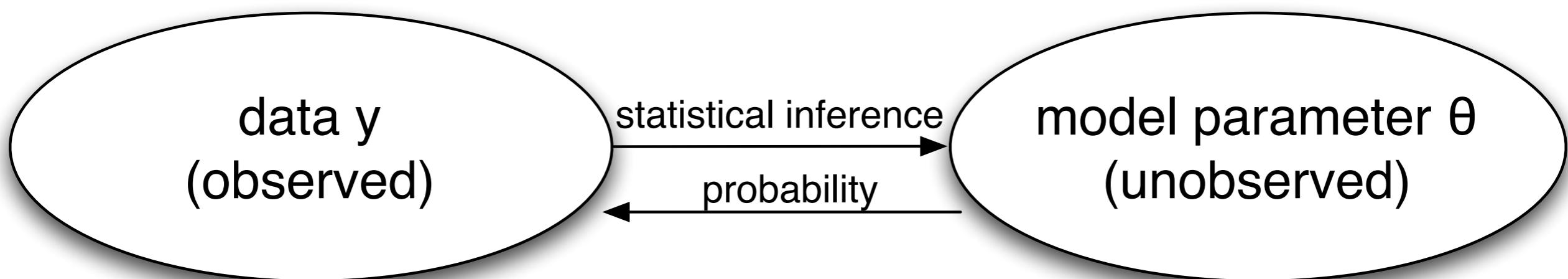


CS I 09/Stat I 2I/AC209/E- I 09

# Data Science

## Statistical Models

Hanspeter Pfister, Joe Blitzstein, and Verena Kaynig



# This Week

- HWI due next Thursday - start last week!
- Section assignments coming soon
- Please avoid posting duplicate questions on Piazza (always search first), and avoid posting code from your homework solutions (see Andrew's post <https://piazza.com/class/icf0cypdc3243c?cid=310> for more)

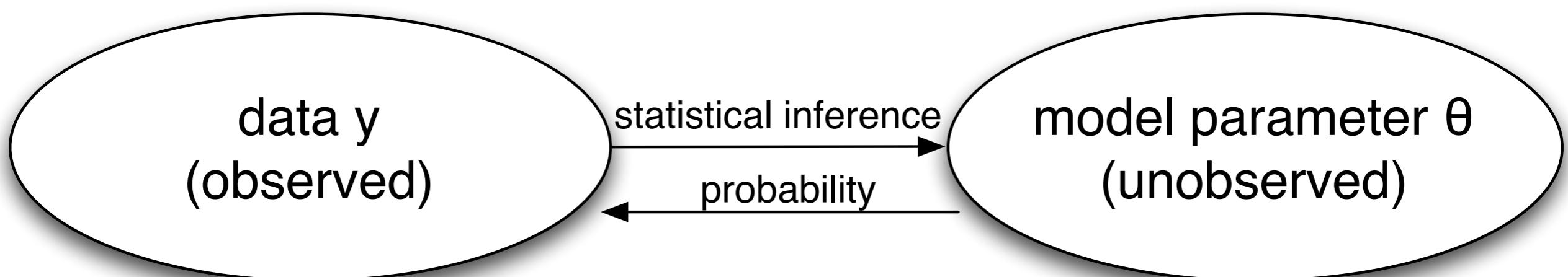
# Drowning in data, but starved for information



source: <http://extensionengine.com/drowning-in-data-the-biggest-hurdle-for-mooc-proliferation/>

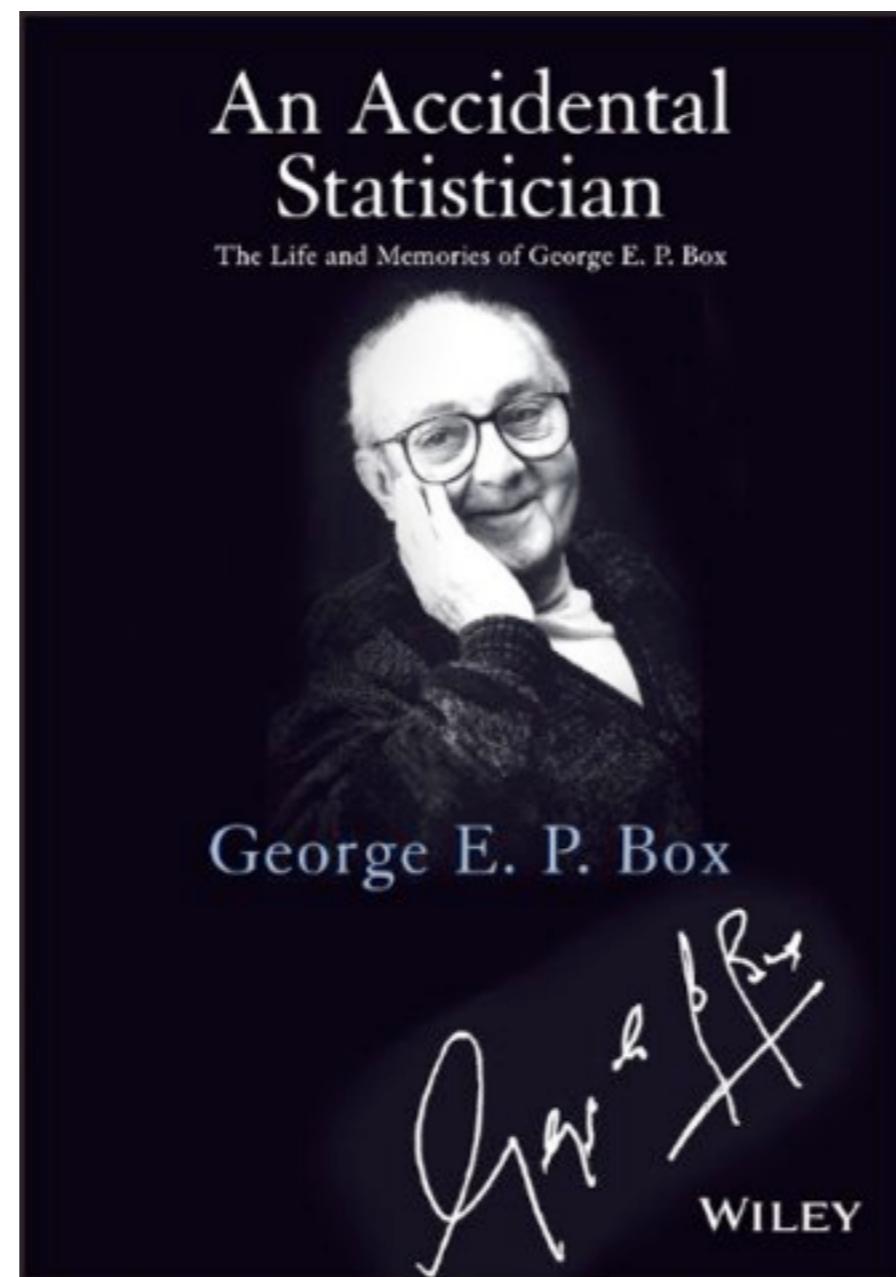
# What is a statistical model?

- a *family* of distributions, indexed by *parameters*
- sharpens distinction between *data* and *parameters*, and between *estimators* and *estimands*
- parametric (e.g., based on Normal, Binomial) vs. nonparametric (e.g., methods like bootstrap, KDE)



# What good is a statistical model?

“All models are wrong, but some models are useful.”  
– George Box (1919-2013)



# Jorge Luis Borges, “On Exactitude in Science”

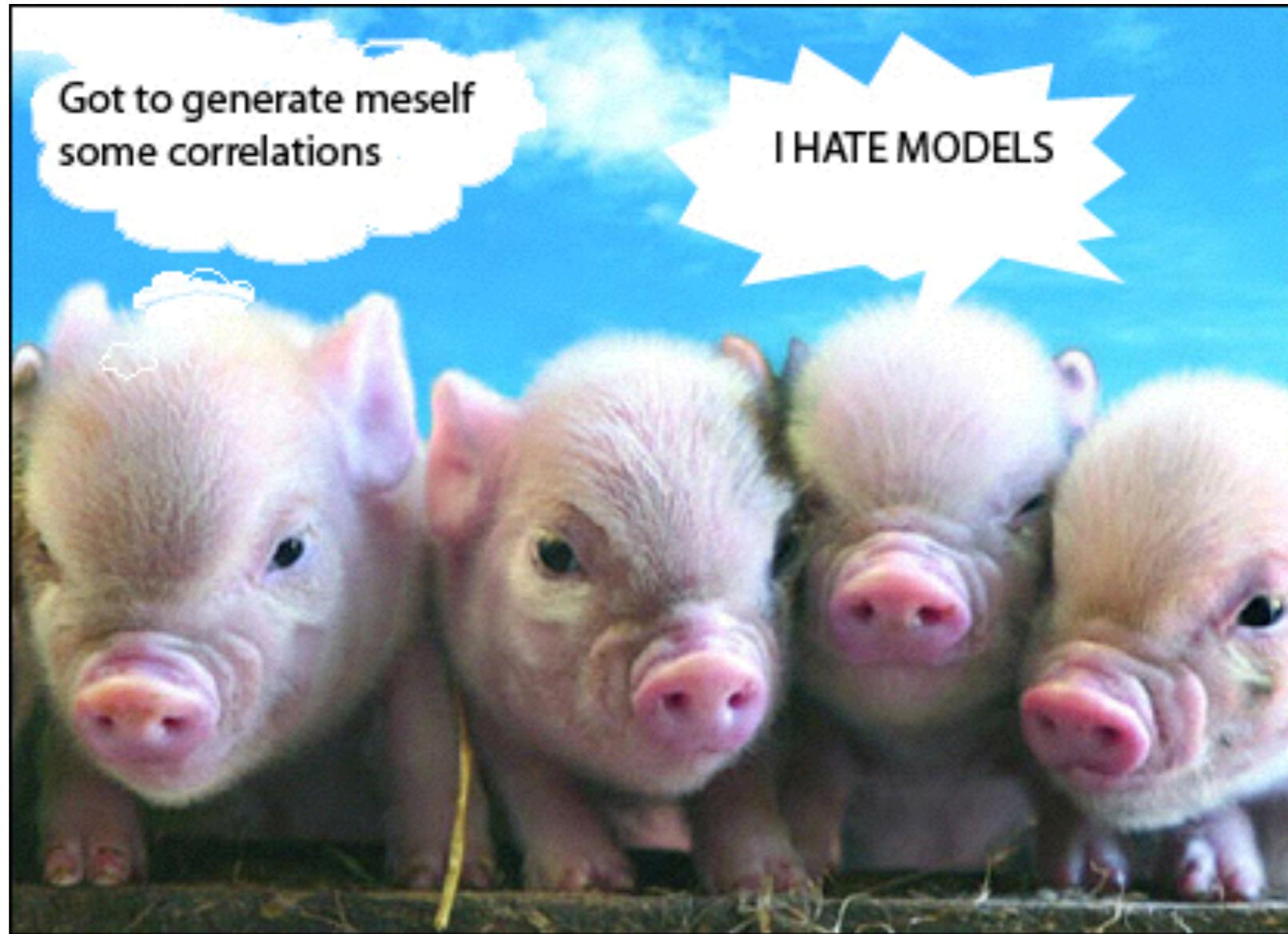
In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guild struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it.



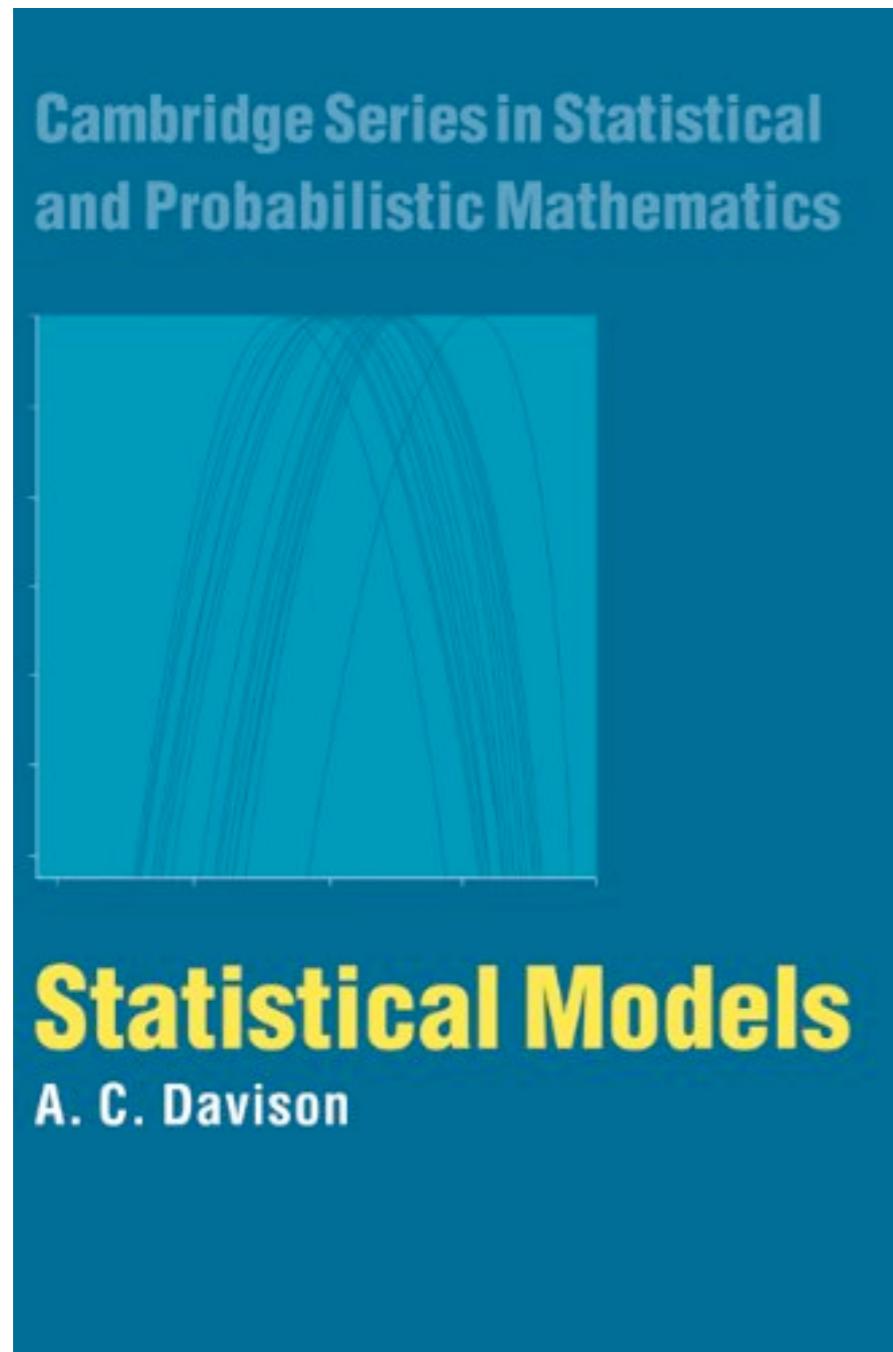
Borges Google Doodle

# “Big Data vs. Pig Data”:

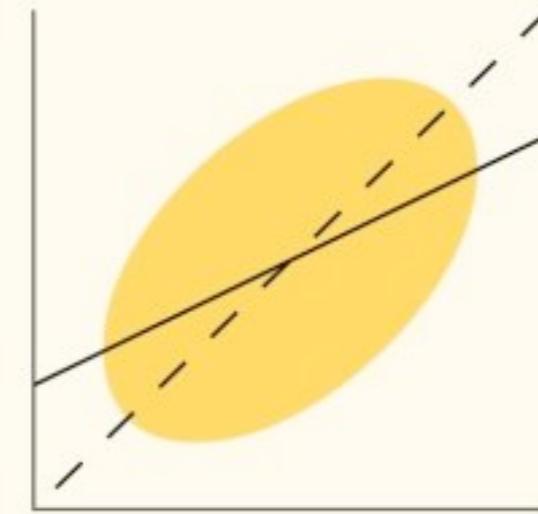
<https://scensci.wordpress.com/2012/12/14/big-data-or-pig-data/>



# Statistical Models: Two Books



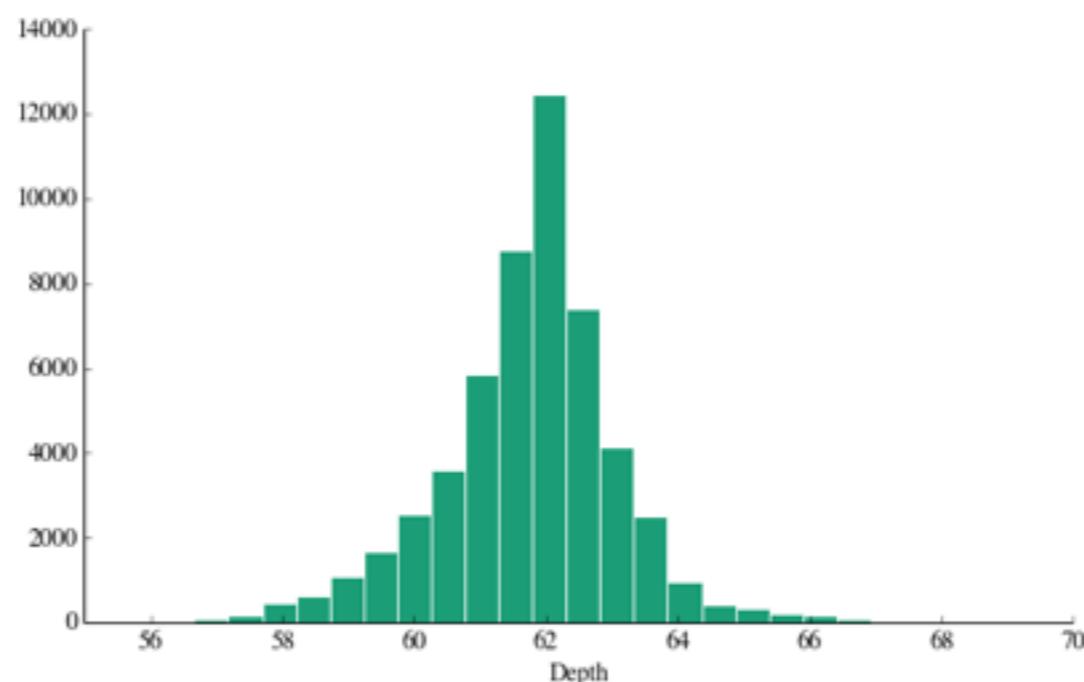
**Statistical Models**  
Theory and Practice  
REVISED EDITION



David A. Freedman

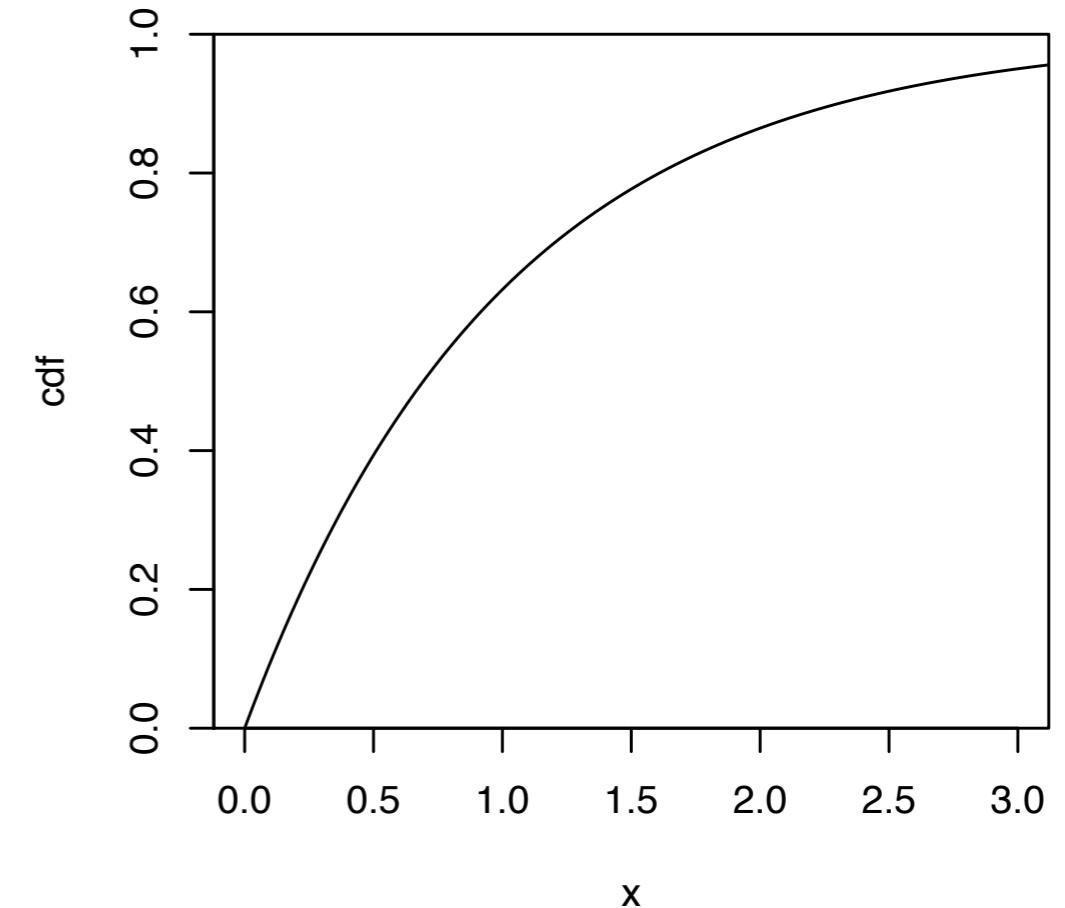
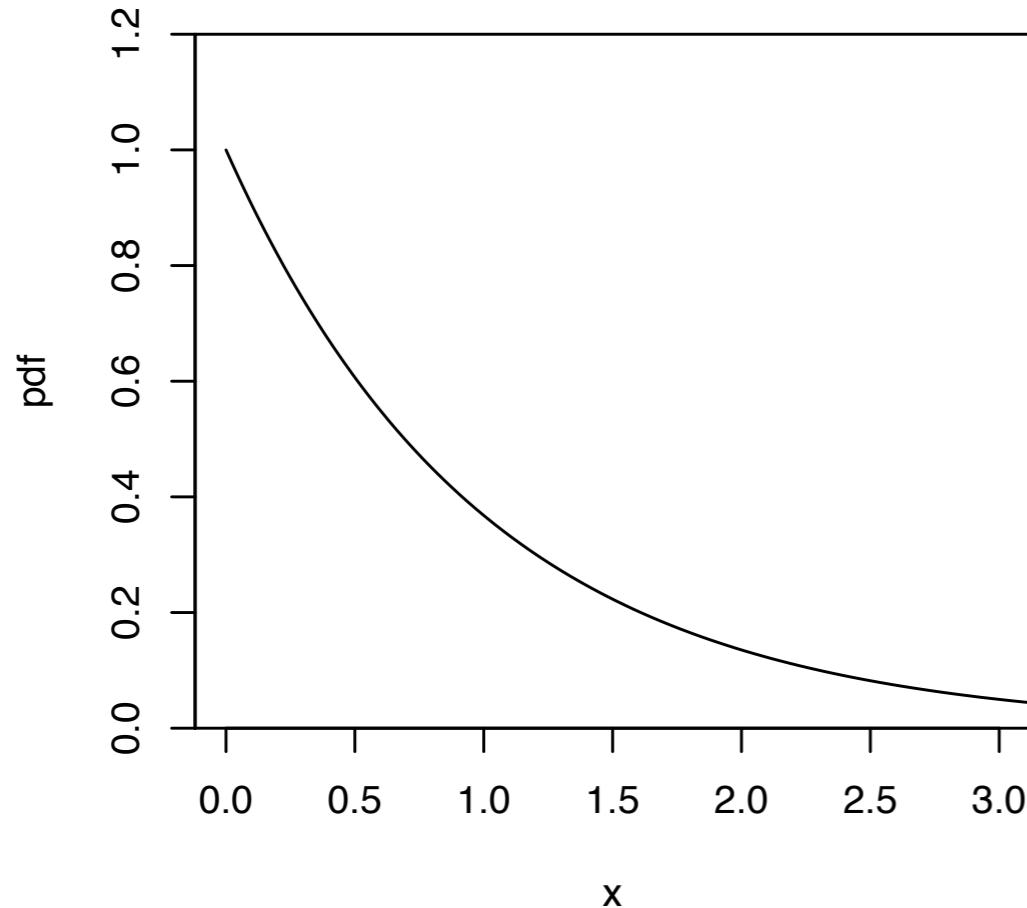
# Parametric vs. Nonparametric

- parametric: finite-dimensional parameter space (e.g., mean and variance for a Normal)
- nonparametric: infinite-dimensional parameter space
- is there anything in between?
- nonparametric is very general, but no free lunch!
- remember to plot and explore the data!



# Parametric Model Example: Exponential Distribution

$$f(x) = \lambda e^{-\lambda x}, x > 0$$



Remember the memoryless property!

# Exponential Distribution

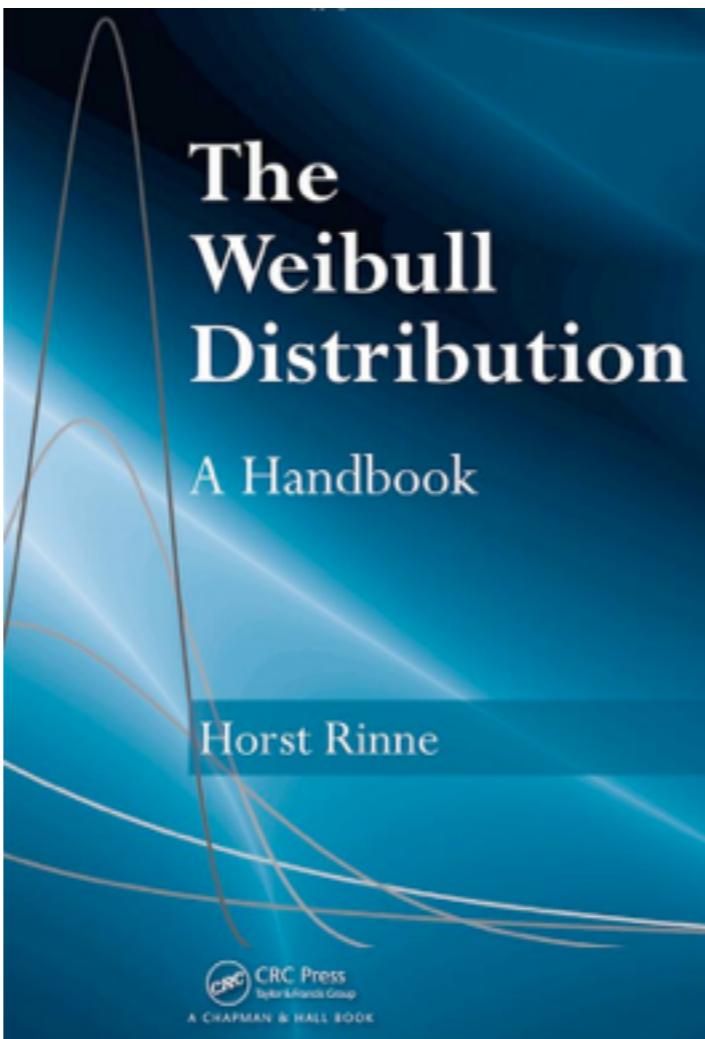
$$f(x) = \lambda e^{-\lambda x}, x > 0$$

- Exponential is *characterized* by memoryless property
- all models are wrong, but some are useful...
- iterate between exploring, the data model-building, model-fitting, and model-checking
- key building block for more realistic models

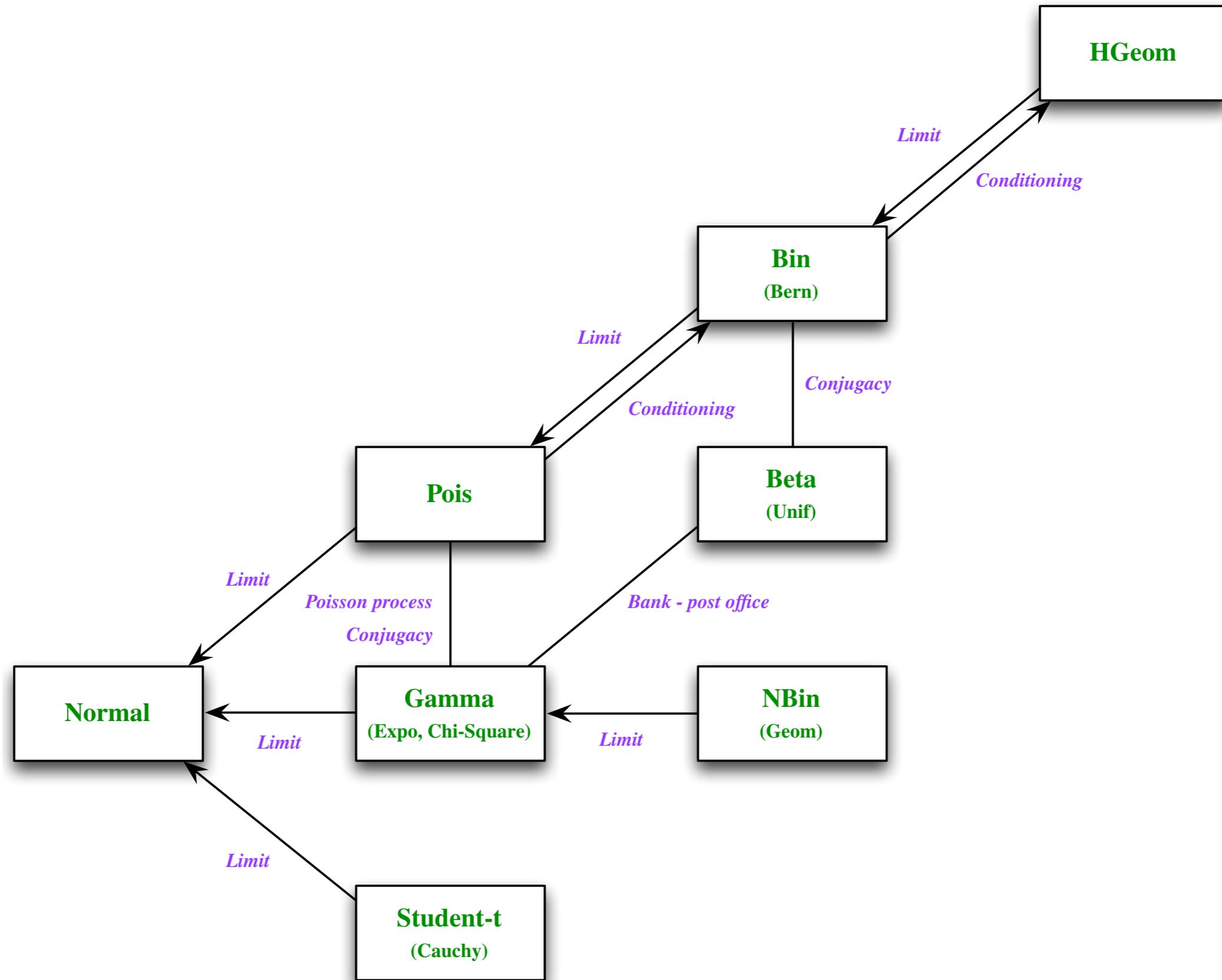
Remember the memoryless property!

# The Weibull Distribution

- Exponential has constant hazard function
- Weibull generalizes this to a hazard that is  $t$  to a power
- much more flexible and realistic than Exponential
- *representation*: a Weibull is an Expo to a power

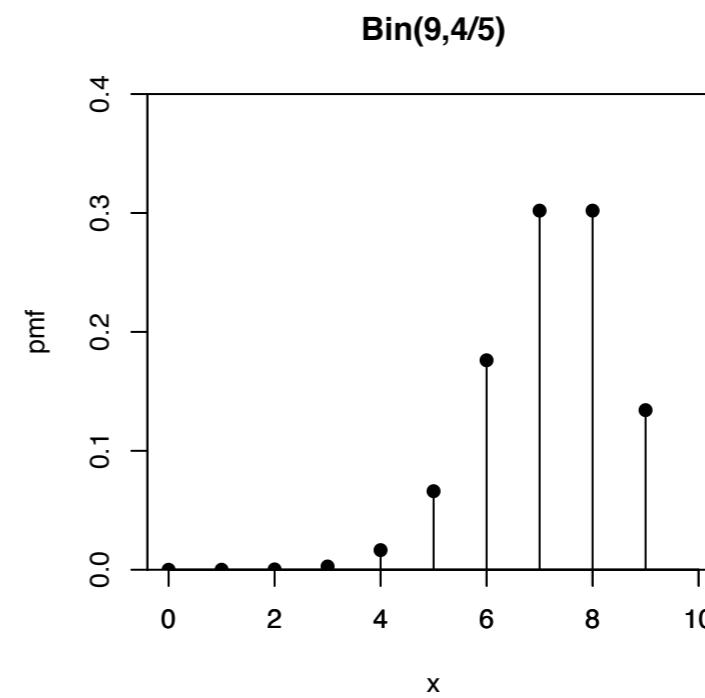
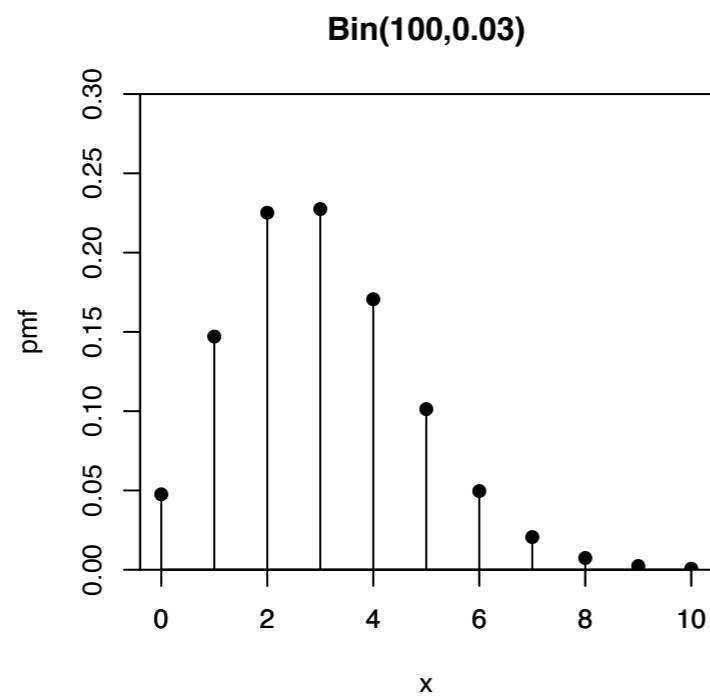
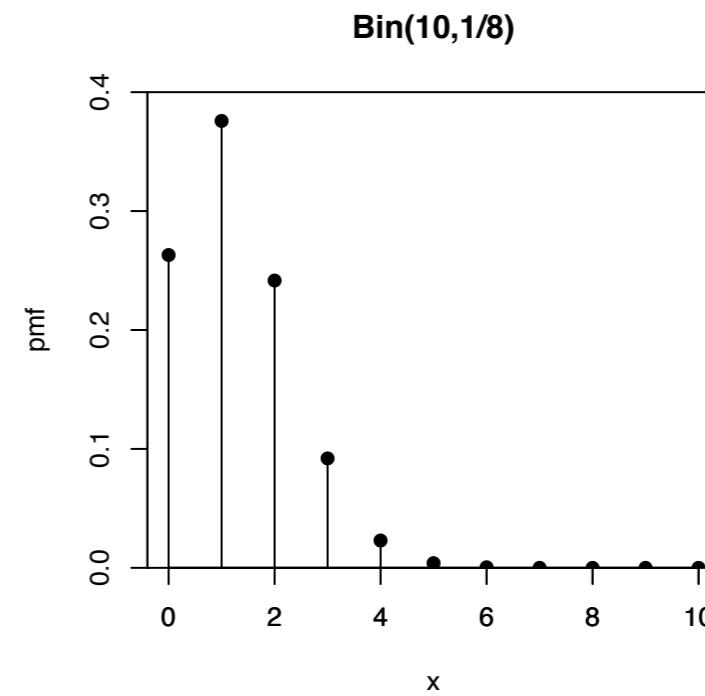
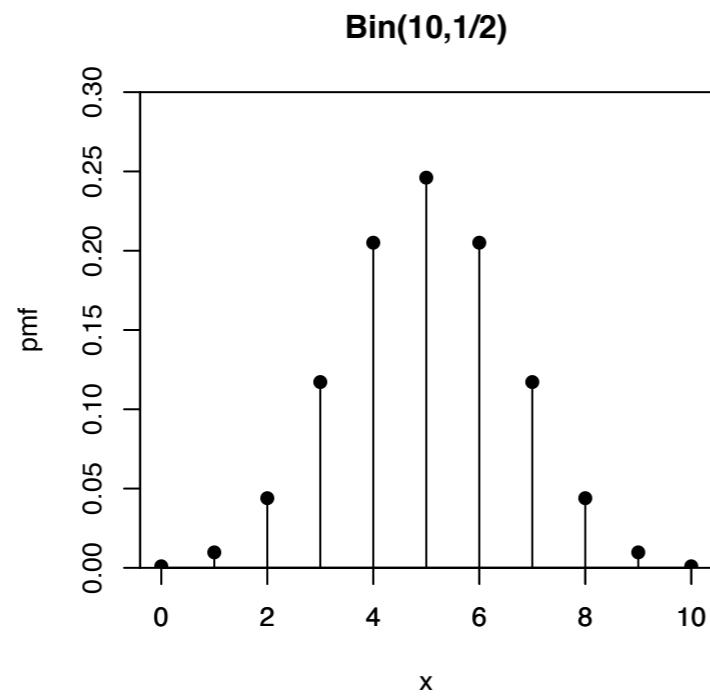


# Family Tree of Parametric Distributions



# Binomial Distribution

**story:**  $X \sim \text{Bin}(n, p)$  is the number of successes in  $n$  independent Bernoulli( $p$ ) trials.



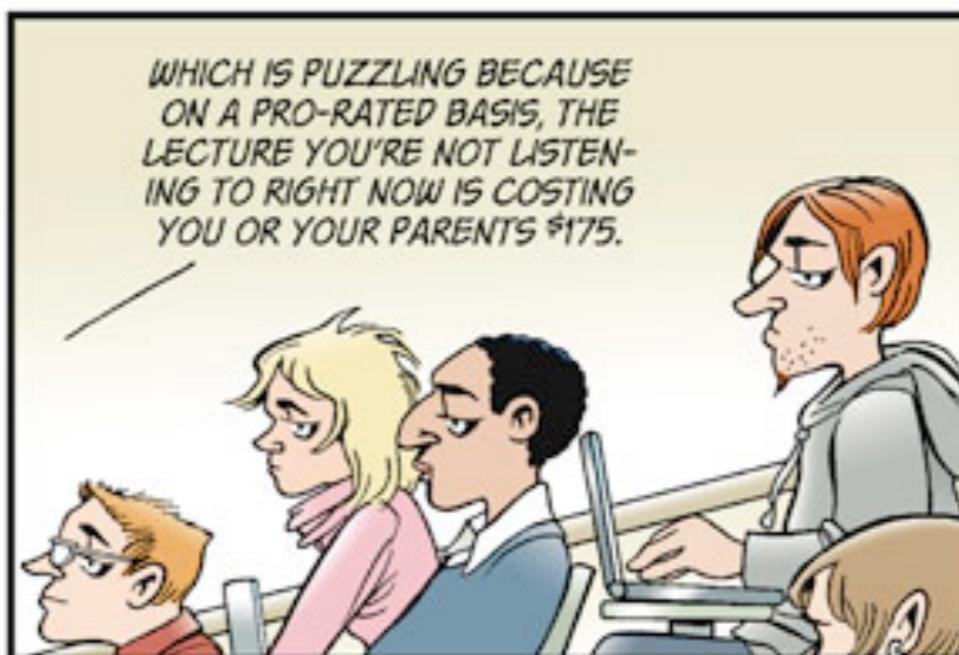
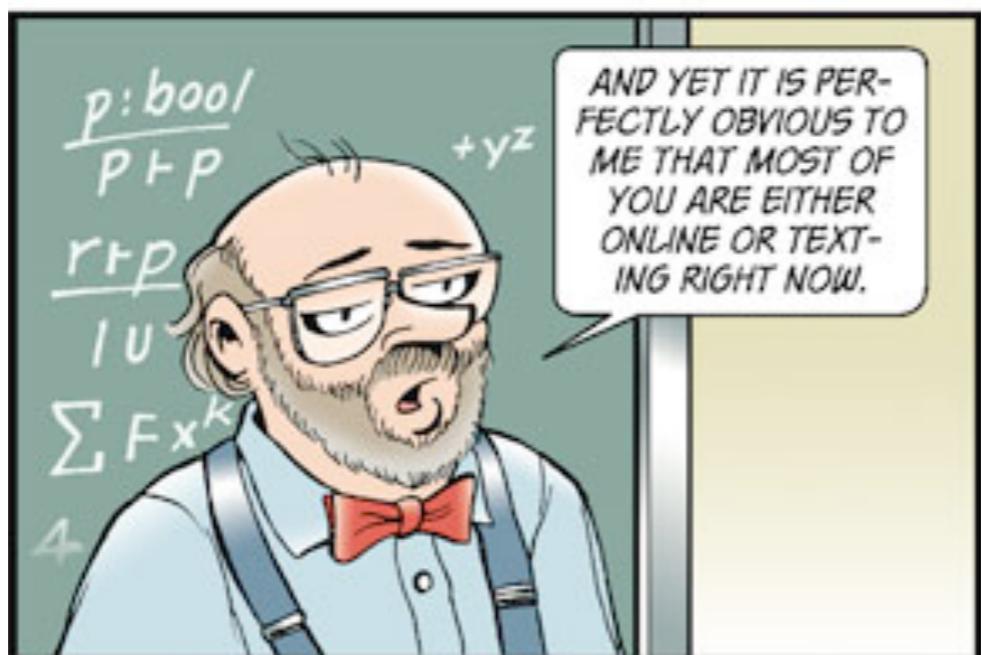
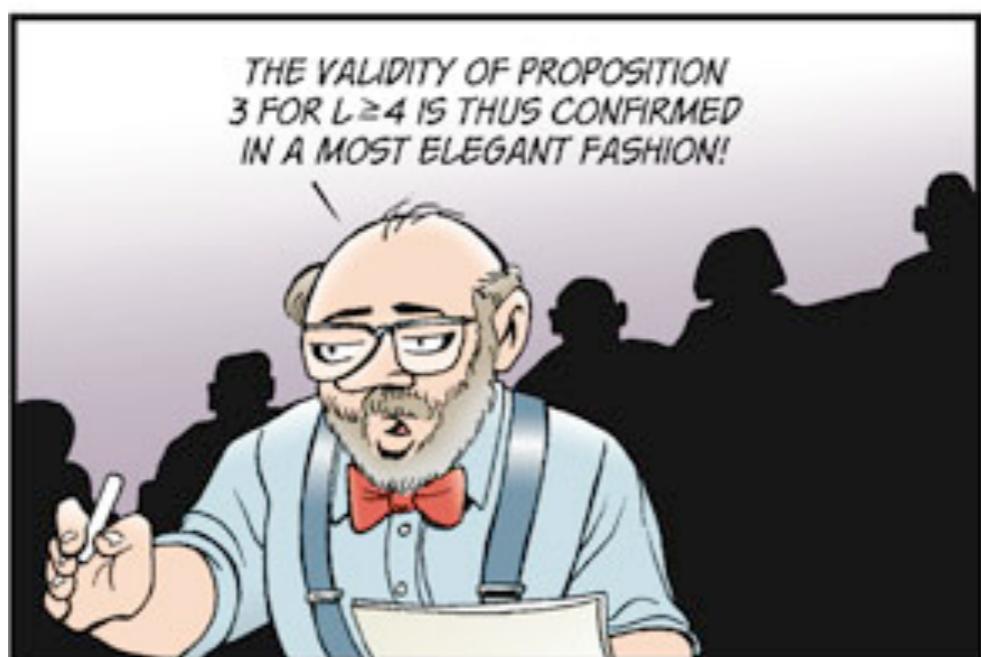
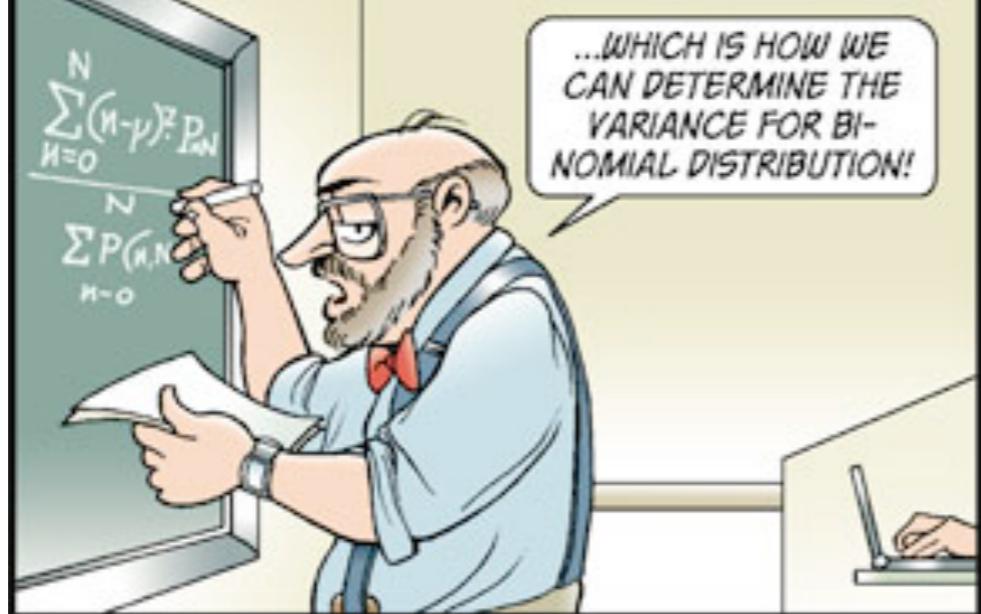
# Binomial Distribution

**story:**  $X \sim \text{Bin}(n, p)$  is the number of successes in  $n$  independent Bernoulli( $p$ ) trials.

**Example:** # votes for candidate A in election with  $n$  voters, where each independently votes for A with probability  $p$

mean is  $np$  (by story and linearity of expectation:  
 $E(X+Y)=E(X)+E(Y)$ )

variance is  $np(1-p)$  (by story and the fact that  
 $\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)$  if  $X, Y$  are uncorrelated)



(Doonesbury)

# Poisson Distribution

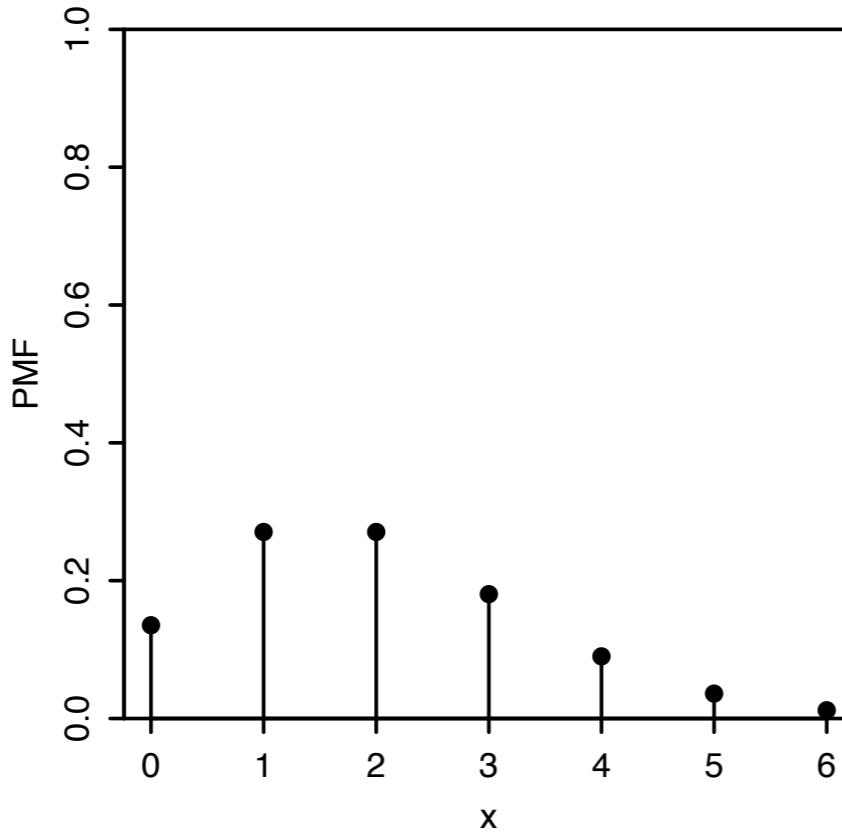
**story:** count number of events that occur, when there are a large number of independent *rare events*.

**Examples:** # mutations in genetics, # of traffic accidents at a certain intersection, # of emails in an inbox  
mean = variance for the Poisson

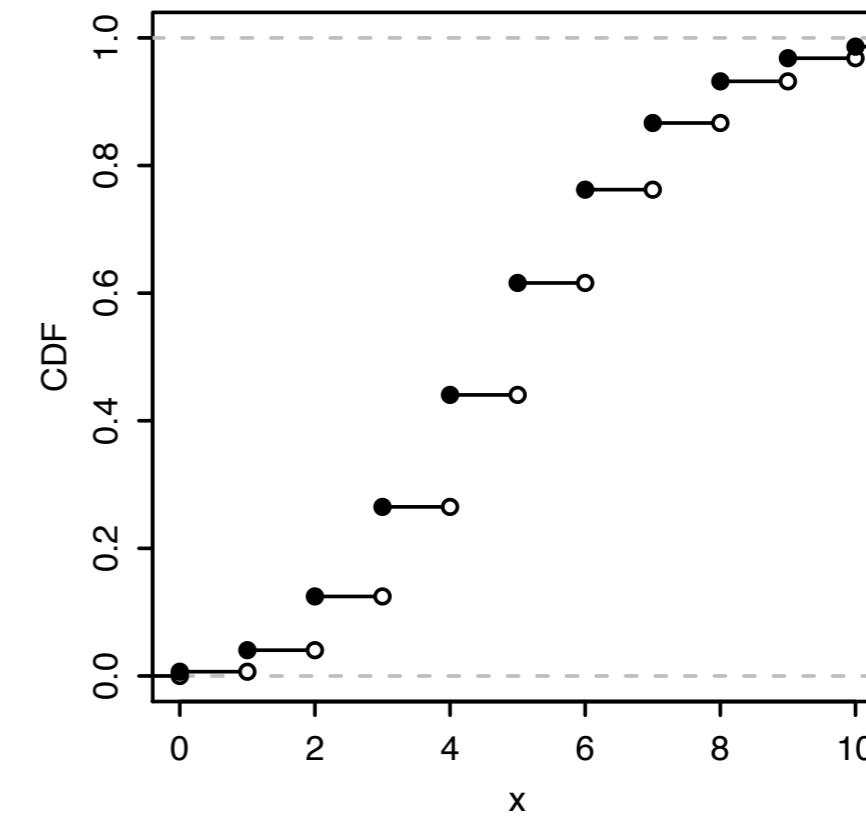
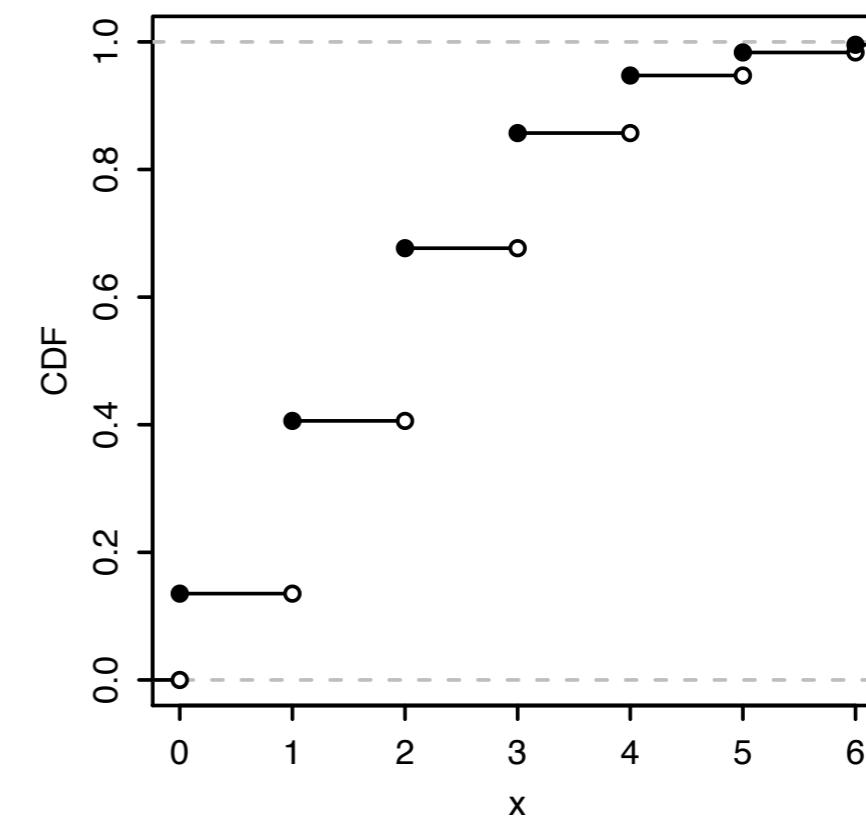
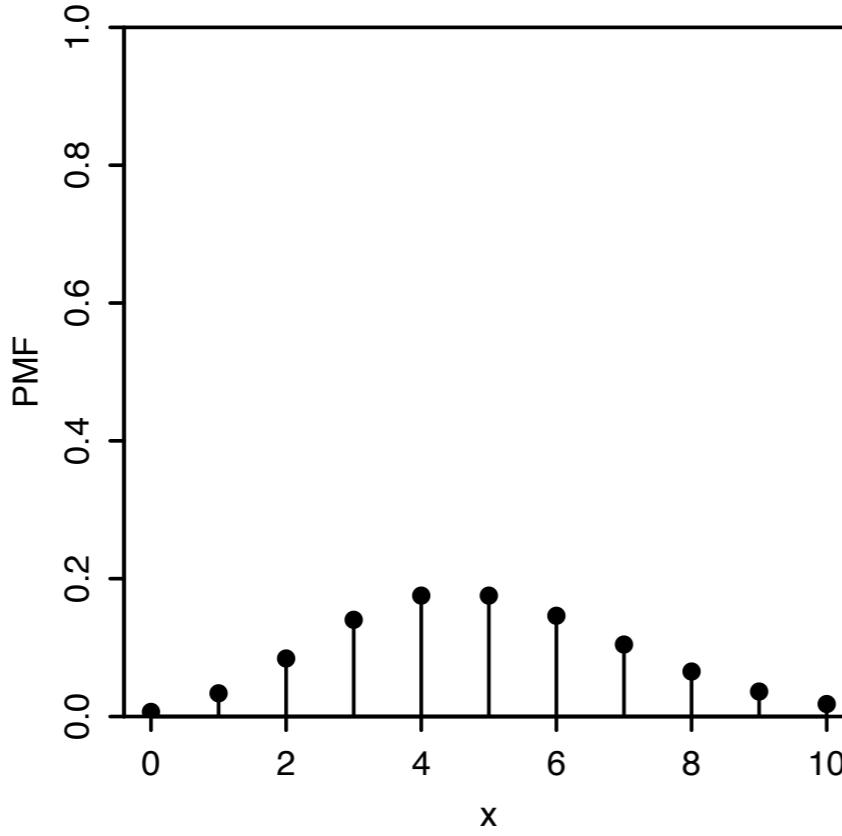
$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

# Poisson PMF, CDF

Pois(2)



Pois(5)



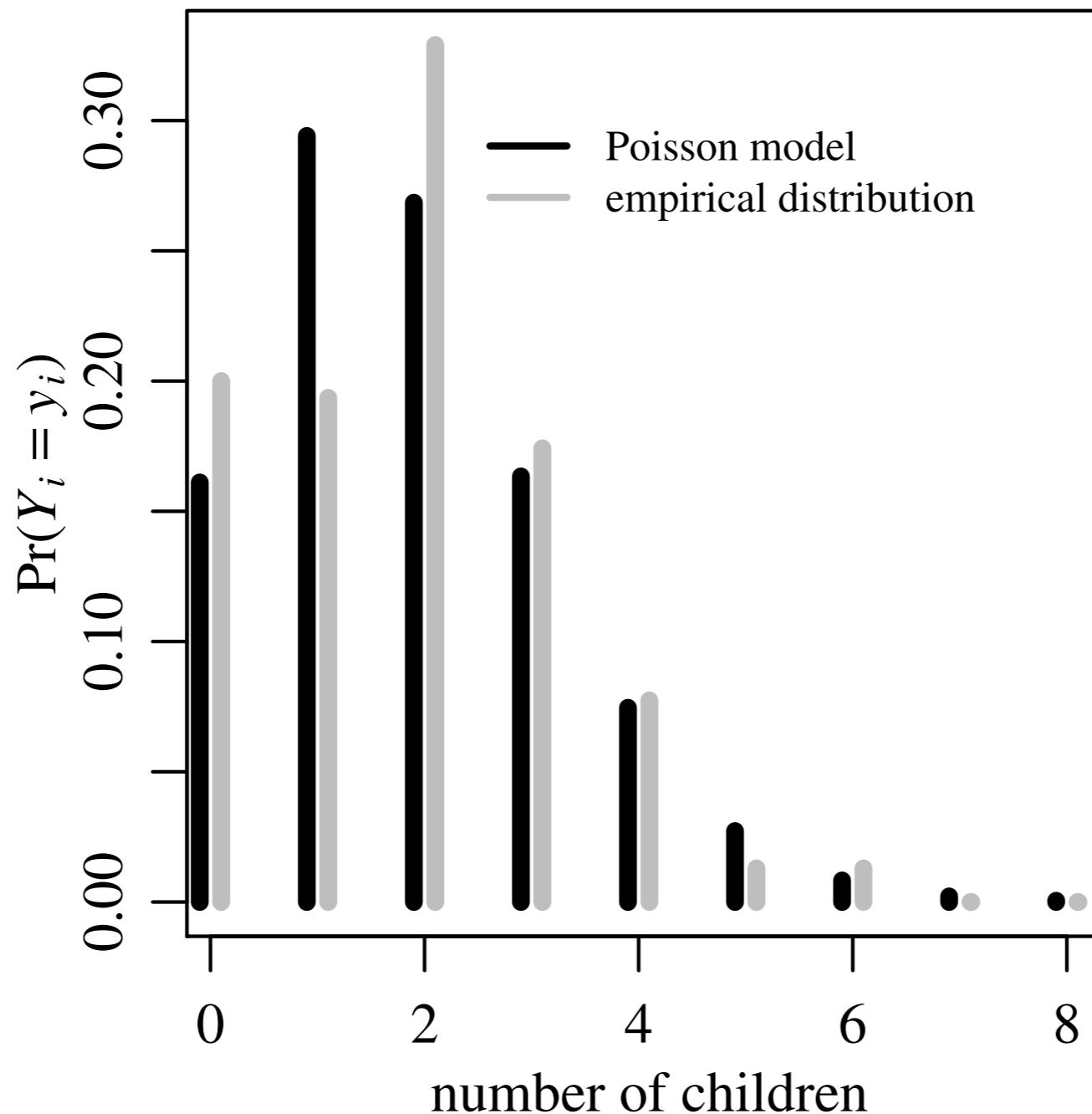
# Poisson Approximation

$$|P(X \in B) - P(N \in B)| \leq \min\left(1, \frac{1}{\lambda}\right) \sum_{j=1}^n p_j^2.$$

if  $X$  is the number of events that occur, where the  $i$ th event has probability  $p_i$ , and  $N \sim \text{Pois}(\lambda)$ , with  $\lambda$  the average number of events that occur.

Example: in matching problem, the probability of no match is approximately  $1/e$ .

# Poisson Model Example

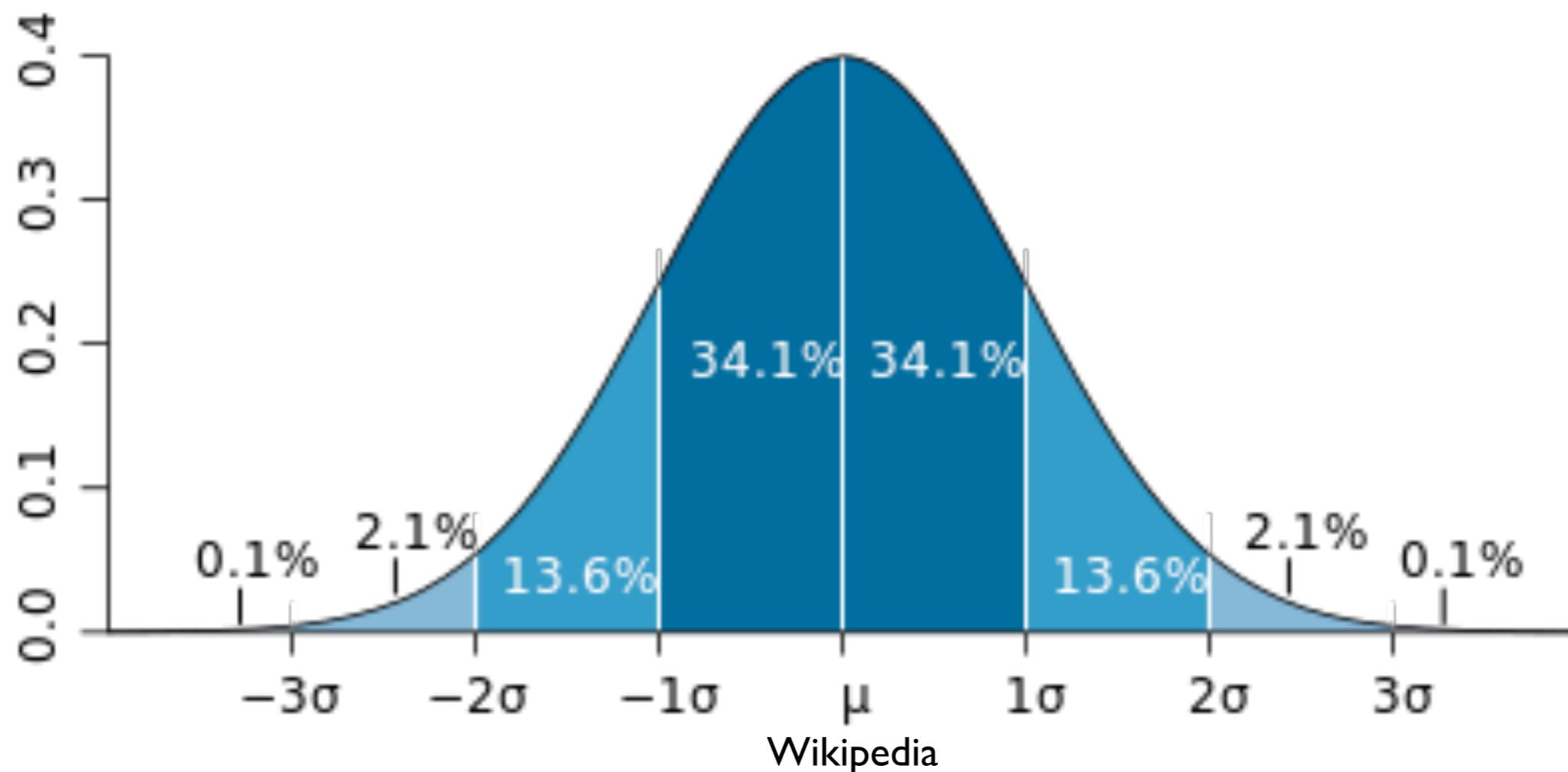


source: Hoff, A First Course in Bayesian Statistical Methods

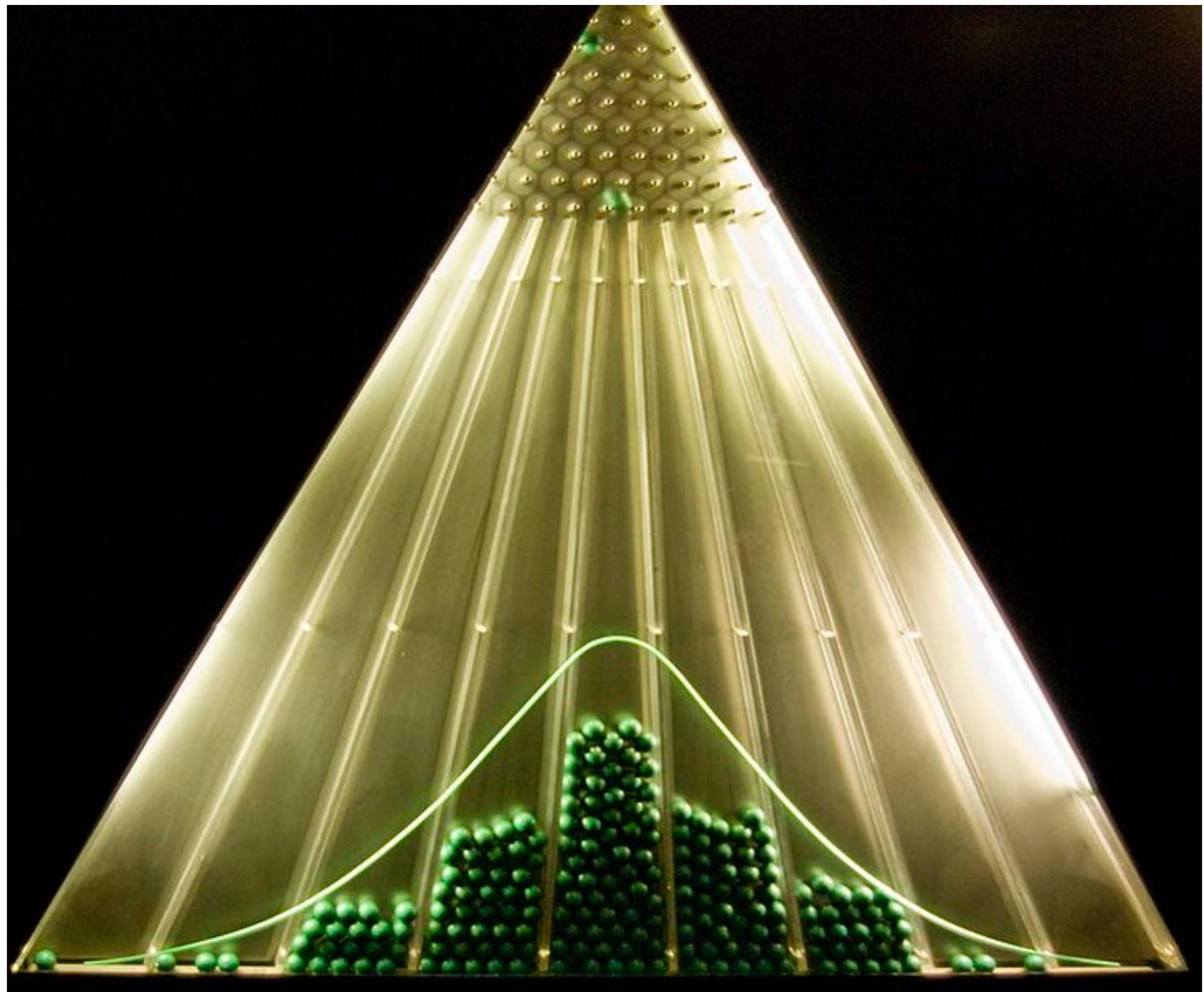
Extensions: Zero-Inflated Poisson, Negative Binomial, ...

# Normal (Gaussian) Distribution

- symmetry
- central limit theorem
- characterizations (e.g., via entropy)
- 68-95-99.7% rule



# Normal Approximation to Binomial



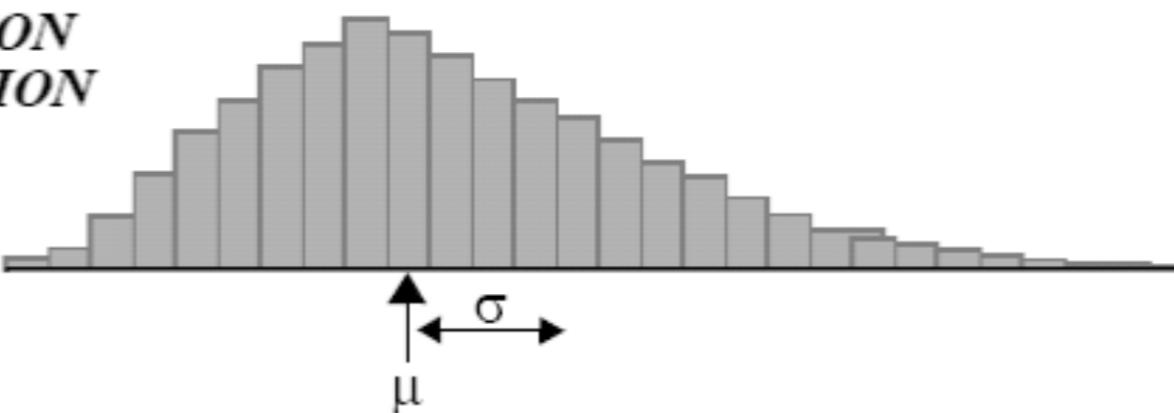
# The Magic of Statistics

---

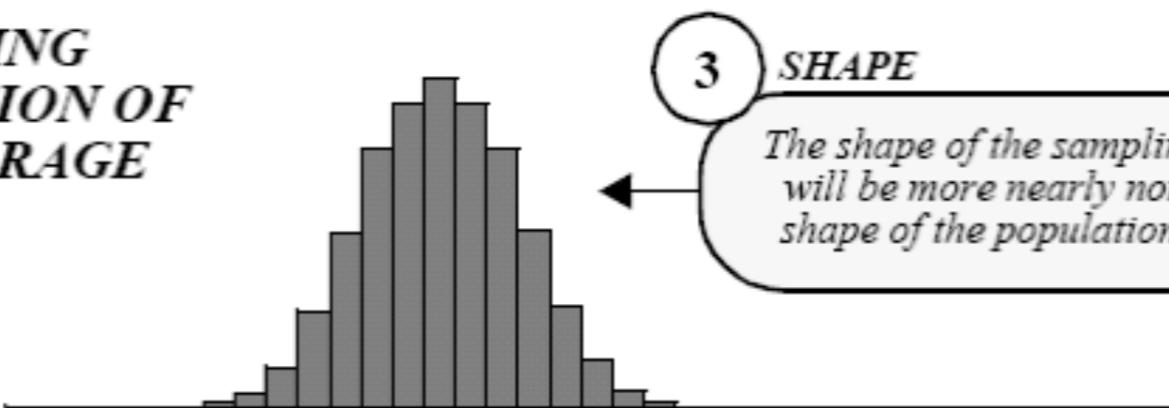
The relationship between the population distribution and the sampling distribution of the average in random sampling

---

**POPULATION  
DISTRIBUTION**



**SAMPLING  
DISTRIBUTION OF  
THE AVERAGE**



3

**SHAPE**

*The shape of the sampling distribution will be more nearly normal than the shape of the population distribution.*

1 **CENTER**

*The sampling distribution is centered on the population mean*

2 **SPREAD**

*Sample averages are closer to the mean than single values; the sampling distribution has*

$$SD(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

# The Evil Cauchy Distribution



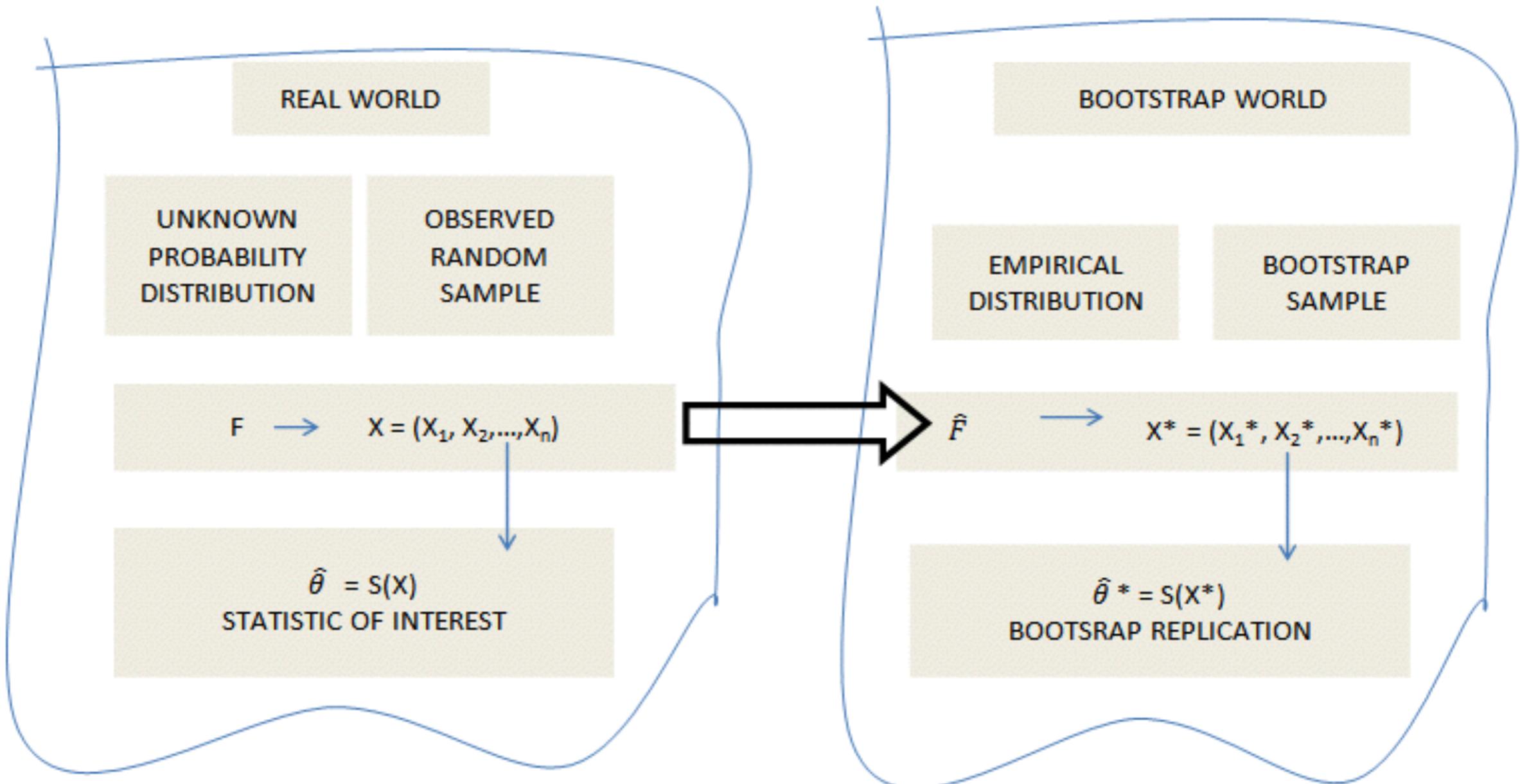
<http://www.etsy.com/shop/NausicaaDistribution>

# Bootstrap (Efron, 1979)

data:	3.142	2.718	1.414	0.693	1.618
	1.414	2.718	0.693	0.693	2.718
	1.618	3.142	1.618	1.414	3.142
reps	1.618	0.693	2.718	2.718	1.414
	0.693	1.414	3.142	1.618	3.142
	2.718	1.618	3.142	2.718	0.693
	1.414	0.693	1.618	3.142	3.142

resample with *replacement*, use empirical distribution to approximate true distribution

# Bootstrap World



source: <http://pubs.sciepub.com/ijfm/3/3/2/>  
which is based on diagram in Efron-Tibshirani book