function: SGD, EM, etc

OR

- calculate an Expectation or a marginalization: numerical integration, monte carlo, MCMC
- two sides of the same coin

Bayesian statistics

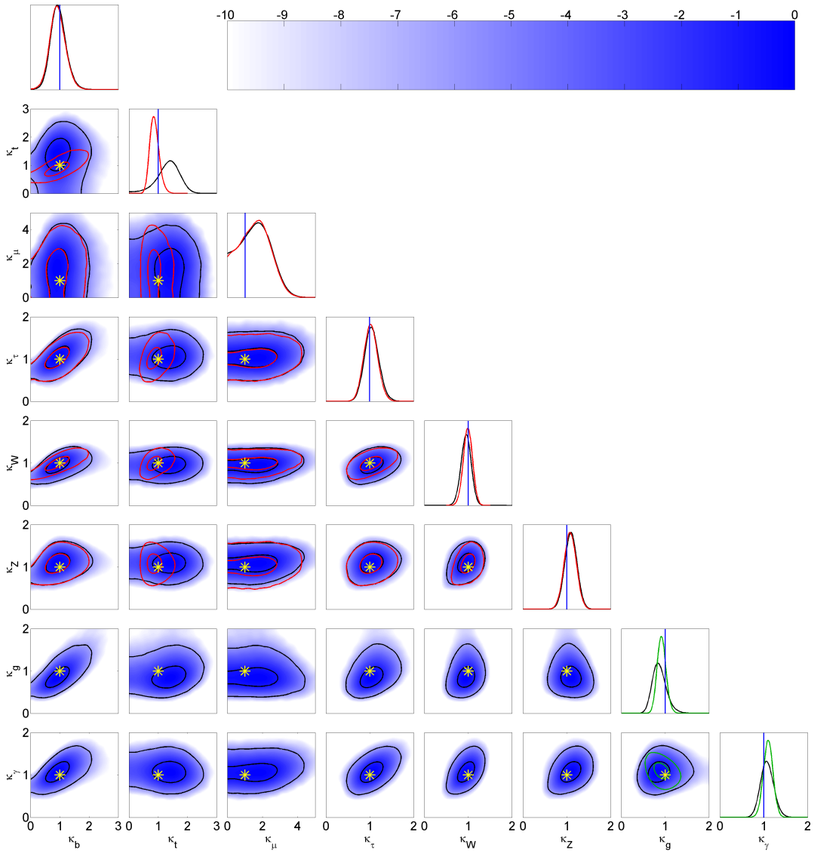
Small world:

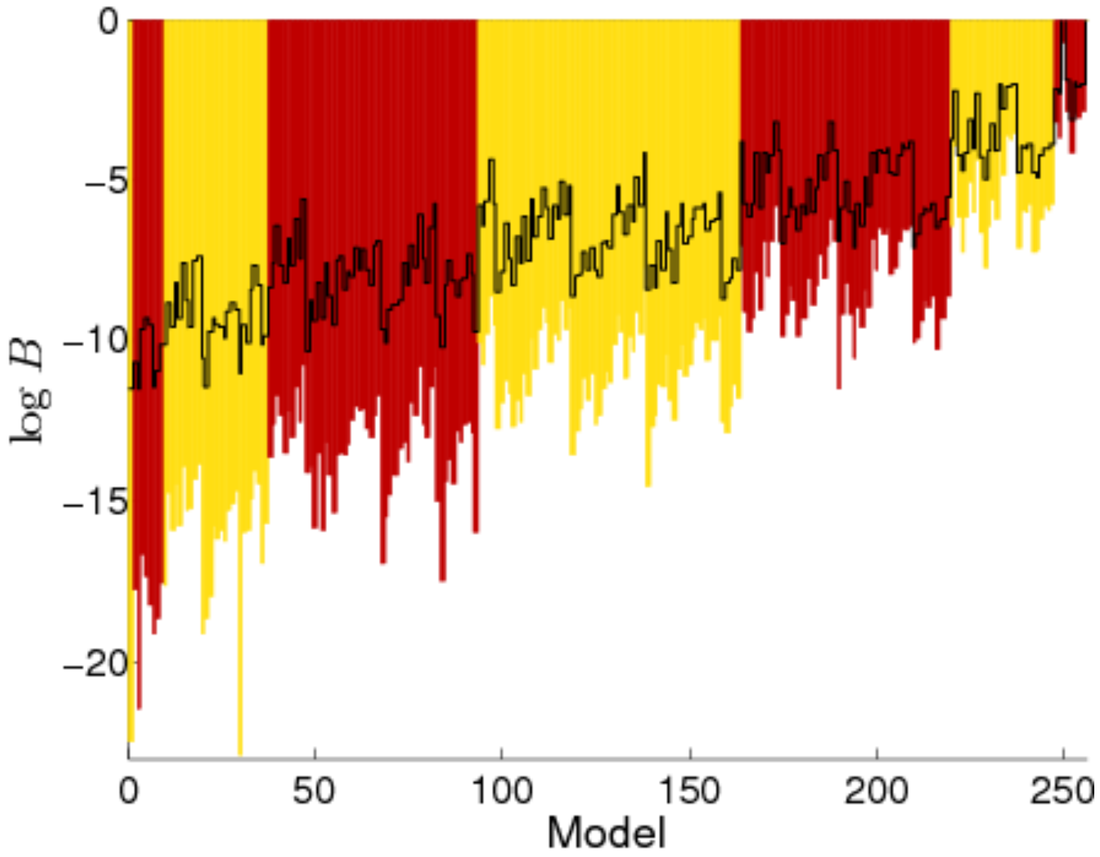
$$P(\theta \mid D) = \frac{P(D \mid \theta) \times P(\theta)}{P(D)}$$

Big World:

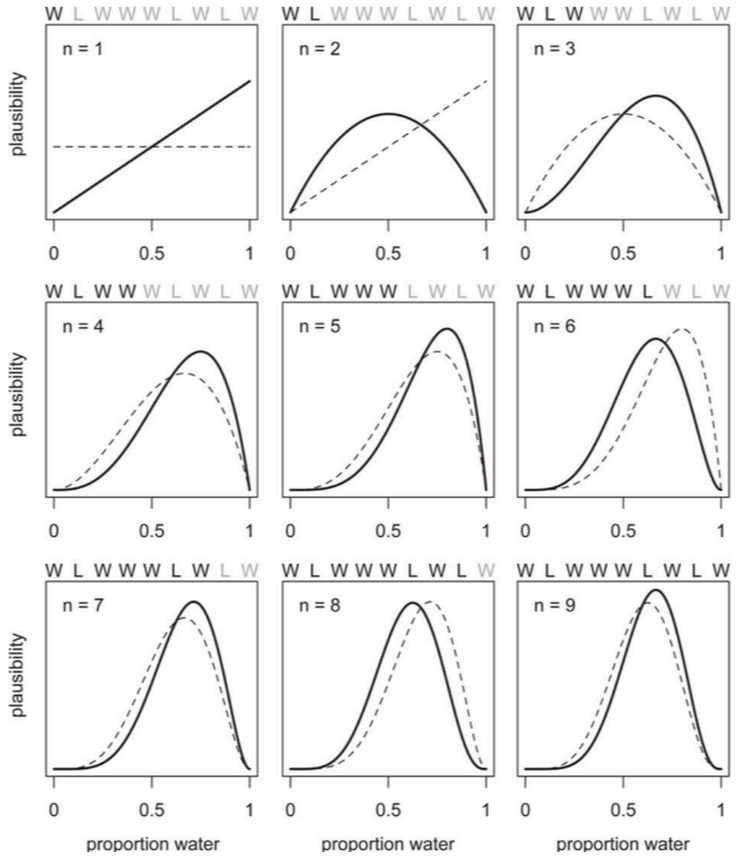
$$P(M \mid D) = \frac{P(D \mid M) \times P(M)}{P(D)}$$

Bayesian Analysis for Higgs

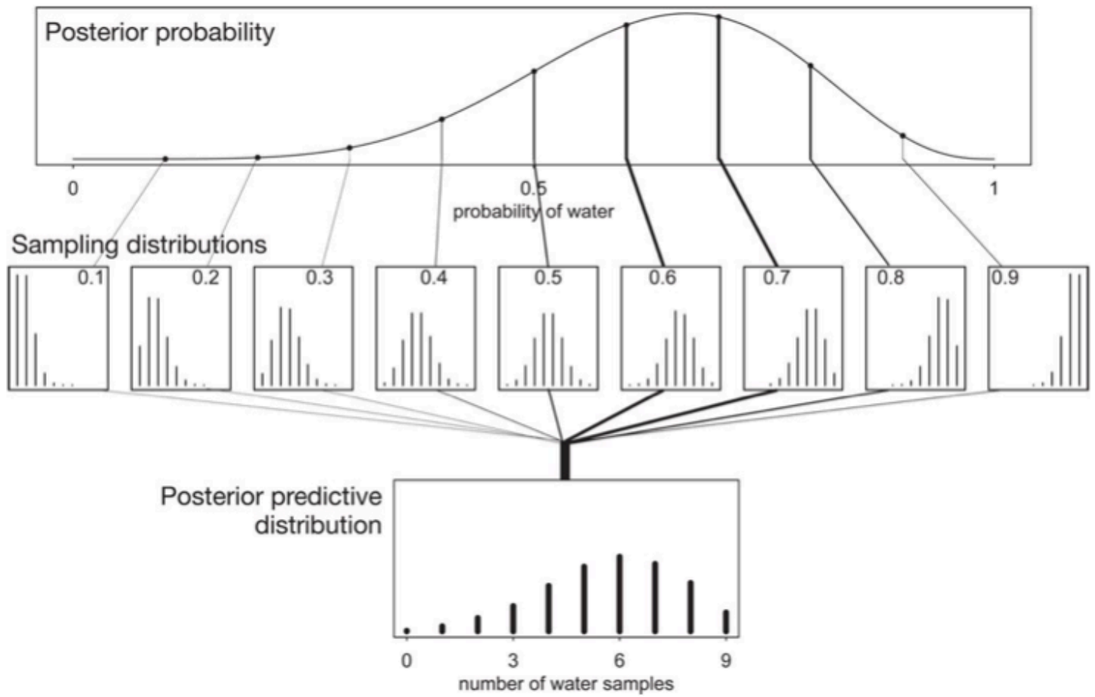


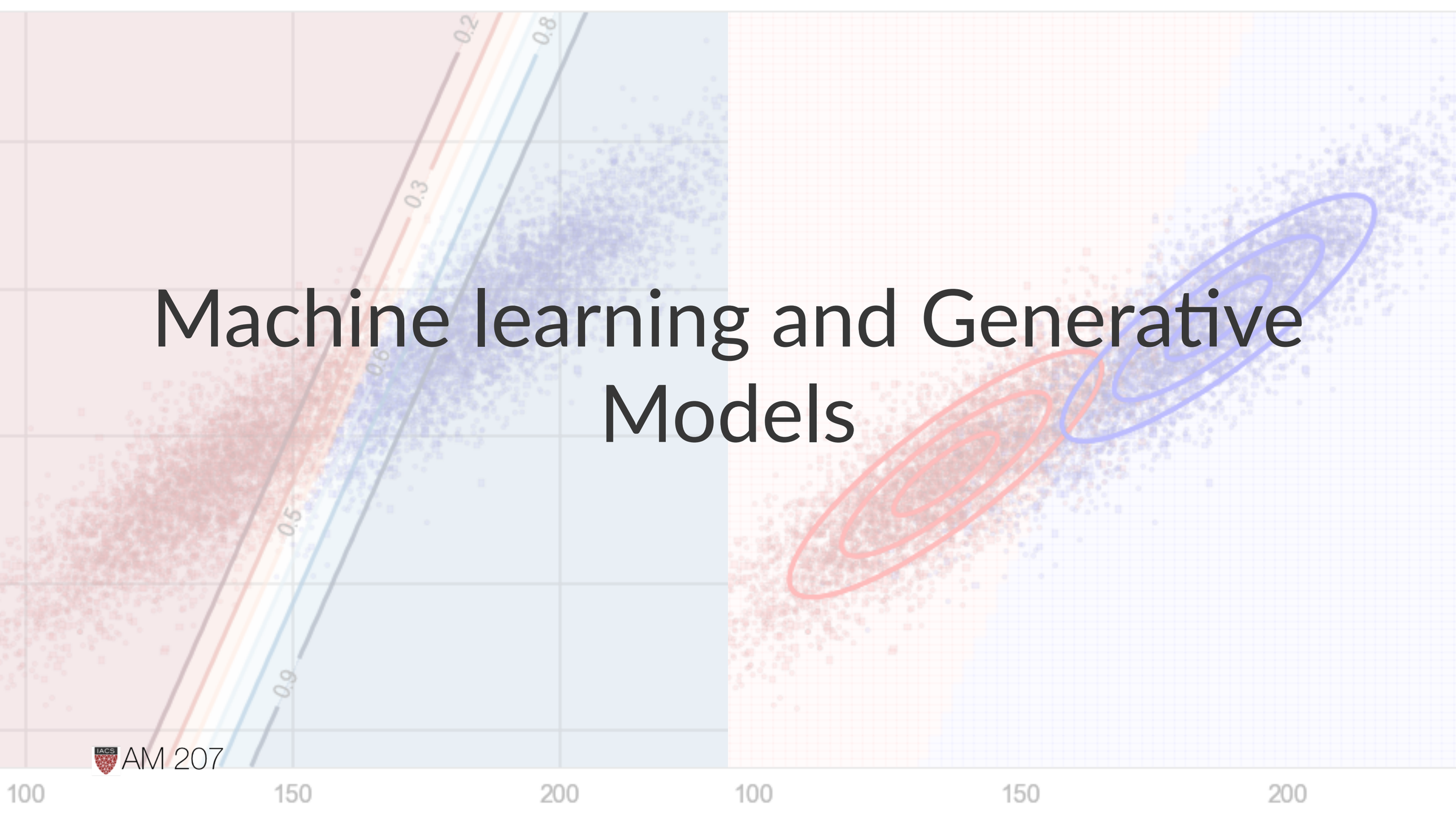


Posterior, updated



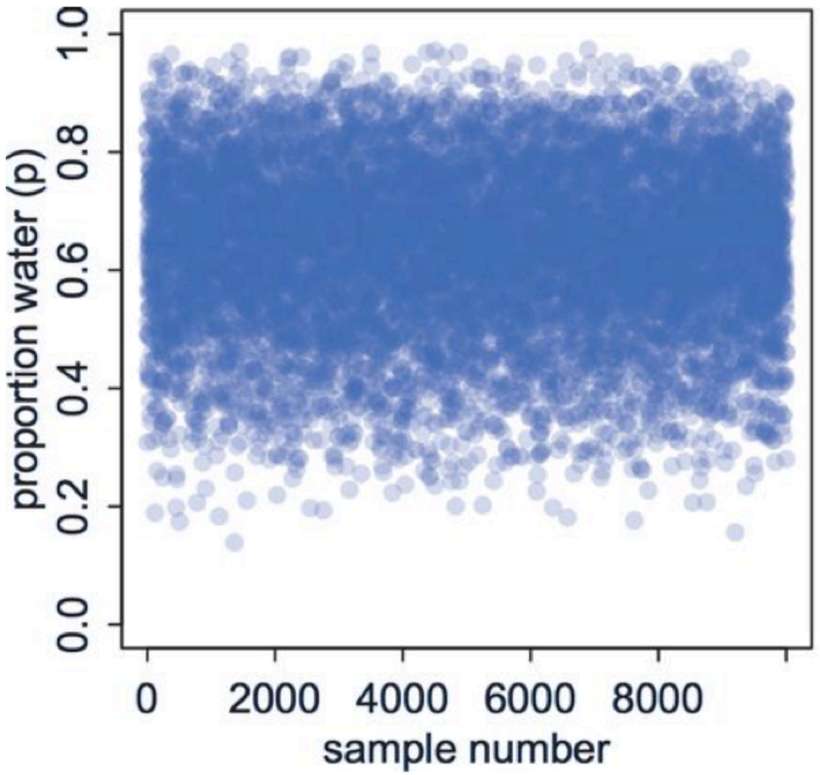
Posterior Predictive

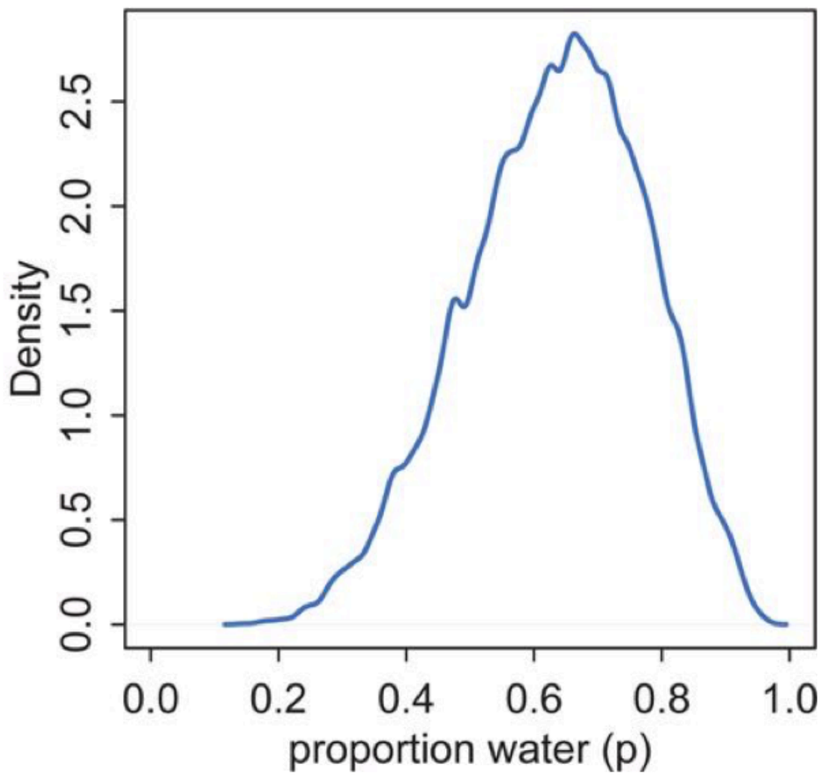




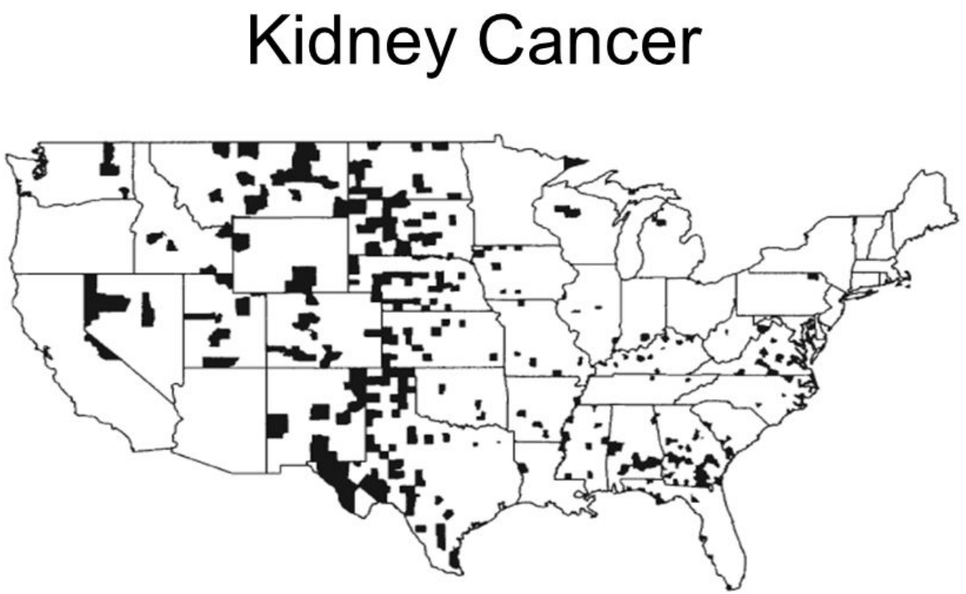
Machine learning and Generative Models

MCMC


proportion water (p)
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
sample number


Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
proportion water (p)

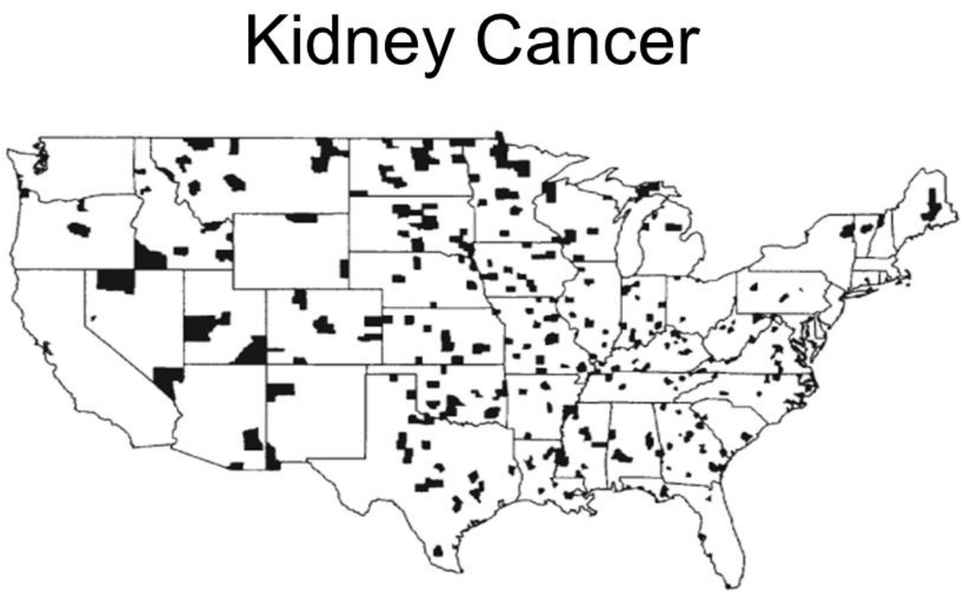
Whats up with these counties?



Counties with the lowest kidney cancer death rates

Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004.

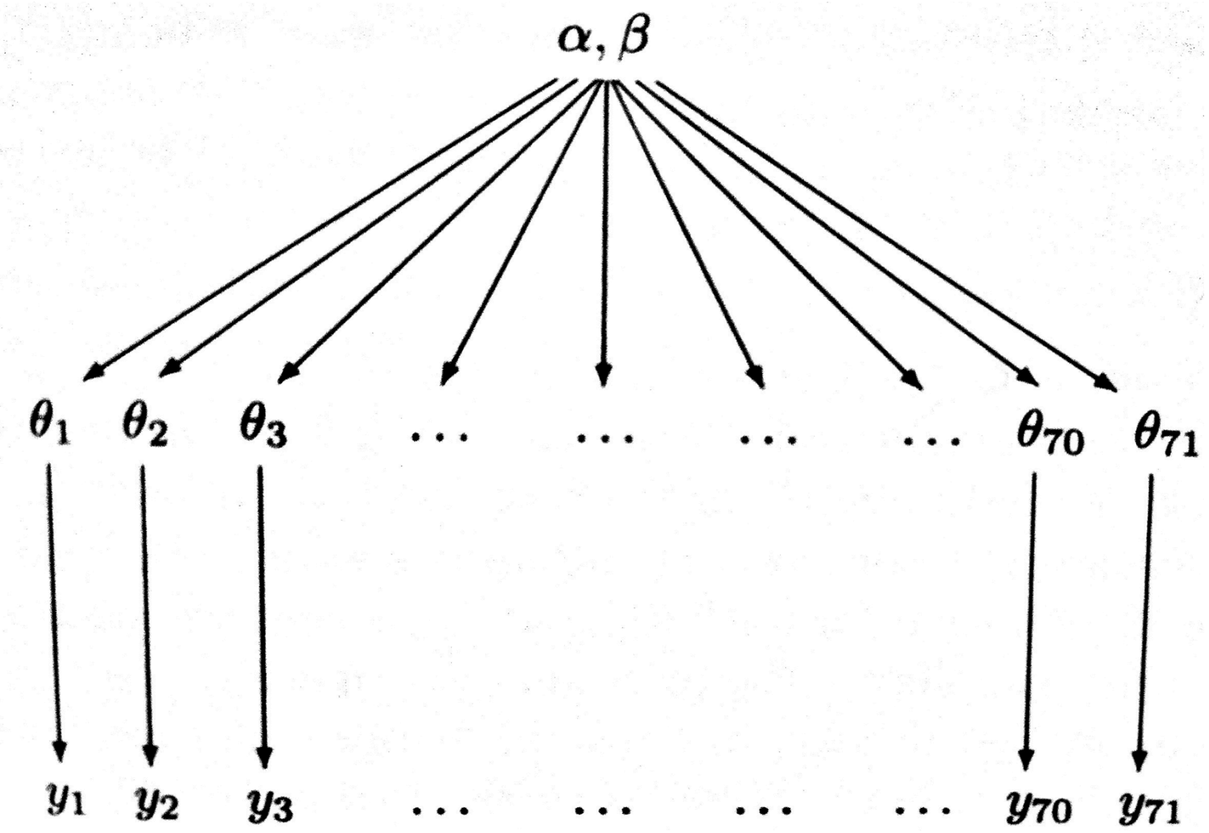
And with these?



Counties with the highest kidney cancer death rates

Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004.

Hierarchical Model with regularization



glms

Monks in monastery i (indicator x_i) produce y_i manuscripts a day.

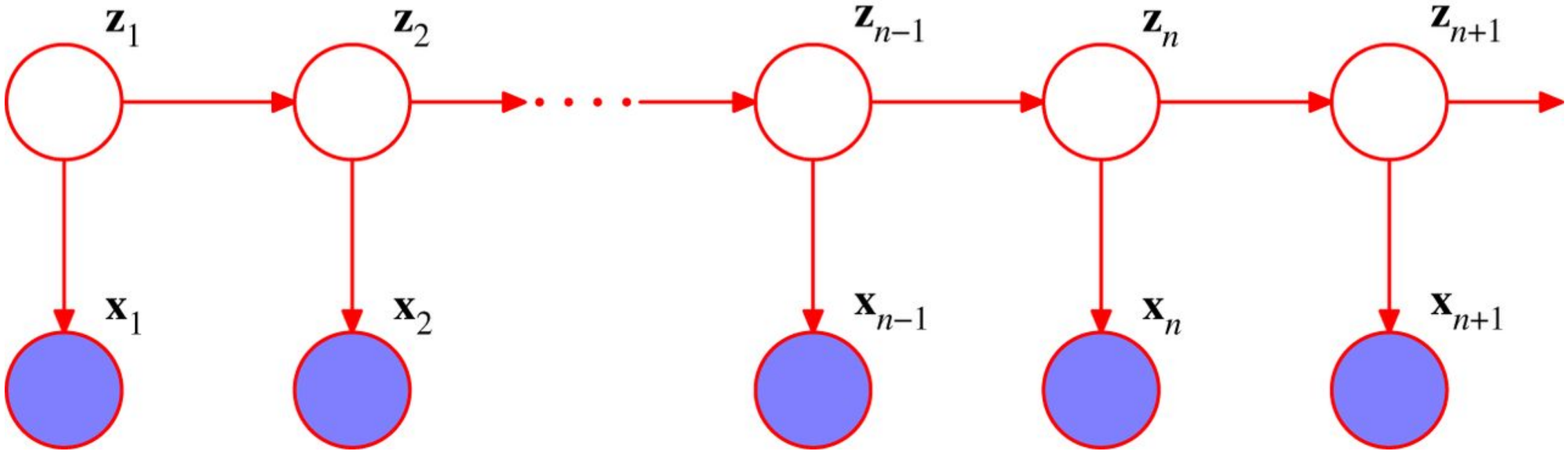
Poisson likelihood and logarithmic link

Model:

$$y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \alpha + \beta x_i$$

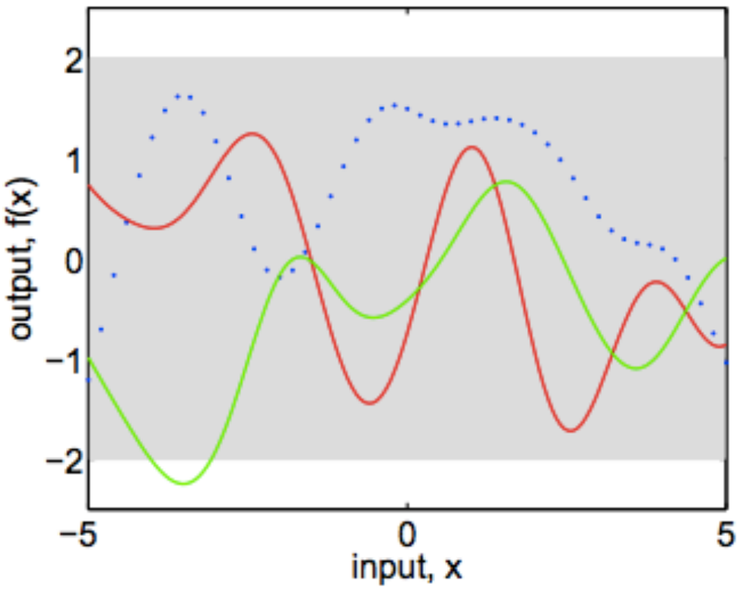
dynamical systems



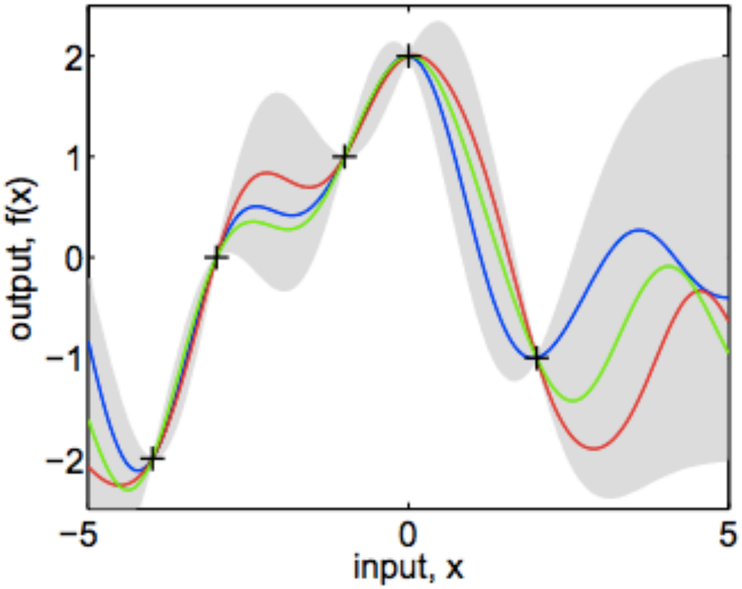
hidden markov models

gaussian processes

nonparametric, prior on functions...



(a), prior



(b), posterior

Concepts running through:

Hidden Variables

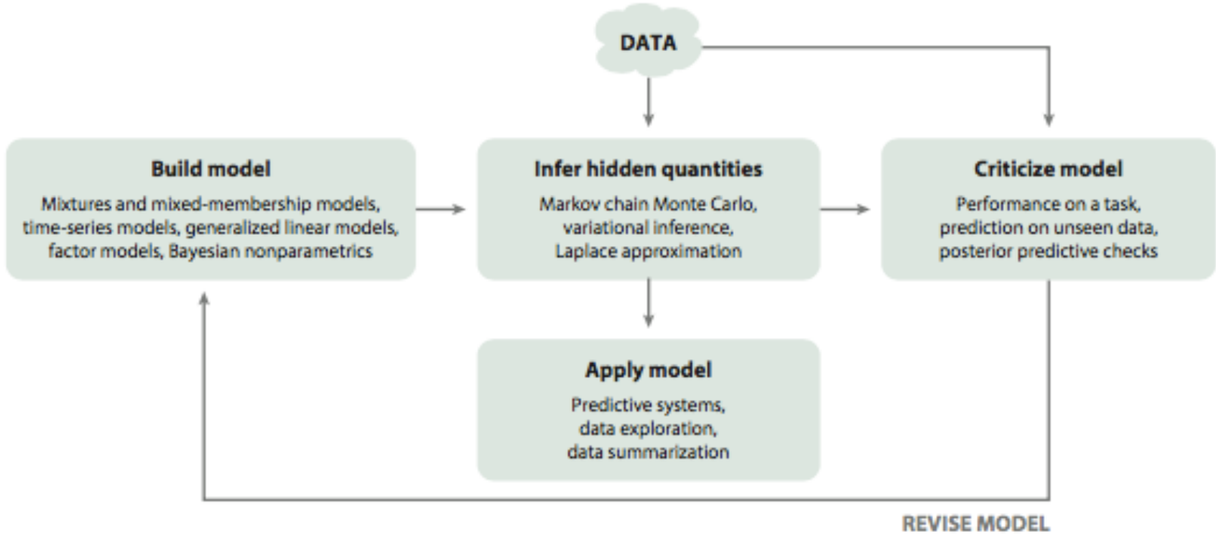
Marginalizing over nuisance parameters

Differentiation vs Integration

Frequentist vs Bayesian

Generative Models

Overall concept: Box's Loop



(image from David Blei's paper on hidden variables)

Requirements

- you will need to know how to program numerical python
- you will need to have a background in stats and simple distributions at least although we will review concepts whenever needed. Its better when you are reviewing concepts than learning it for the first time
- you should be comfortable with matrix manipulations and calculus. You should have a passing knowledge of multivariate calculus.

What kind of course?

- grad level course though nothing is really grad level hard
- if you have machine learning background you will make a lot of mental connections.
- i am your emcee; its my job to incorporate info and understanding from various places.
- probably harder than cs181 but simpler than cs281. Ideal in-between course.

Structure of the course

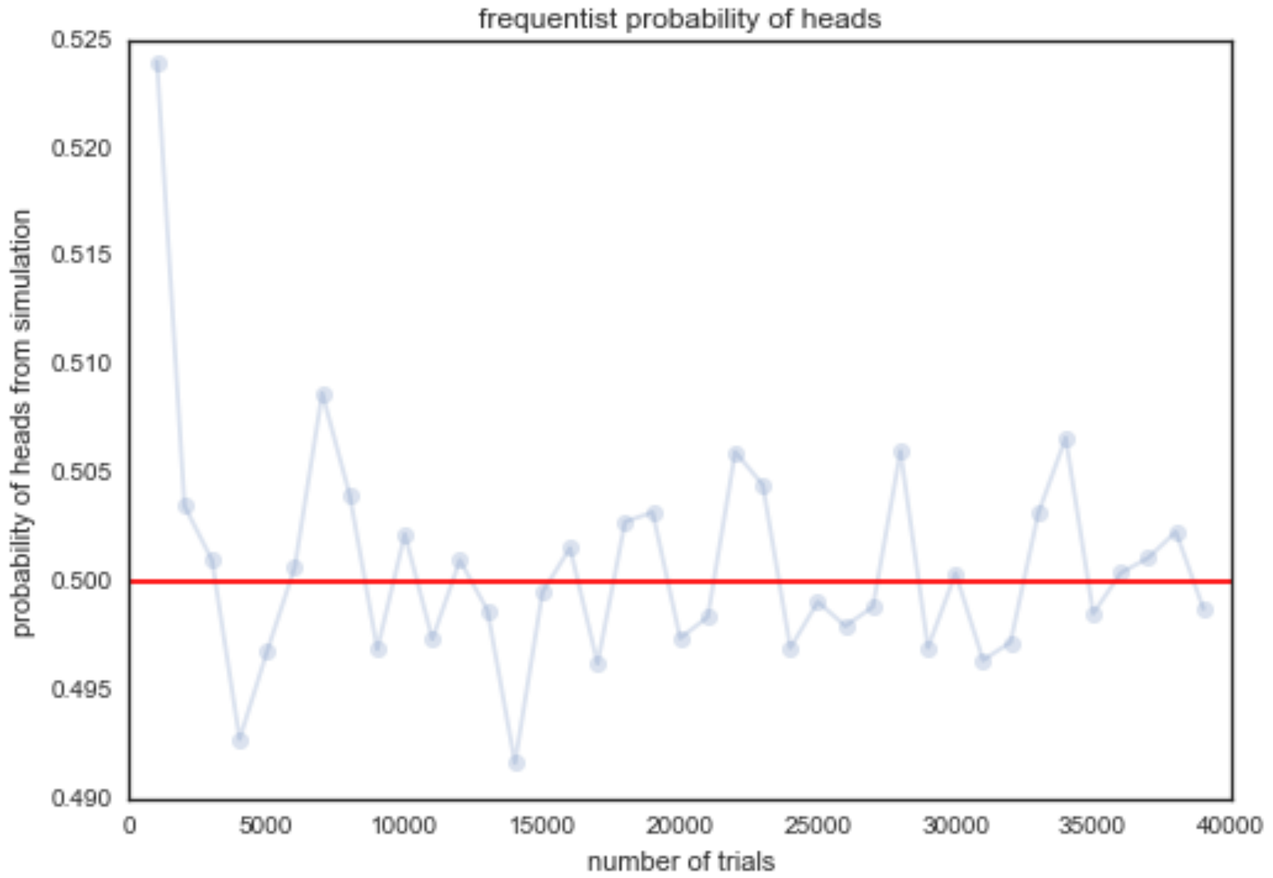
- lectures (2 per week), compulsory
- lab (you will play), compulsory
- homework (every week)
- paper
- final exam (a glorified project-ish homework)
- readings

- there will be readings most weeks, some made available a lecture or two ahead
- preliminary notes will made available a lecture or two ahead..you should read these before class
- notes will be updated towards the time of the lecture
- lecture slides will be made available just before or after the lecture

- homework will be made available every week Fri evening or Saturday Morning; is due every week friday 11.59pm. should take 7-8 hours
- expect another 6-7 hours of reading, including both before and after lecture

Probability

- from symmetry
- from a model, and combining beliefs and data: Bayesian Probability
- from long run frequency



Random Variables

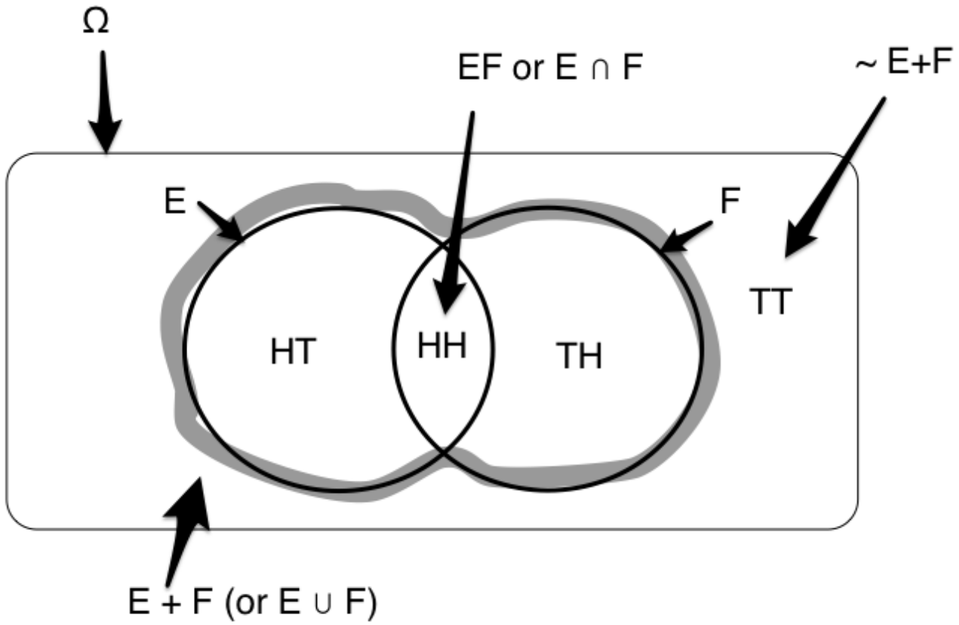
Definition. A random variable is a mapping

$$X : \Omega \to \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω.

- Ω is the sample space. Points
- ω in Ω are called sample outcomes, realizations, or elements.
- Subsets of Ω are called Events.

An example with 2 coin tosses



- E is the event of getting a heads in a first coin toss, and F is the same for a second coin toss.
- Ω is the set of all possibilities that can happen when you toss two coins: {HH,HT,TH,TT}

Fundamental rules of probability:

1. $p(X) >= 0$; probability must be non-negative
2. $0 \leq p(X) \leq 1$
3. $p(X) + p(X^{-}) = 1$ either happen or not happen.
4. $p(X + Y) = p(X) + p(Y) - p(X, Y)$

- Say $\omega = HHTTTTHTT$ then $X(\omega) = 3$ if defined as number of heads in the sequence ω.
- We will assign a real number P(A) to every event A, called the probability of A.
- We also call P a probability distribution or a probability measure.

A Murder Mystery

(from the book: Model Based Machine Learning)

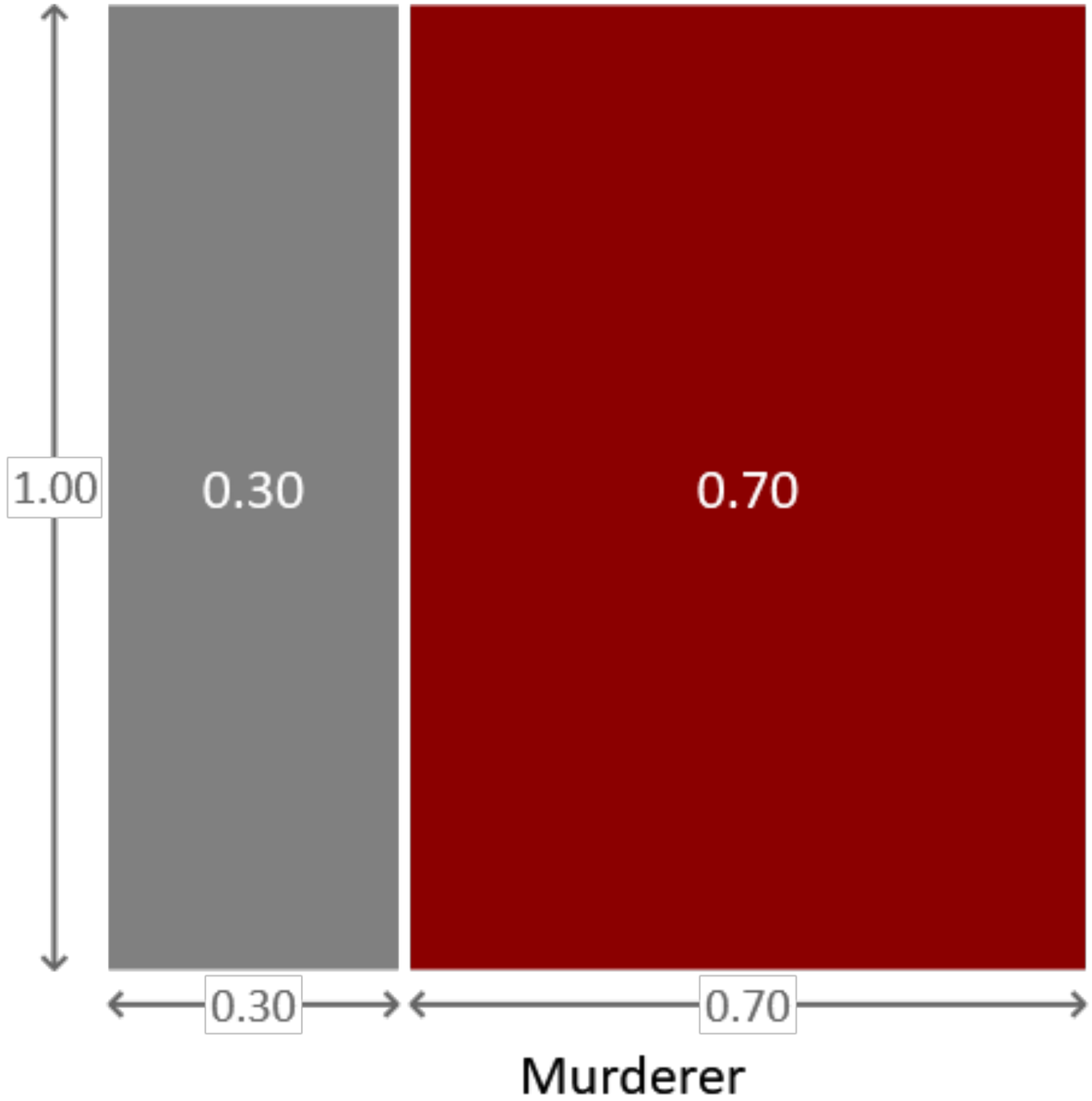
"As the clock strikes midnight in the Old Tudor Mansion, a raging storm rattles the shutters and fills the house with the sound of thunder. The dead body of Mr Black lies slumped on the floor of the library, blood still oozing from the fatal wound. Quick to arrive on the scene is the famous sleuth Dr Bayes, who observes that there were only two other people in the Mansion at the time of the murder. So who committed this dastardly crime? Was it the fine upstanding pillar of the establishment Major Grey? Or was it the mysterious and alluring femme fatale Miss Auburn?

We represent the murderer with a random variable `murderer` whose value we dont know. This variable equals either Auburn or Grey.

Priors

"From what we know about our two characters, we might think it is unlikely that someone with the impeccable credentials of Major Grey could commit such a heinous crime, and therefore our suspicion is directed towards the enigmatic Miss Auburn. Therefore, we might assume that the probability that Miss Auburn committed the crime is 70%, or equivalently 0.7."

$$p(murderer = Auburn) = 0.7$$



Bernoulli Distribution

The "prior" distribution for `murder` is the Bernoulli.

Say a coin flip represented as X, where $X = 1$ is heads, and $X = 0$ is tails. The parameter is probability of heads p.

$$X \sim Bernoulli(p)$$

is to be read as X **has distribution** $Bernoulli(p)$.

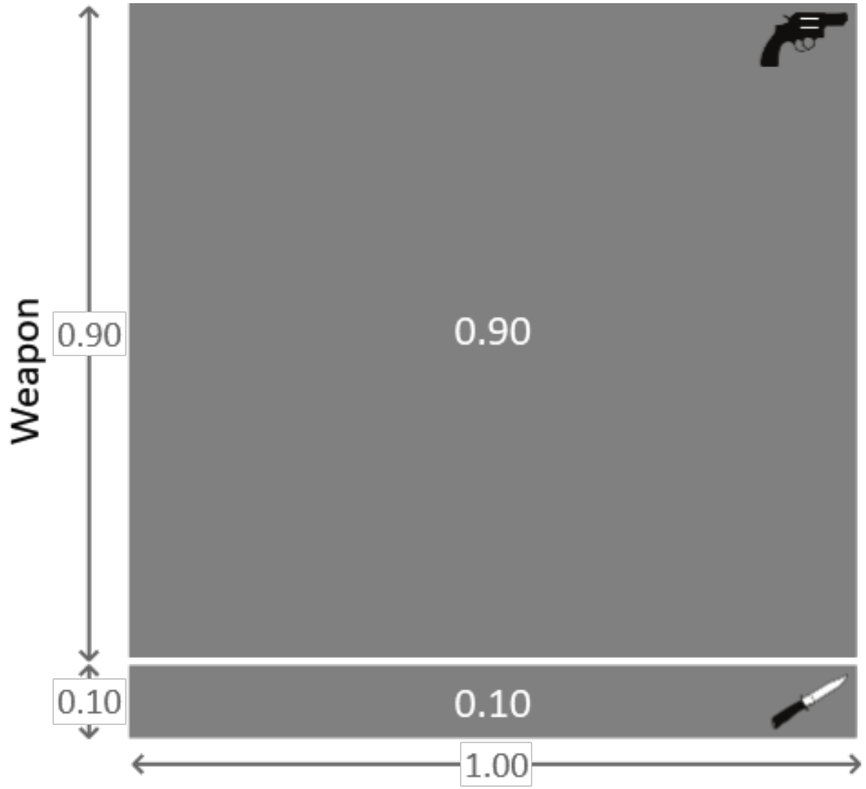
"Dr Bayes searches the mansion thoroughly. He finds that the only weapons available are an ornate ceremonial dagger and an old army revolver. “One of these must be the murder weapon”, he concludes."

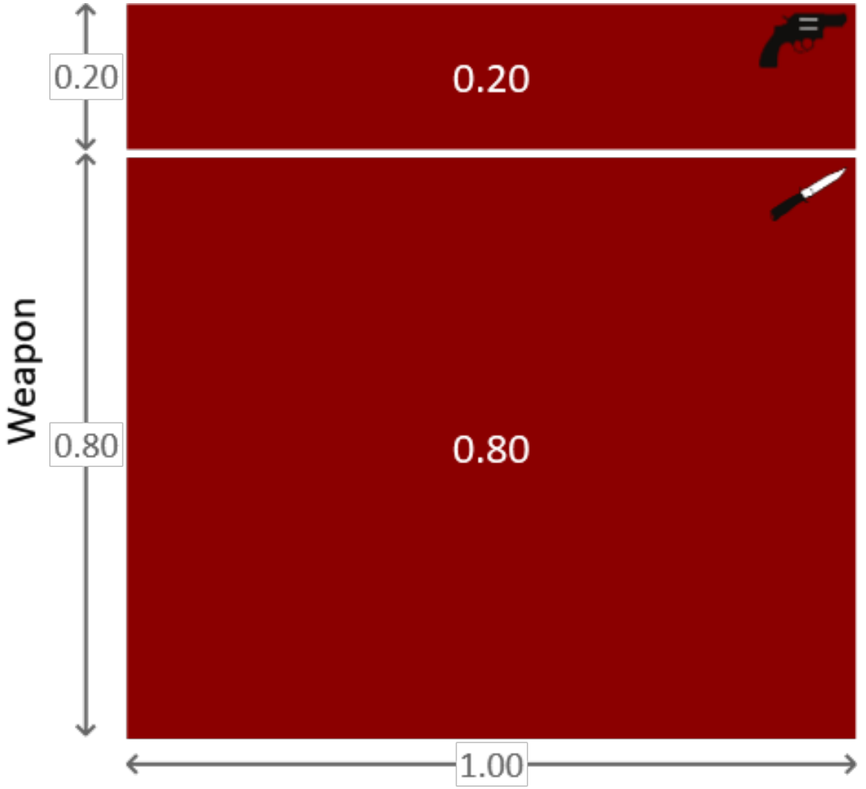
Evidence and conditional probability

We thus introduce a new random variable `weapon`, in addition to the existing random variable `murderer`.

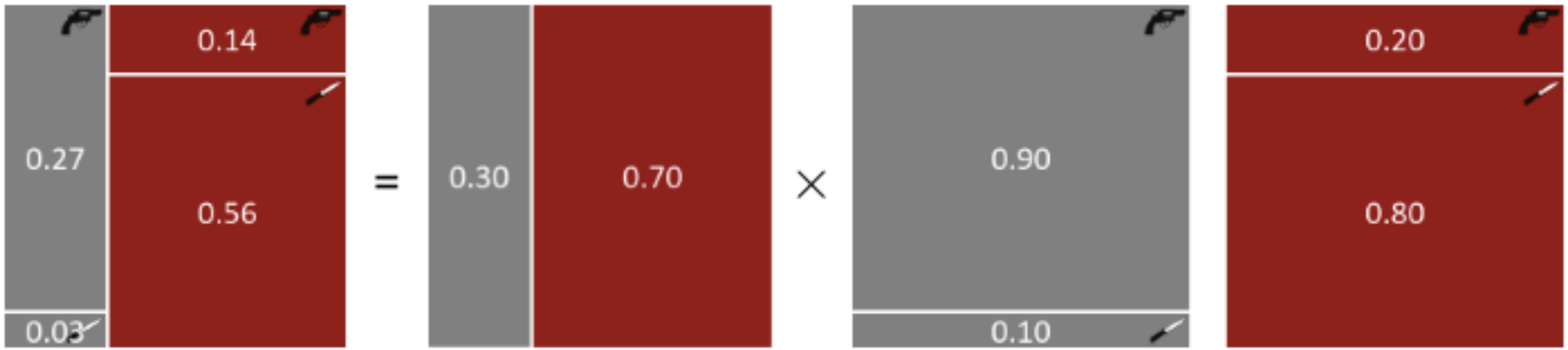
"Suppose Major Grey were the murderer. We might believe that the probability of his choosing a revolver rather than a dagger for the murder is, say, 90% on the basis that he is ex-military and would be familiar with the use of guns. But if instead Miss Auburn were the murderer, we might think the probability of her using a revolver would be much smaller, say 20%, on the basis that she is unlikely to be familiar with the operation of an old revolver and is therefore more likely to choose the dagger. This means that the probability distribution over the random variable weapon depends on whether the murderer is Major Grey or Miss Auburn. This is known as a conditional probability distribution because the probability values it gives vary depending on another random variable, in this case murderer."

$$p(weapon = revolver \mid murderer = grey) = 0.9$$
$$p(weapon = revolver \mid murderer = auburn) = 0.2$$





The joint Probability distribution



A probabilistic model is:

- A set of random variables,
- A joint probability distribution over these variables (i.e. a distribution that assigns a probability to every configuration of these variables such that the probabilities add up to 1 over all possible configurations).

Now we condition on some random variables and learn the values of others.

(paraphrased from Model Based Machine Learning)

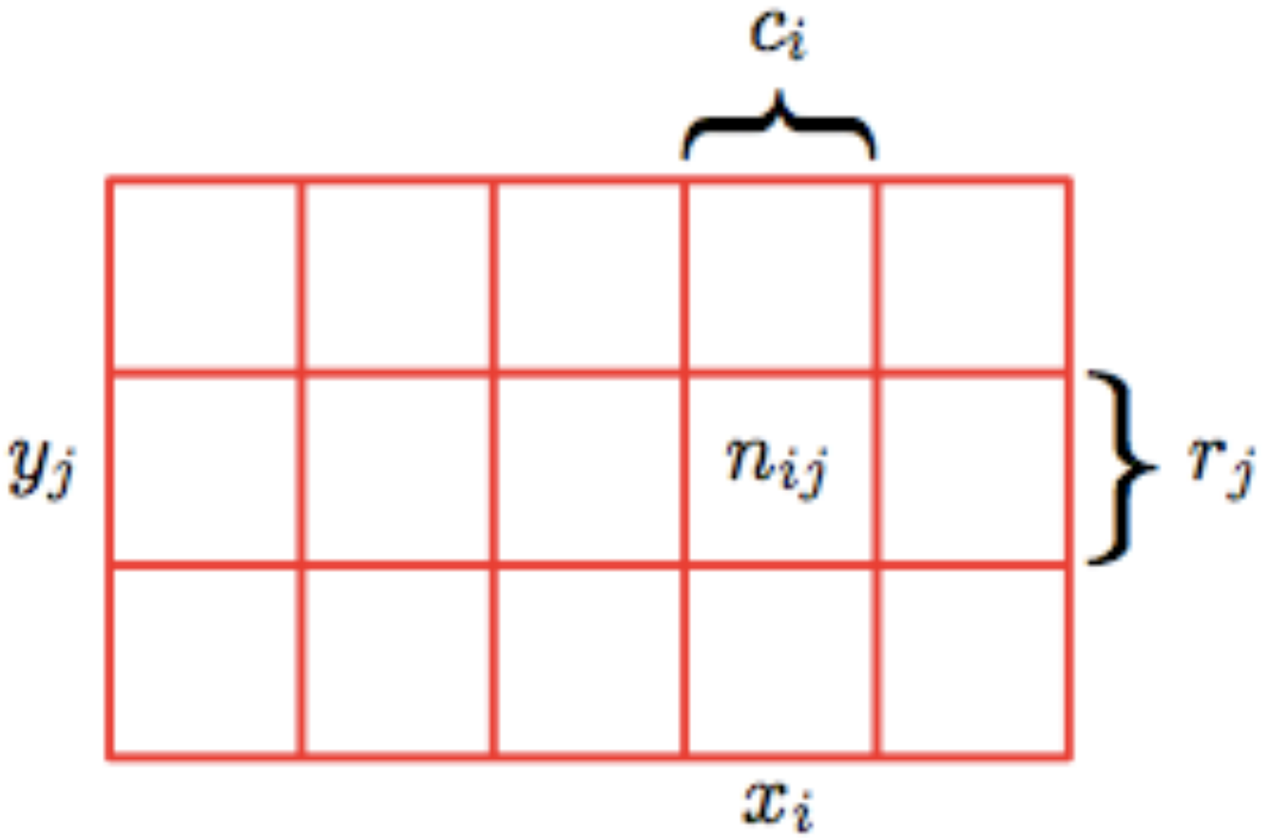
Rules

1. $P(A, B) = P(A \mid B)P(B)$

2. $P(A) = \sum_B P(A, B) = \sum_B P(A \mid B)P(B)$

$P(A)$ is called the **marginal** distribution of A, obtained by summing or marginalizing over B.

Marginals and Conditionals



$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$$

$$p(Y = y_j \mid X = x_i) \times p(X = x_i) = p(X = x_i, Y = y_j).$$

More generally for hidden variables z:

$$p(x) = \sum_z p(x, z) = \sum_z p(x|z)p(z)$$

Observation

Searching carefully around the library, Dr Bayes spots a bullet lodged in the book case. "Hmm, interesting", he says, "I think this could be an important clue".

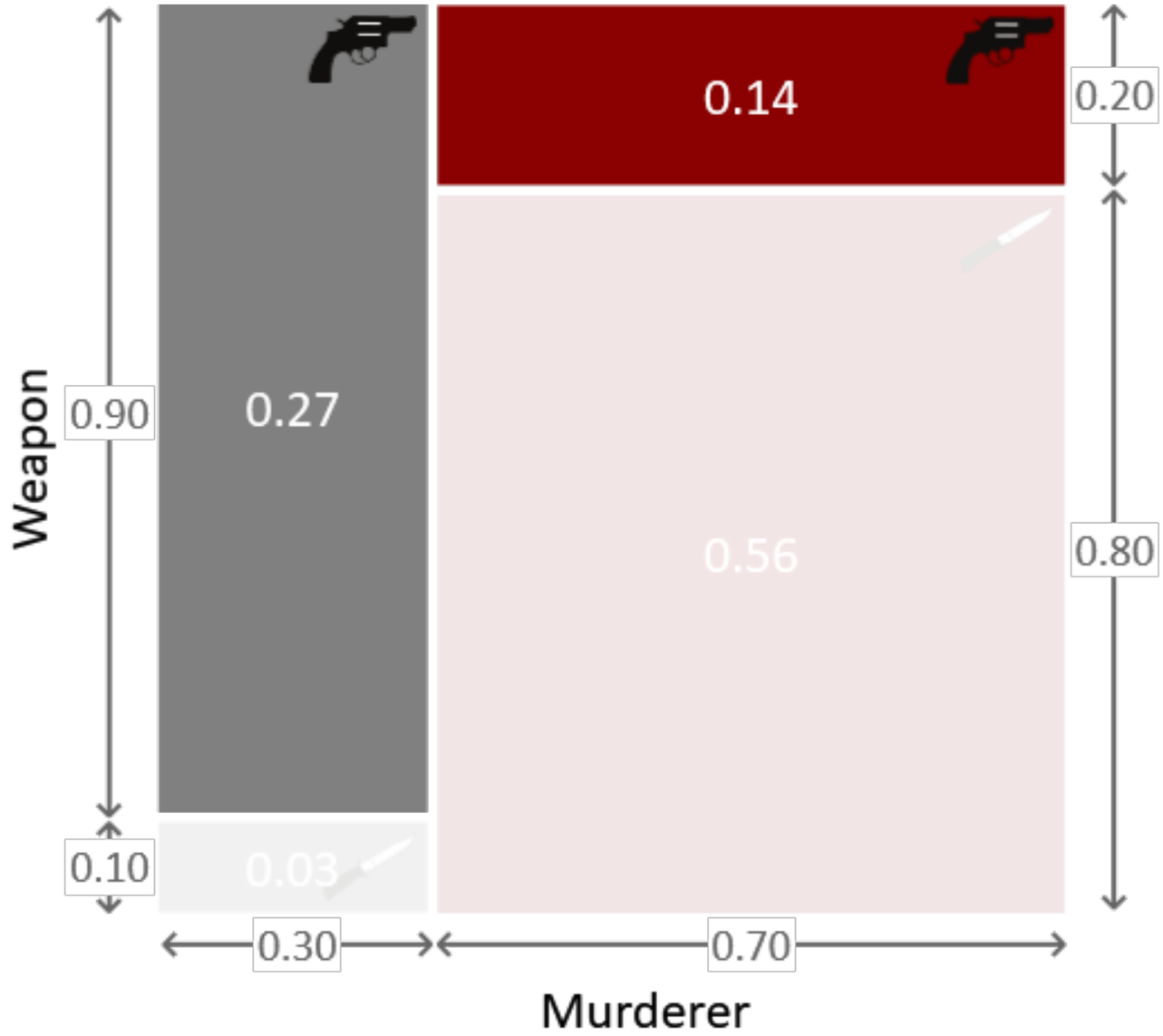
Inference

The process of computing revised probability distributions after we have observed the values of some the random variables, is called inference.

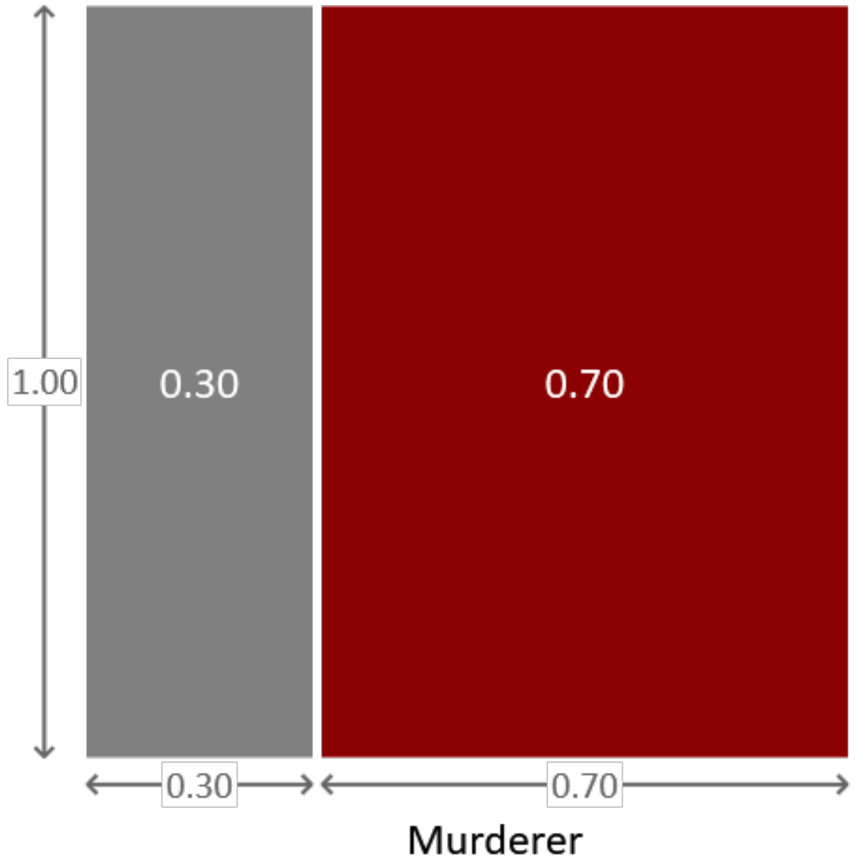
$$P(murderer = Grey | weapon = revolver) =$$

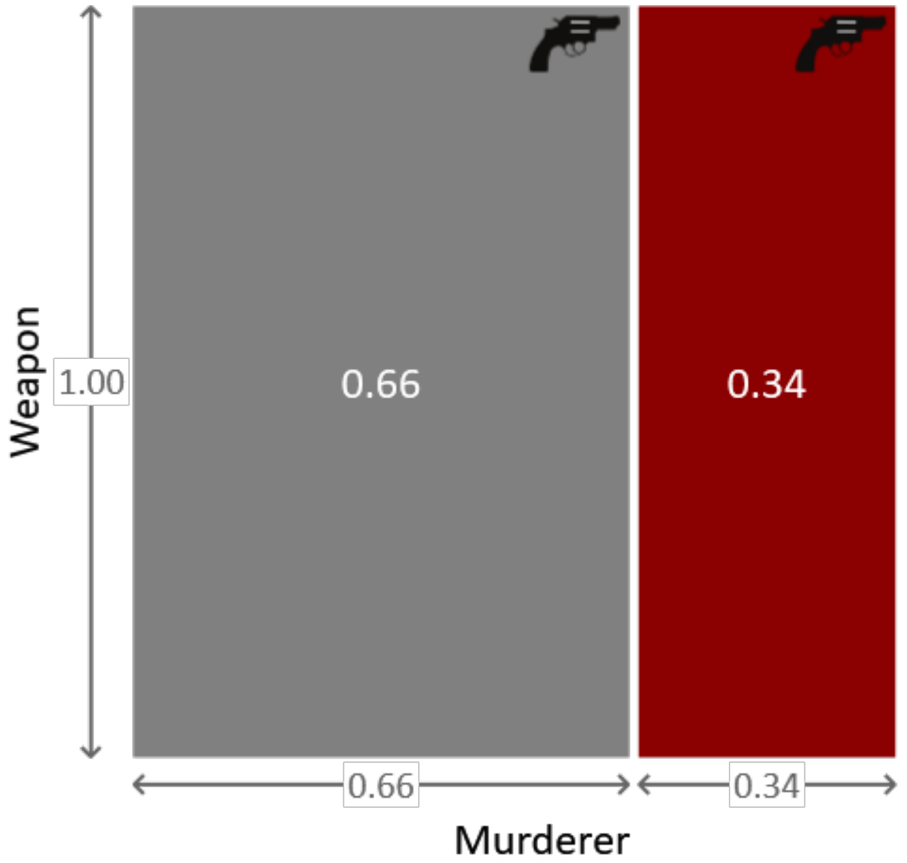
$$\frac{0.27}{0.27 + 0.14} \simeq 0.66$$

This **posterior probability** is higher than prior 0.3.



Inference: a pricipled way from prior to posterior





Bayes Theorem: Inference without computing the joint distribution

Why? The joint can be computationally hard. Sometimes there are two many "factors"

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)} = \frac{p(x \mid y)\, p(y)}{\sum_{y'} p(x, y')} = \frac{p(x \mid y)\, p(y)}{\sum_{y'} p(x \mid y')p(y')}$$

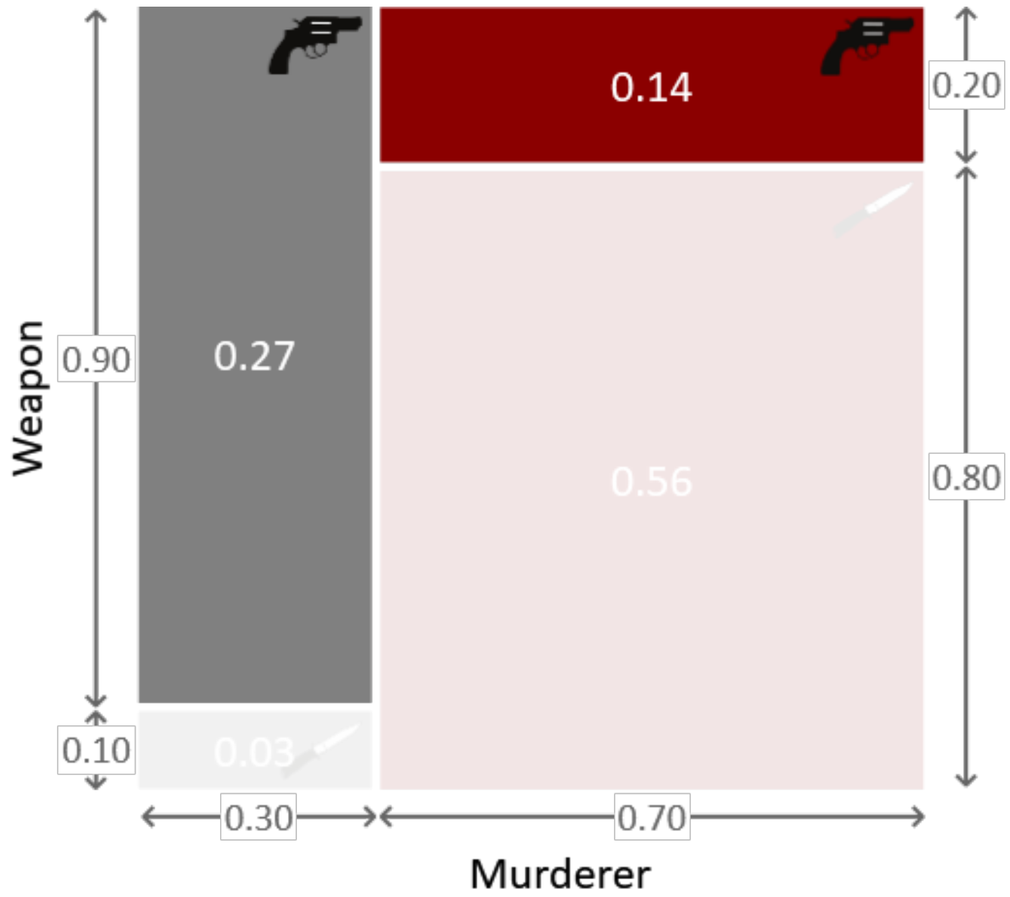
$$P(murderer|weapon) = \frac{P(weapon|murderer)P(murderer)}{P(weapon)}.$$

$$P(weapon) = \sum_{murderer} P(weapon|murderer)P(murderer)$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

The `evidence` is just a normalizer and can often be ignored.

The `likelihood function` is NOT a probability distribution over `weapon` (which is known!). It is a function of the random variable `murderer`.



Just ignore the fact that we are in a square!

Lets get precise

Cumulative distribution Function

The **cumulative distribution function**, or the **CDF**, is a function

$$F_X : \mathbb{R} \to [0, 1],$$

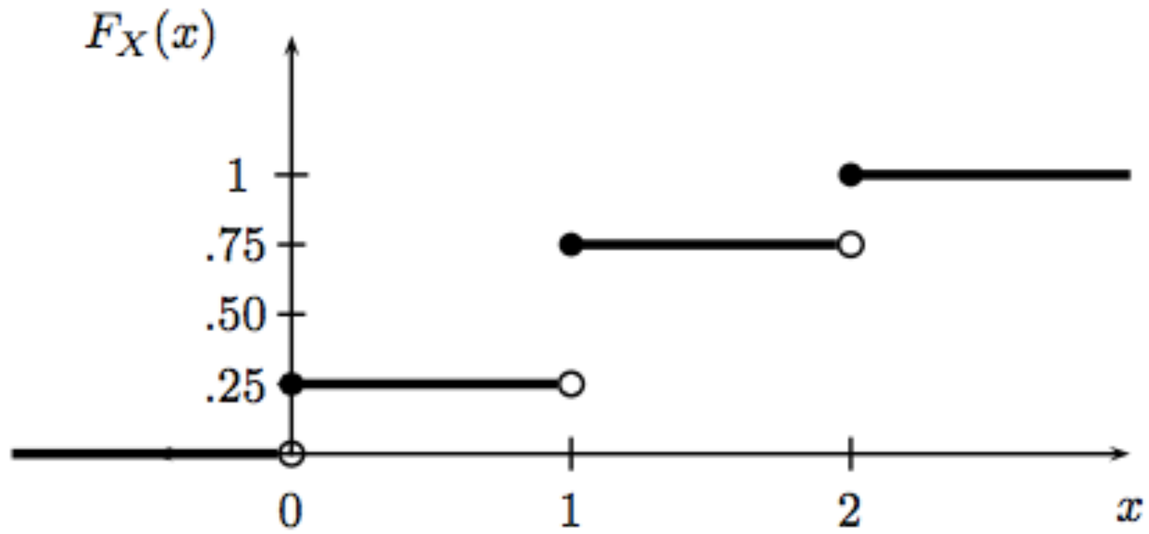
defined by

$$F_X(x) = p(X \leq x).$$

Sometimes also just called *distribution*.

Let X be the random variable representing the number of heads in two coin tosses. Then x = 0, 1 or 2.

CDF:



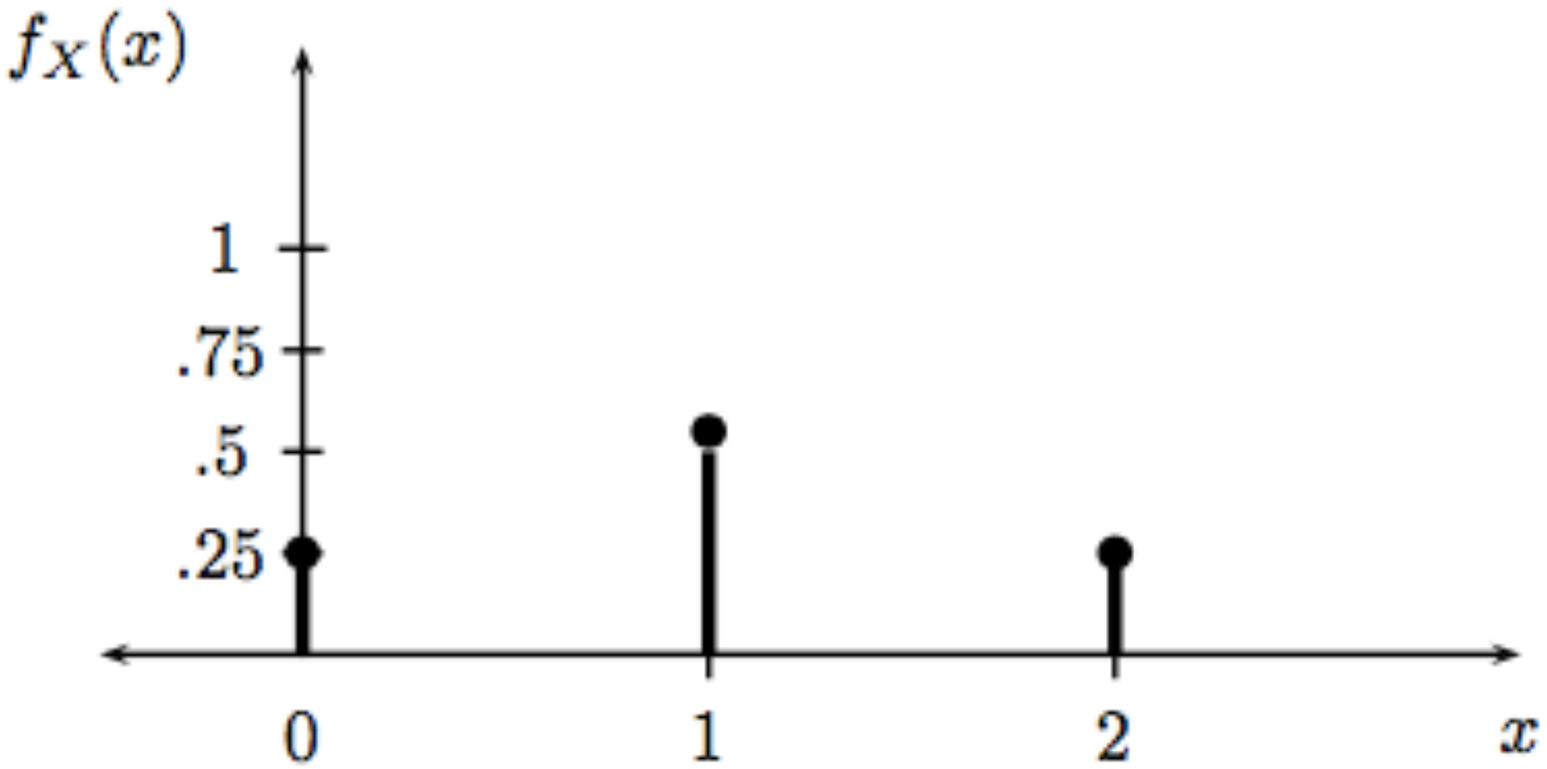
Probability Mass Function

X is called a **discrete random variable** if it takes countably many values $\{x_1, x_2, \ldots\}$.

We define the **probability function** or the **probability mass function** (**pmf**) for X by:

$$f_X(x) = p(X = x)$$

The pmf for the number of heads in two coin tosses:



Probability Density function (pdf)

A random variable is called a **continuous random variable** if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for every a ≤ b,

$$p(a < X < b) = \int_a^b f_X(x)dx$$

Note: $p(X = x) = 0$ for every x. Confusing!

CDF for continuous random variables

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = \frac{dF_X(x)}{dx}$ at all points x at which F_X is differentiable.

Continuous pdfs can be > 1. cdfs bounded in [0,1].

A continuous example: the Uniform(0,1) Distribution

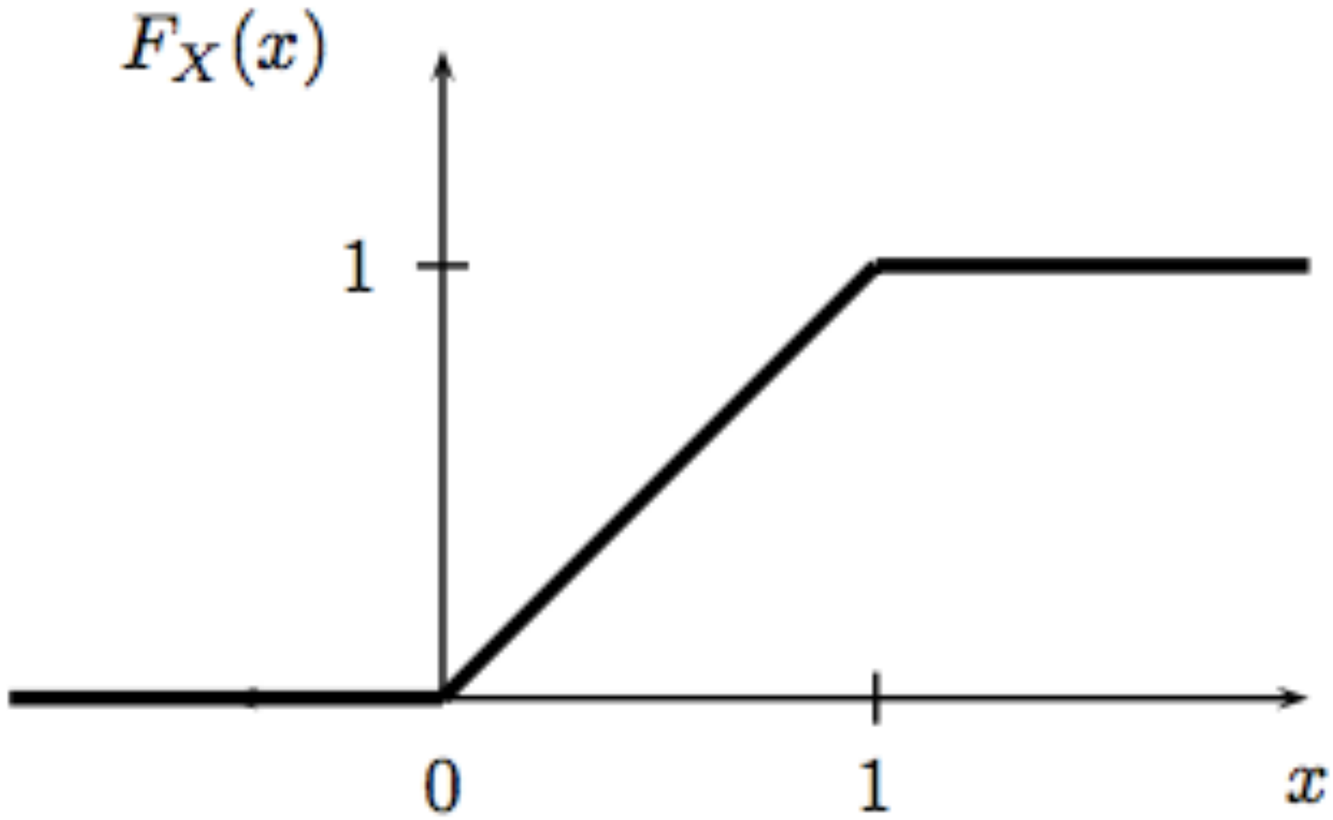
pdf:

$$f_X(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

cdf:

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1. \end{cases}$$

cdf:



Marginals

Marginal mass functions are defined in analog to probabilities:

$$f_X(x) = p(X = x) = \sum_y f(x, y); \;\; f_Y(y) = p(Y = y) = \sum_x f(x, y).$$

Marginal densities are defined using integrals:

$$f_X(x) = \int dy f(x, y); \;\; f_Y(y) = \int dx f(x, y).$$

Conditionals

Conditional mass function is a conditional probability:

$$f_{X|Y}(x \mid y) = p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{f_{XY}(x, y)}{f_Y(y)}$$

The same formula holds for densities with some additional requirements $f_Y(y) > 0$ and interpretation:

$$p(X \in A \mid Y = y) = \int_{x \in A} f_{X|Y}(x, y) dx.$$

Bernoulli pmf:

$$f(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1. \end{cases}$$

for p in the range 0 to 1.

$$f(x) = p^x (1-p)^{1-x}$$

for x in the set {0,1}.

What is the cdf?

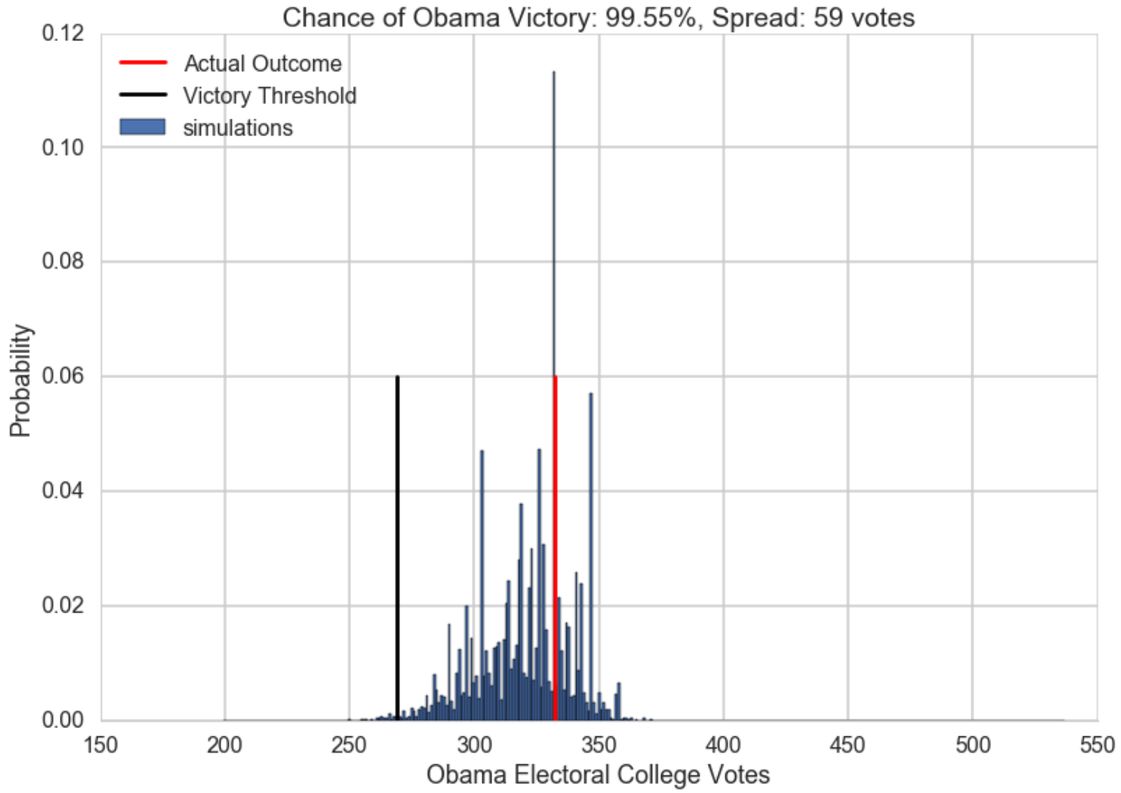
```
from scipy.stats import bernoulli
#bernoulli random variable
brv=bernoulli(p=0.3)
print(brv.rvs(size=20))
```

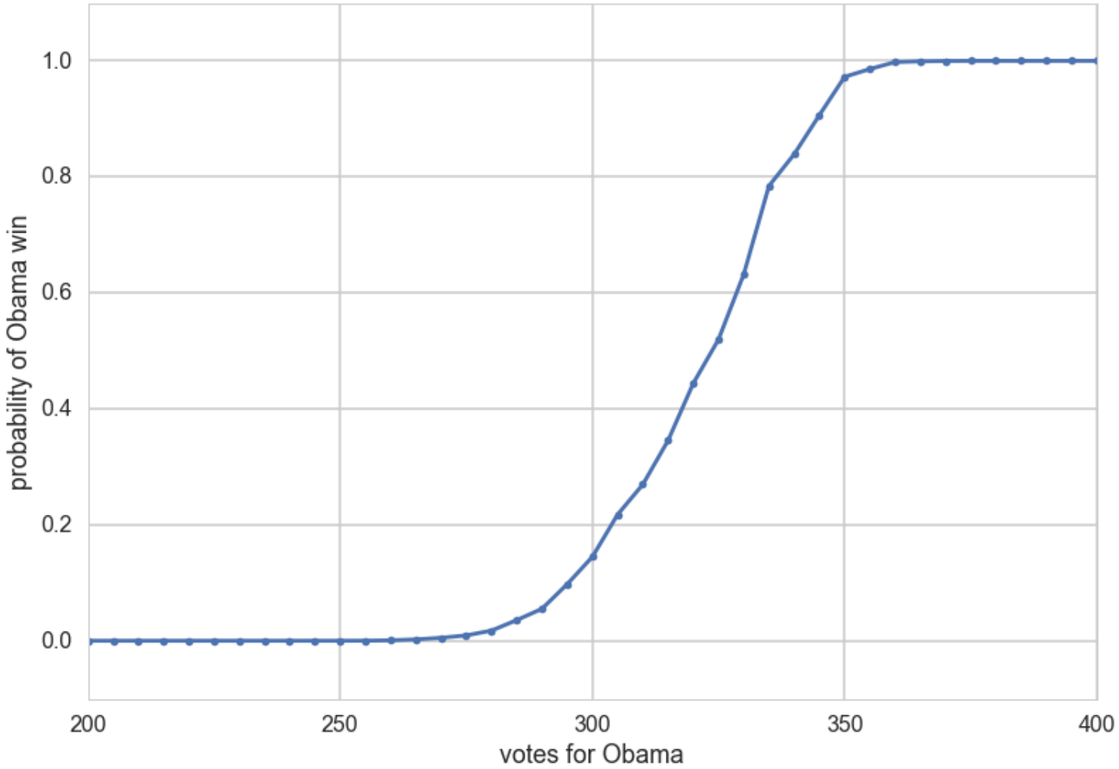
```
[1 0 0 0 1 0 0 1 1 0 0 0 0 0 1 1 0 0 1 0]
```

Election forecasting

- Each state has a Bernoulli coin.
- p for each state can come from prediction markets, models, polls
- Many simulations for each state. In each simulation:
  - $rv = Uniform(0, 1)$ If. $rv < p$ say Obama wins
  - or $rv = Bernoulli(p)$. 1=Obama.

Empirical pmf and cdf





The big Ideas

- a data story
- inference using the data story

Data story

- a story of how the data came to be.
- may be a causal story, or a descriptive one (correlational, associative).
- **The story must be sufficient to specify an algorithm to simulate new data**.
- a **formal probability model**.

tossing a globe in the air experiment

- toss and catch it. When you catch it, see whats under index finger
- mark `W` for water, `L` for land.
- figure how much of the earth is covered in water
- thus the "data" is the fraction of `W` tosses

Probabilistic Model

1. The true proportion of water is p.
2. Bernoulli probability for each globe toss, where p is thus the probability that you get a `W`. This assumption is one of being **Identically Distributed**.
3. Each globe toss is **Independent** of the other.

Assumptions 2 and 3 taken together are called **IID**, or **Independent and Identially Distributed** Data.

Next time

- Expectation values
- Law of large numbers
- How it enables empirical distributions
- And Monte Carlo
- Central Limit theorem for sampling and error on expectations