Lecture 2

Distributions and Frequentist Statistics

So far:

- Probability and Bayes Theorem

Today:

- Distributions
- Frequentist Statistics
- Maximum Likelihood Estimation
- Sampling Distribution

A Random Variable is a mapping

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

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

Example: the number of heads in a sequence of N coin tosses.

Cumulative distribution Function

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

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

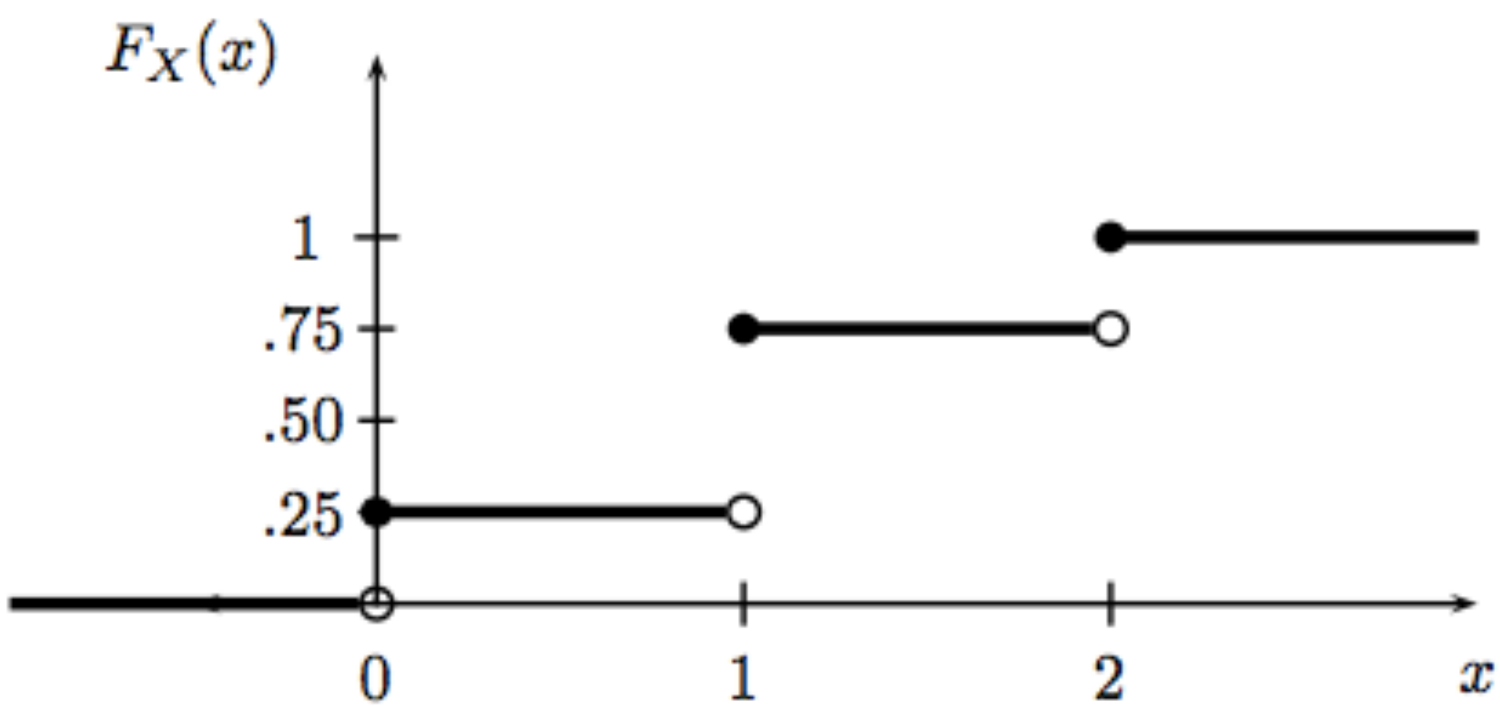
defined by

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

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:



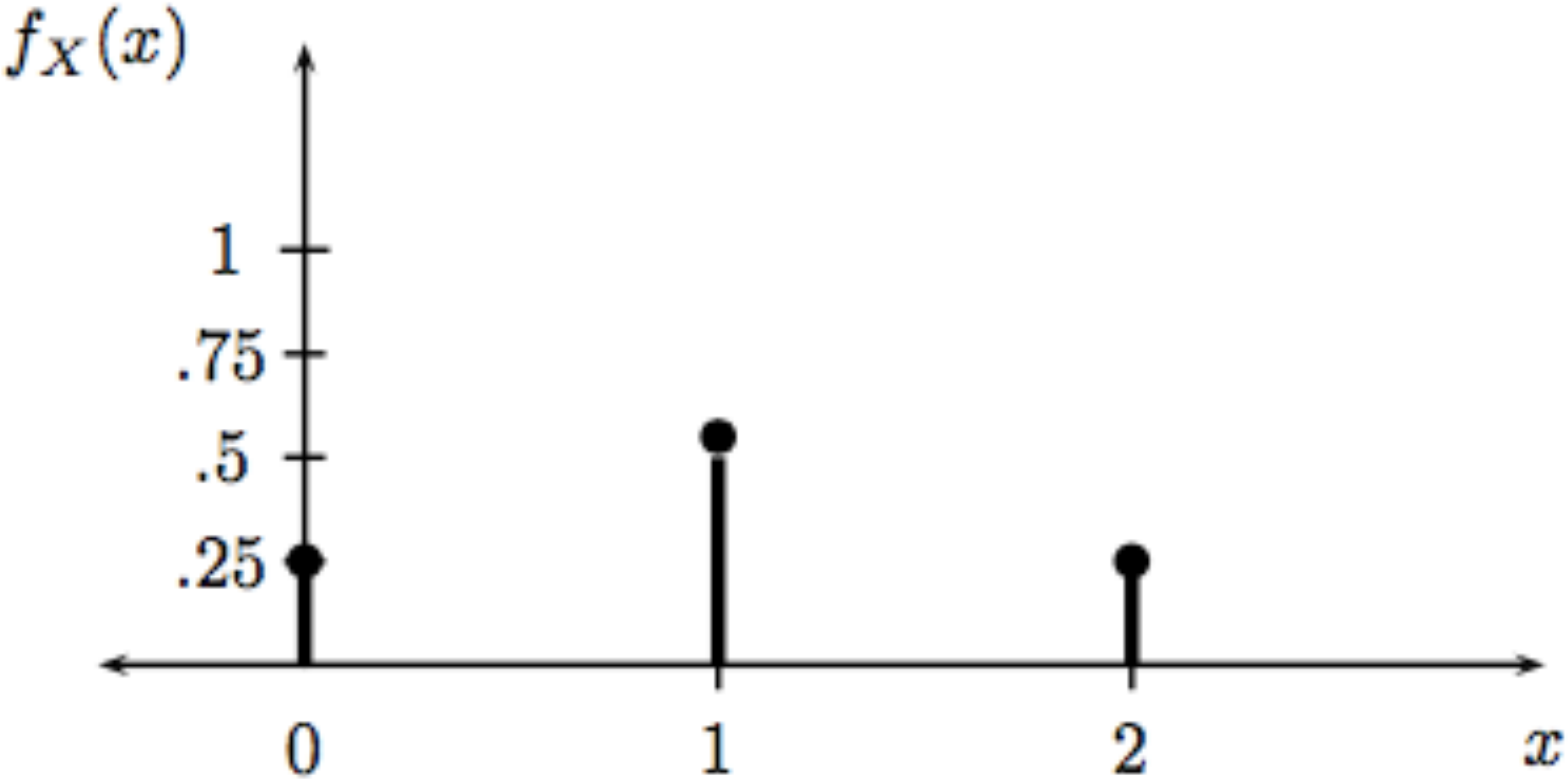
Probability Mass Function

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

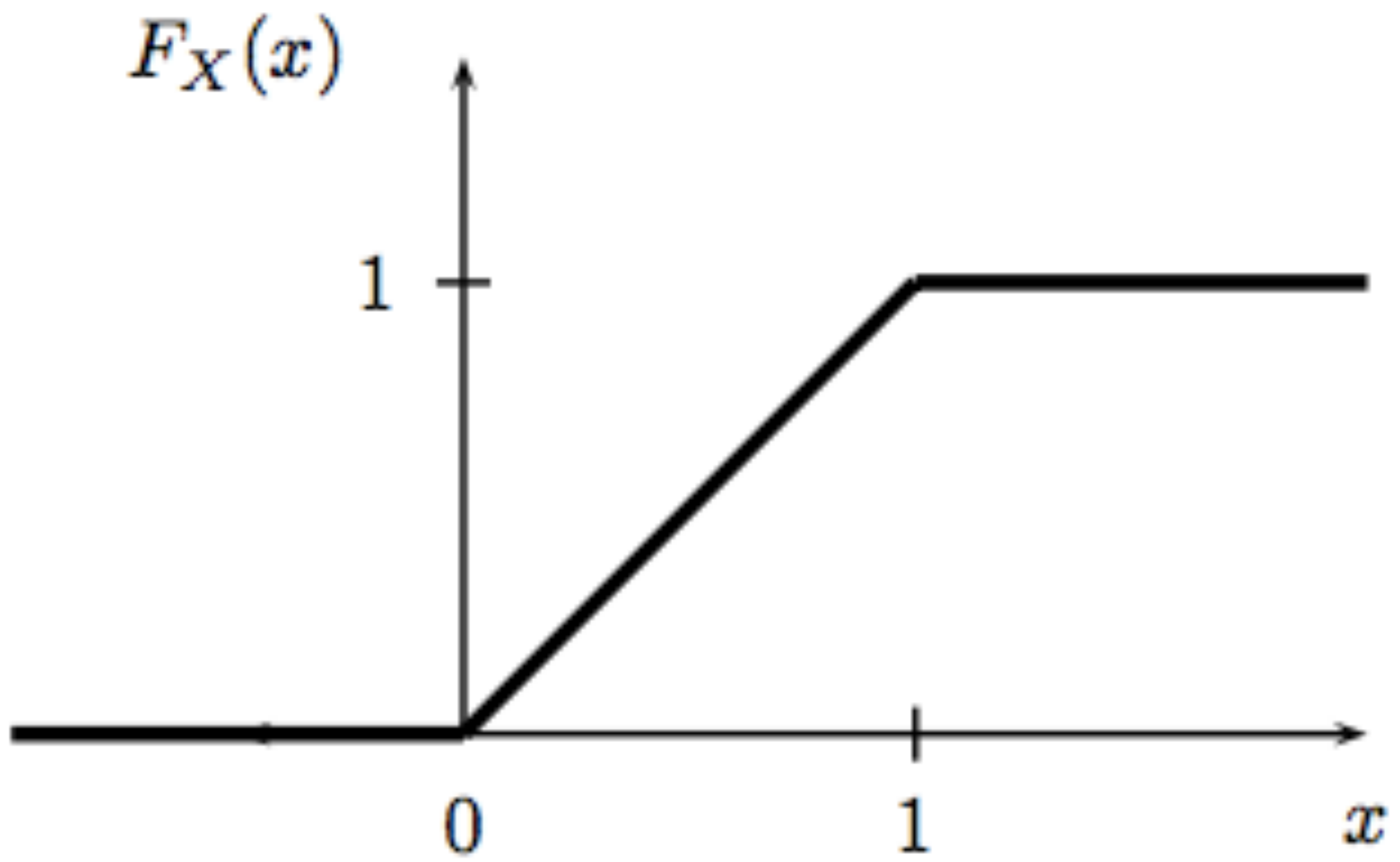
pdf:

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

cdf:

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

cdf:



Bernoulli Distribution

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

$$X \sim Bernoulli(p)$$

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

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?

```python
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]
```

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:

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

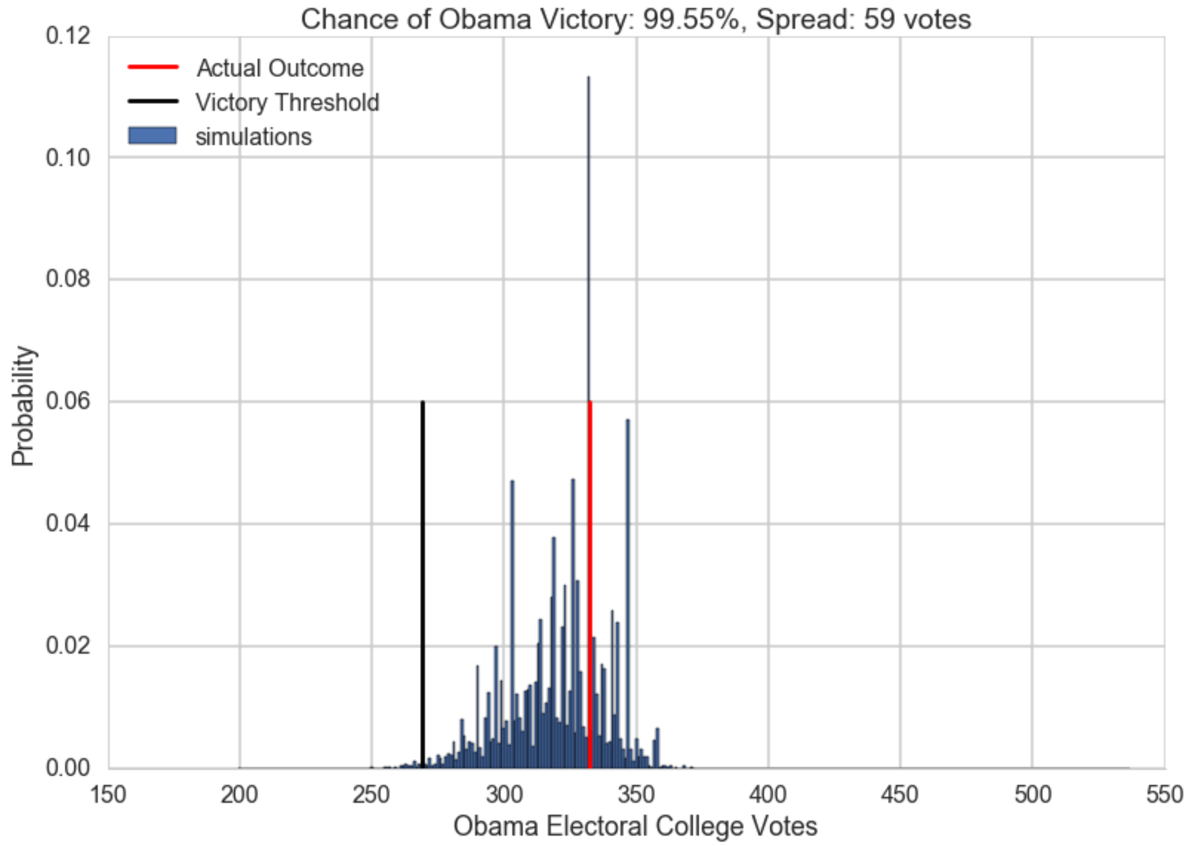
The same formula holds for densities with some additional requirements $f_Y(y) > 0$ and interpretation:

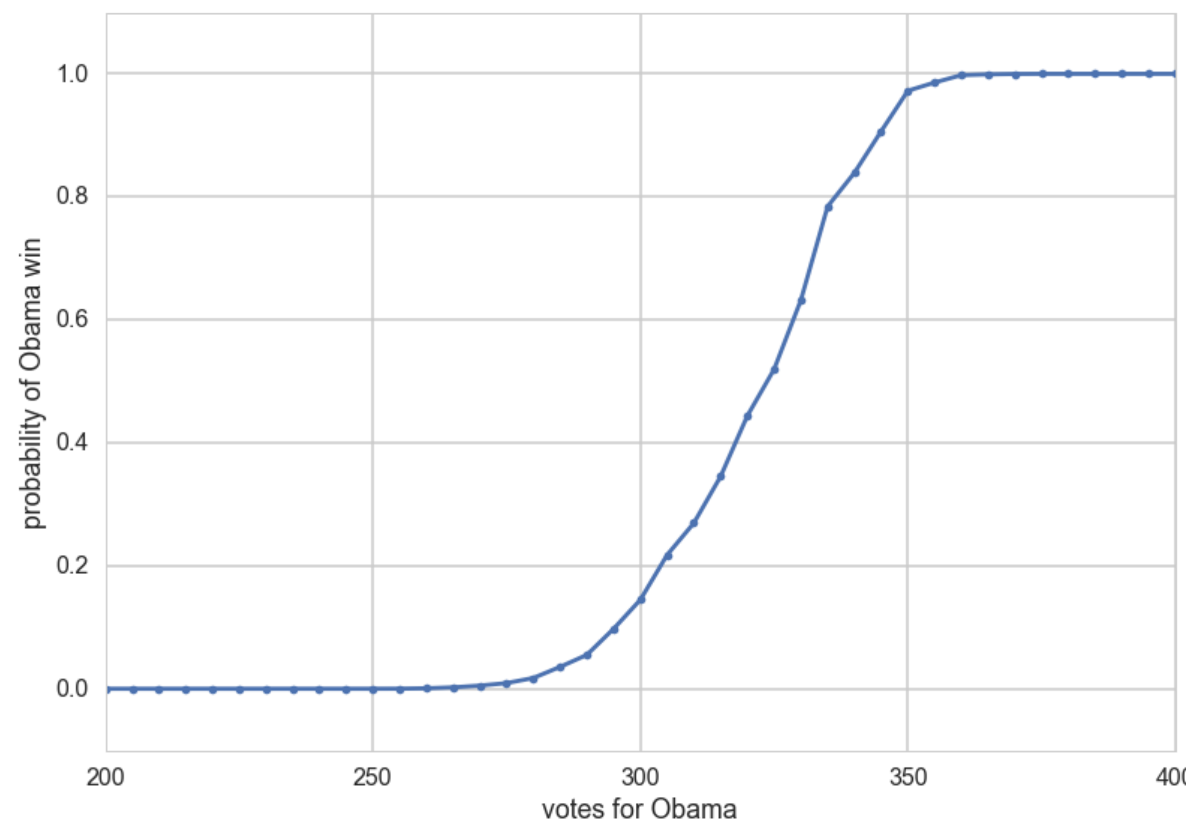
$$p(X \in A \mid Y = y) = \int_{x \in A} f_{X|Y}(x, y) dx.$$

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





Frequentist Statistics

Answers the question: **What is Data?** with

"data is a **sample** from an existing **population**"

- data is stochastic, variable
- model the sample. The model may have parameters
- find parameters for our sample. The parameters are considered fixed.

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**.

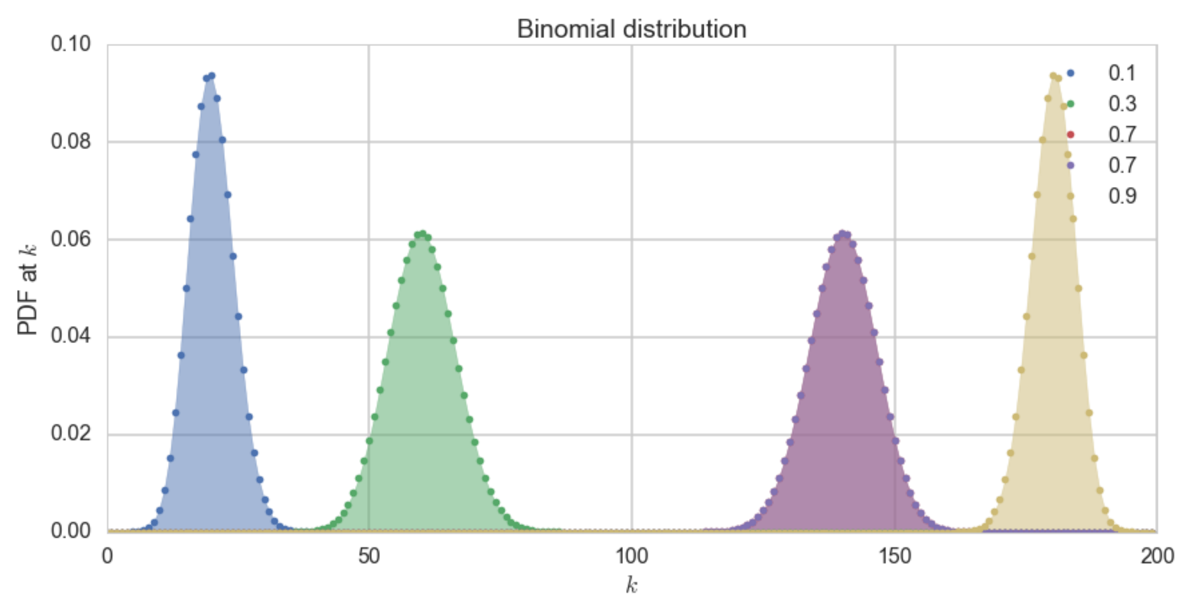
3. Each globe toss is **Independent** of the other.

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

Likelihood

How likely it is to observe k `W` given the parameter p?

$$P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$



Likelihood

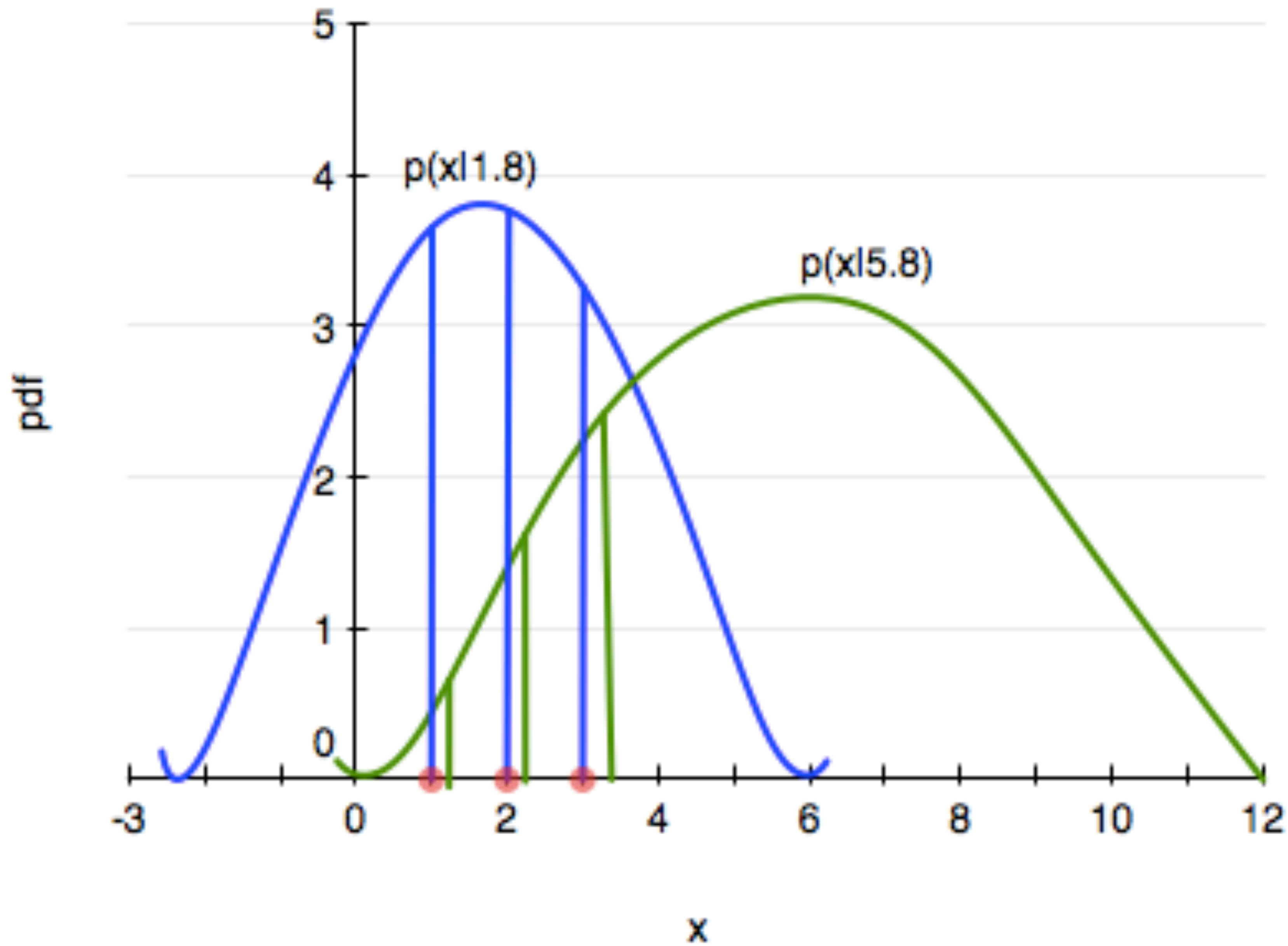
How likely it is to observe values $x_1, \ldots, x_n$ given the parameters λ?

$$L(\lambda) = \prod_{i=1}^{n} P(x_i | \lambda)$$

How likely are the observations if the model is true?

Or, how likely is it to observe k out of n `W`

Maximum Likelihood estimation



Example Exponential Distribution Model

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Describes the time between events in a homogeneous Poisson process (events occur at a constant average rate). Eg time between buses arriving.

log-likelihood

Maximize the likelihood, or more often (easier and more numerically stable), the log-likelihood

$$\ell(\lambda) = \sum_{i=1}^{n} ln(P(x_i \mid \lambda))$$

In the case of the exponential distribution we have:

$$\ell(lambda) = \sum_{i=1}^{n} ln(\lambda e^{-\lambda x_i}) = \sum_{i=1}^{n} (ln(\lambda) - \lambda x_i) \,.$$

Maximizing this:

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

and thus:

$$\hat{\lambda_{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is the sample mean of our sample.

Globe Toss Model

$$P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\ell = log(\binom{n}{k}) + klog(p) + (n-k)log(1-p)$$

$$\frac{d\ell}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0$$

thus $p_{MLE} = \frac{k}{n}$

Point Estimates

If we want to calculate some quantity of the population, like say the mean, we estimate it on the sample by applying an estimator F to the sample data D, so $\hat{\mu} = F(D)$.

Remember, **The parameter is viewed as fixed and the data as random, which is the exact opposite of the Bayesian approach which you will learn later in this class.**

True vs estimated

If your model describes the true generating process for the data, then there is some true μ^*.

We dont know this. The best we can do is to estimate $\hat{\mu}$.

Now, imagine that God gives you some M data sets **drawn** from the population, and you can now find μ on each such dataset.

So, we'd have M estimates.

Sampling distribution

As we let $M \to \infty$, the distribution induced on $\hat{\mu}$ is the empirical **sampling distribution of the estimator**.

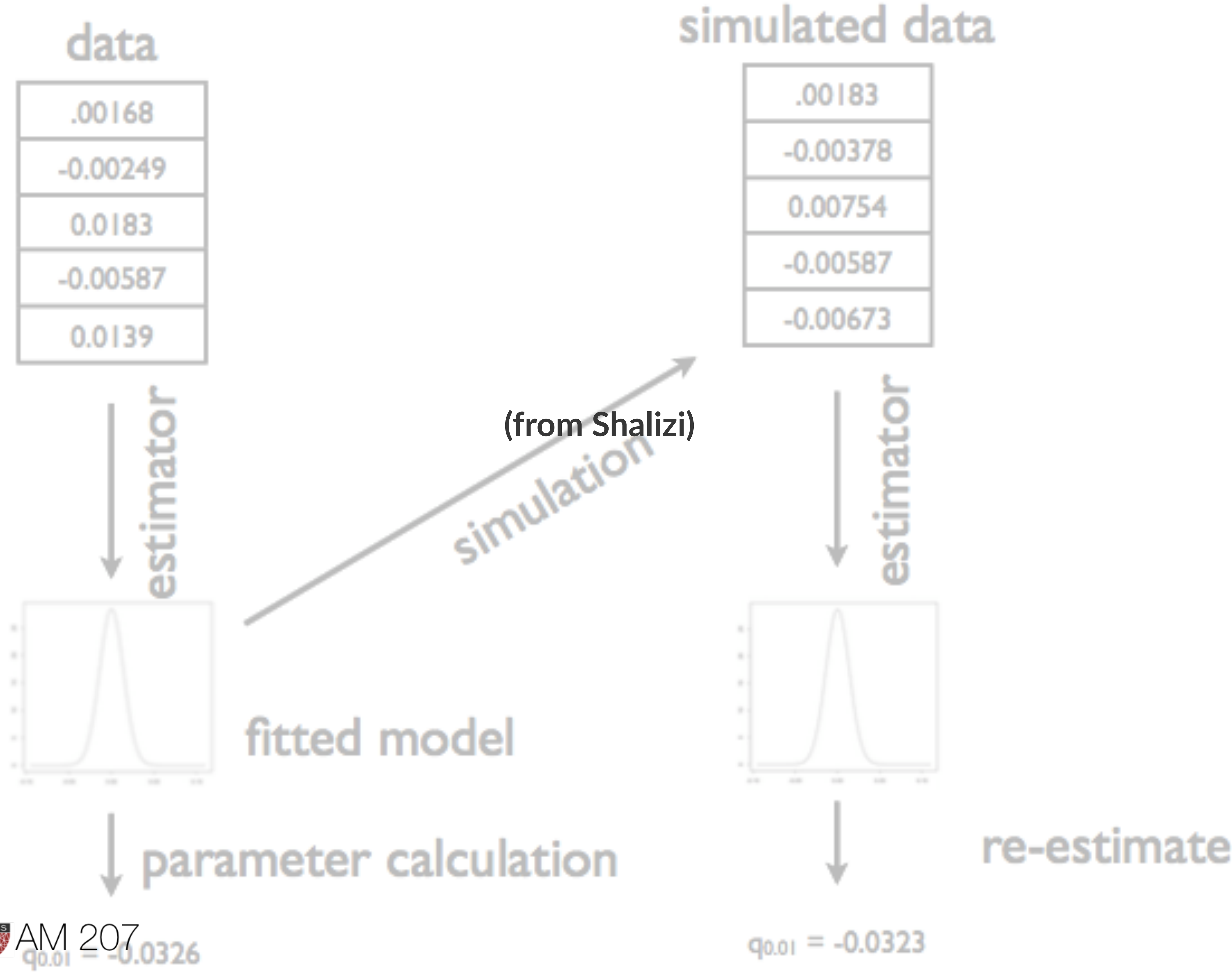
μ could be λ, our parameter, or a mean, a variance, etc

We could use the sampling distribution to get confidence intervals on λ.

But we dont have M samples. What to do?

Bootstrap

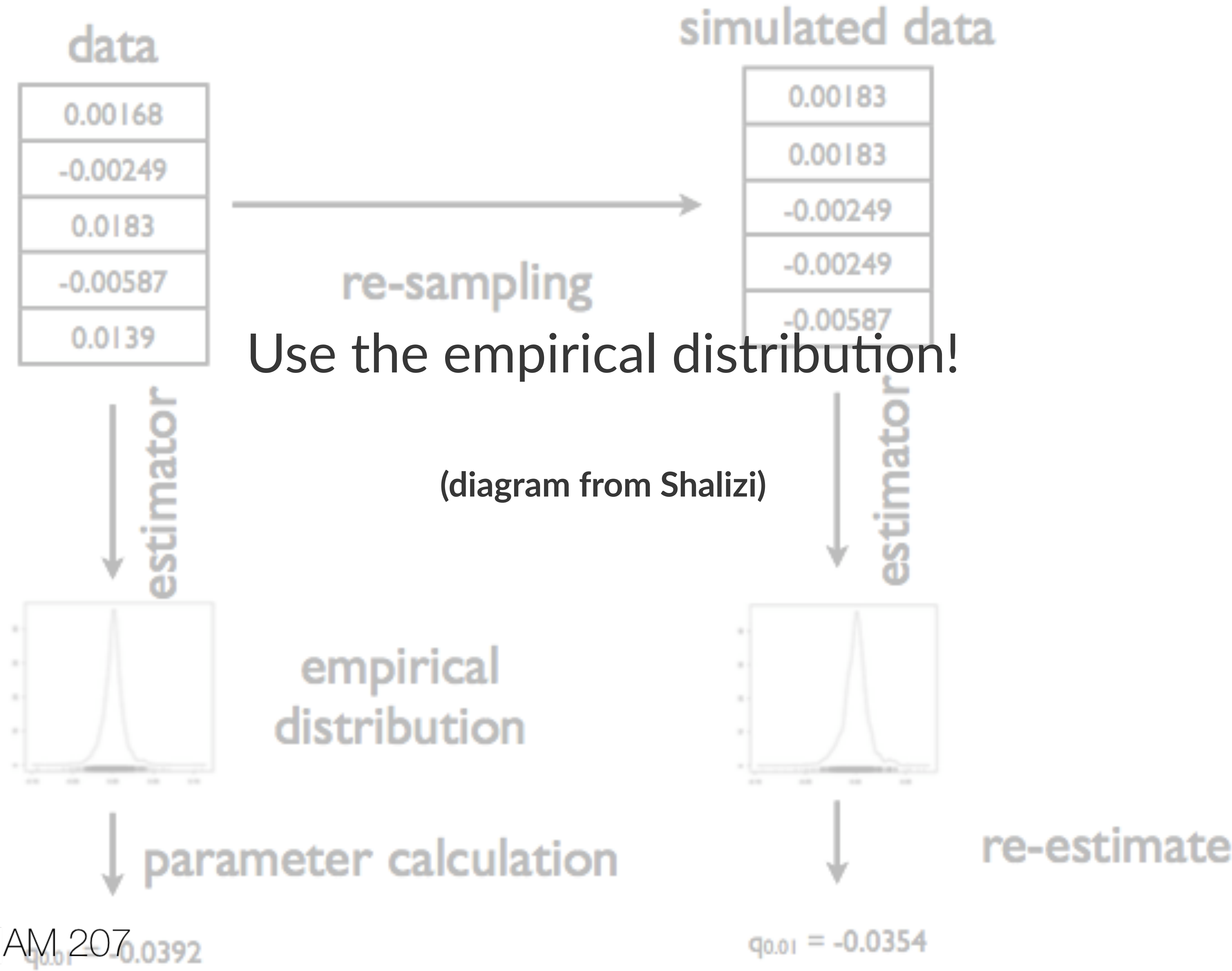
- If we knew the true parameters of the population, we could generate M fake datasets.
- we dont, so we use our estimate $lam\hat{b}da$ to generate the datasets
- this is called the Parametric Bootstrap
- usually best for statistics that are variations around truth



(from Shalizi)

Problems

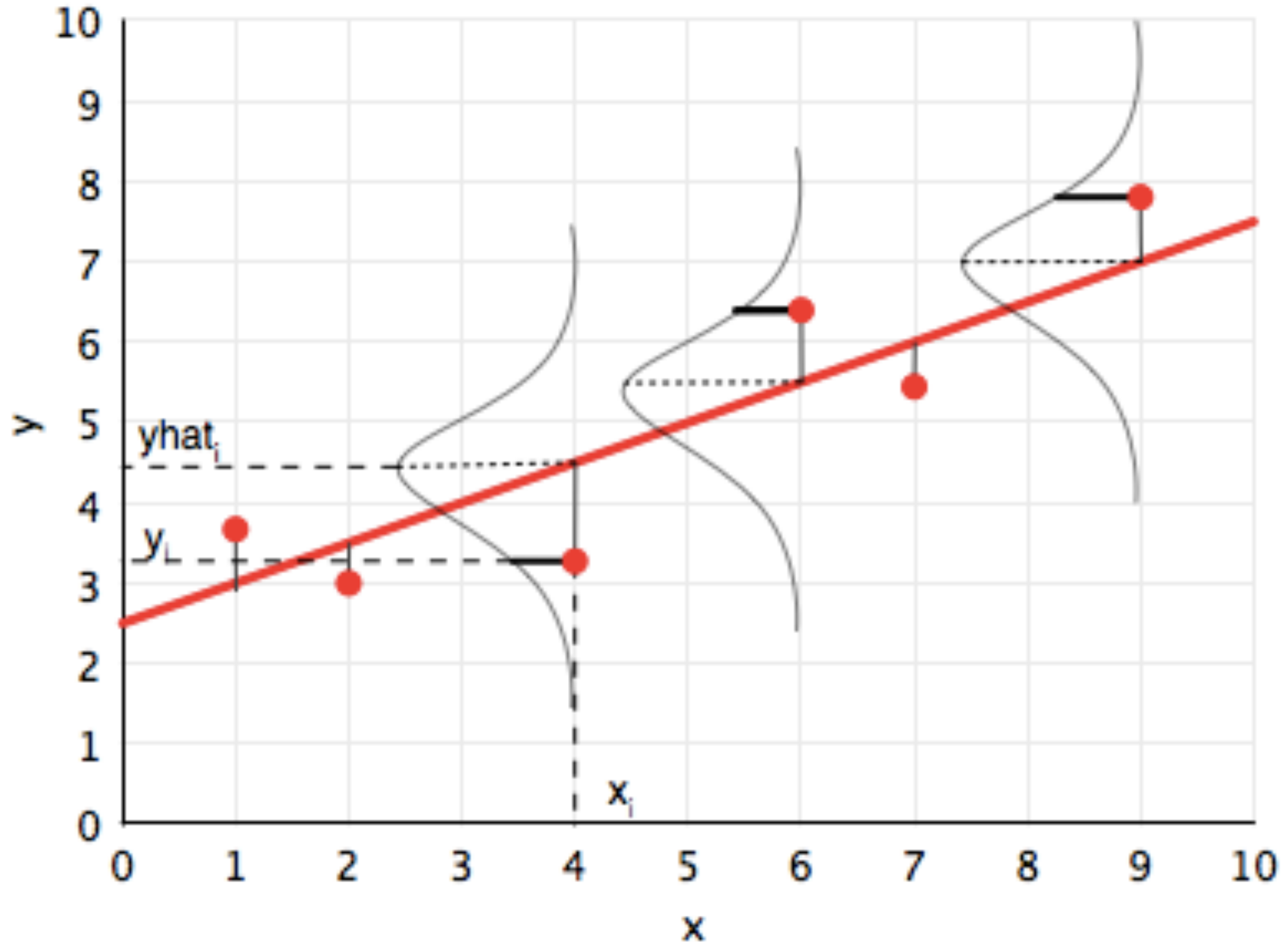
- simulation error: the number of samples M is finite. *Go large M*.
- statistical error: resampling from an estimated parameter is not the "true" data generating process. *Subtraction helps*.
- specification error: the model isnt quite good. *Use the parametric bootstrap*: sample with replacement the X from our original sample D, generating many fake datasets.



Use the empirical distribution!

(diagram from Shalizi)

Linear Regression MLE



Gaussian Distribution assumption

Each y_i is gaussian distributed with mean $\mathbf{w} \cdot \mathbf{x}_i$ (the y predicted by the regression line) and variance σ^2:

$$y_i \sim N(\mathbf{w} \cdot \mathbf{x}_i, \sigma^2).$$

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2},$$

We can then write the likelihood:

$$\mathcal{L} = p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_i p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}, \sigma)$$

$$\mathcal{L} = (2\pi\sigma^2)^{(-n/2)} e^{\frac{-1}{2\sigma^2} \sum_i (y_i - \mathbf{w}\cdot\mathbf{x}_i)^2}.$$

The log likelihood ℓ then is given by:

$$\ell = \frac{-n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2.$$

Maximizing gives:

$$\mathbf{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

where we stack rows to get:

$$\mathbf{X} = stack(\{\mathbf{x}_i\})$$

$$\sigma^2_{MLE} = \frac{1}{n}\sum_i (y_i - \mathbf{w}\cdot\mathbf{x}_i)^2.$$

Next time

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