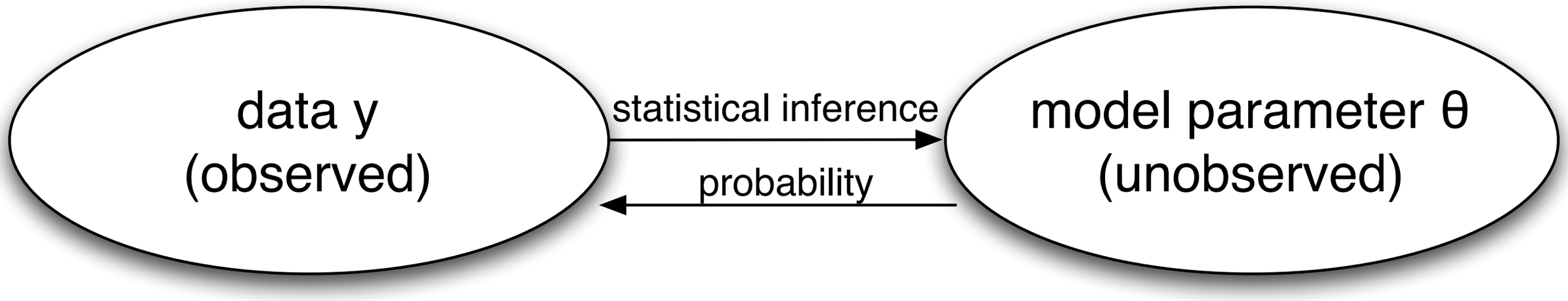
CS109/Stat121/AC209/E-109
Data Science
Statistical Models

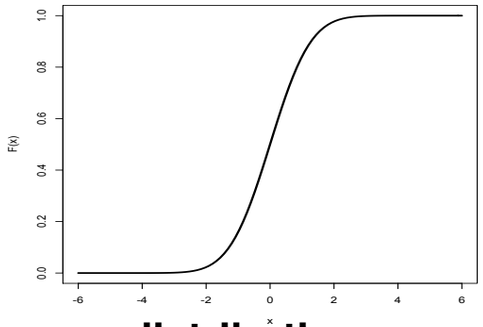
Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



This Week

- HW1 due this Thursday - start last week!
- Course dropbox is now active at http://isites.harvard.edu/k99240 (Harvard ID required). Please follow the submission instructions carefully, and do a test well in advance of the HW1 deadline.
- Friday lab **10-11:30 am** in MD G115
  - *Pandas* with Rahul, Brandon, and Steffen

Road Map to Probability

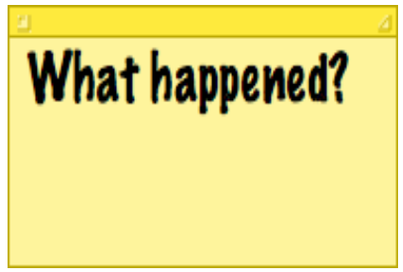


distributions

CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF

random variables

X



events

X≤x
X=x

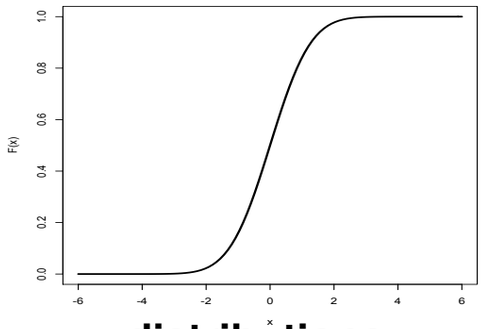
numbers

P(X≤x)=F(x)
P(X=x)

E(X),Var(X),SD(X)

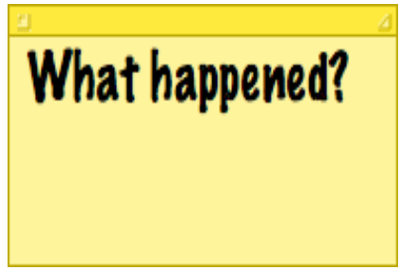
for more about probability: stat110.net

Road Map to Probability



distributions

random variables



events

numbers

CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF

generate →

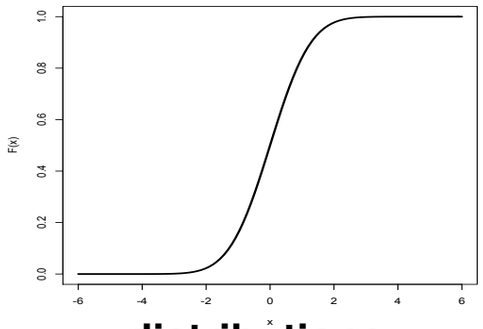
X

X≤x
X=x

P(X≤x)=F(x)
P(X=x)

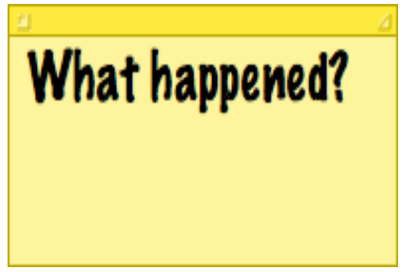
E(X),Var(X),SD(X)

Road Map to Probability



distributions

random variables



events

numbers

CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF

generate →

X

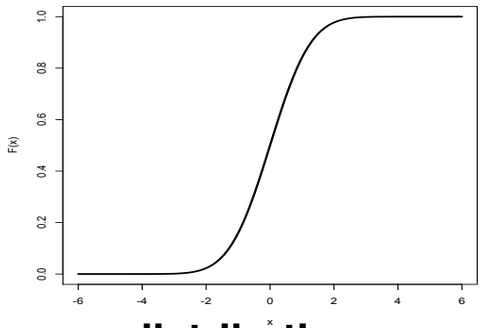
→

X≤x
X=x

P(X≤x)=F(x)
P(X=x)

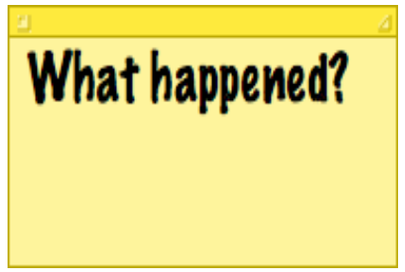
E(X),Var(X),SD(X)

Road Map to Probability



distributions

random variables



events

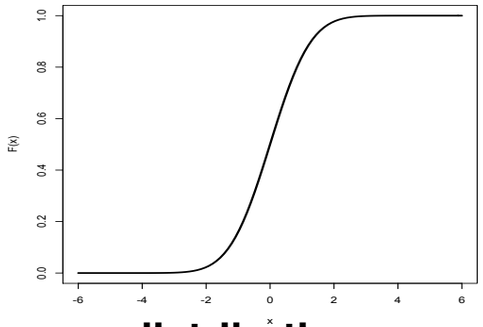
numbers

CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF

generate → **X** → **X≤x** **X=x** → P → **P(X≤x)=F(x)** **P(X=x)**

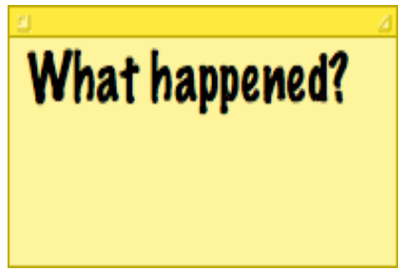
E(X),Var(X),SD(X)

Road Map to Probability



distributions

random variables



events

numbers

CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF

generate

X

X≤x
X=x

P

P(X≤x)=F(x)
P(X=x)

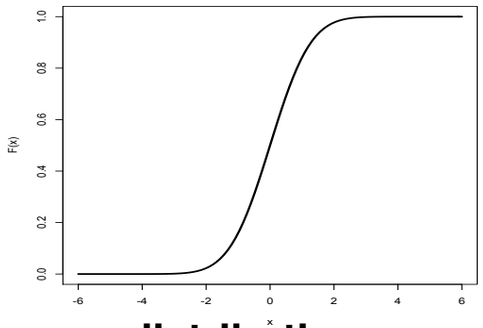
function of r.v.

X,X^2,X^3,…
g(X)

E(X),Var(X),SD(X)

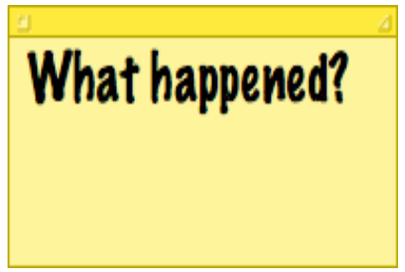
E(X),E(X^2),E(X^3),…
E(g(X))

Road Map to Probability



distributions

random variables



events

numbers

CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF

generate

X

X≤x
X=x

P

P(X≤x)=F(x)
P(X=x)

E(X),Var(X),SD(X)

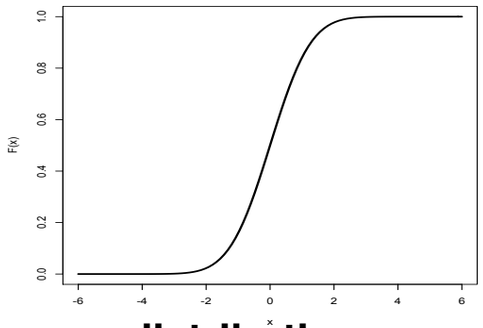
function of r.v.

$X, X^2, X^3, \ldots$
g(X)

LOTUS

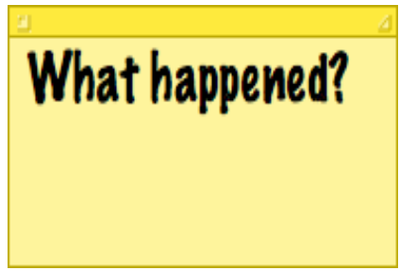
$E(X), E(X^2), E(X^3), \ldots$
E(g(X))

Road Map to Probability



distributions

random variables



events

numbers

CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF

generate

X

X≤x
X=x

P

P(X≤x)=F(x)
P(X=x)

E(X),Var(X),SD(X)

function of r.v.

X,X^2,X^3,…
g(X)

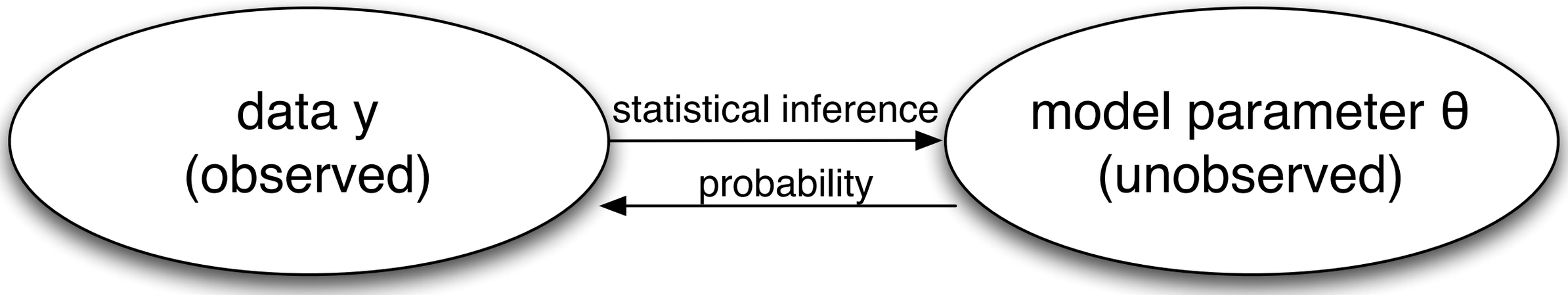
LOTUS

E(X),E(X^2),E(X^3),…
E(g(X))

find the MGF

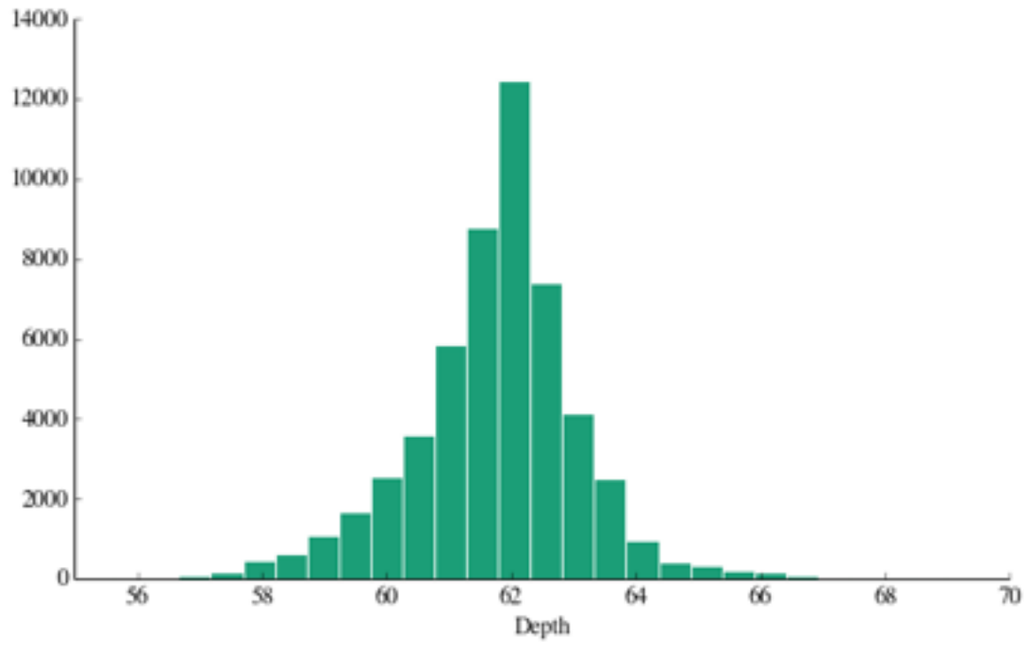
What is a statistical model?

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



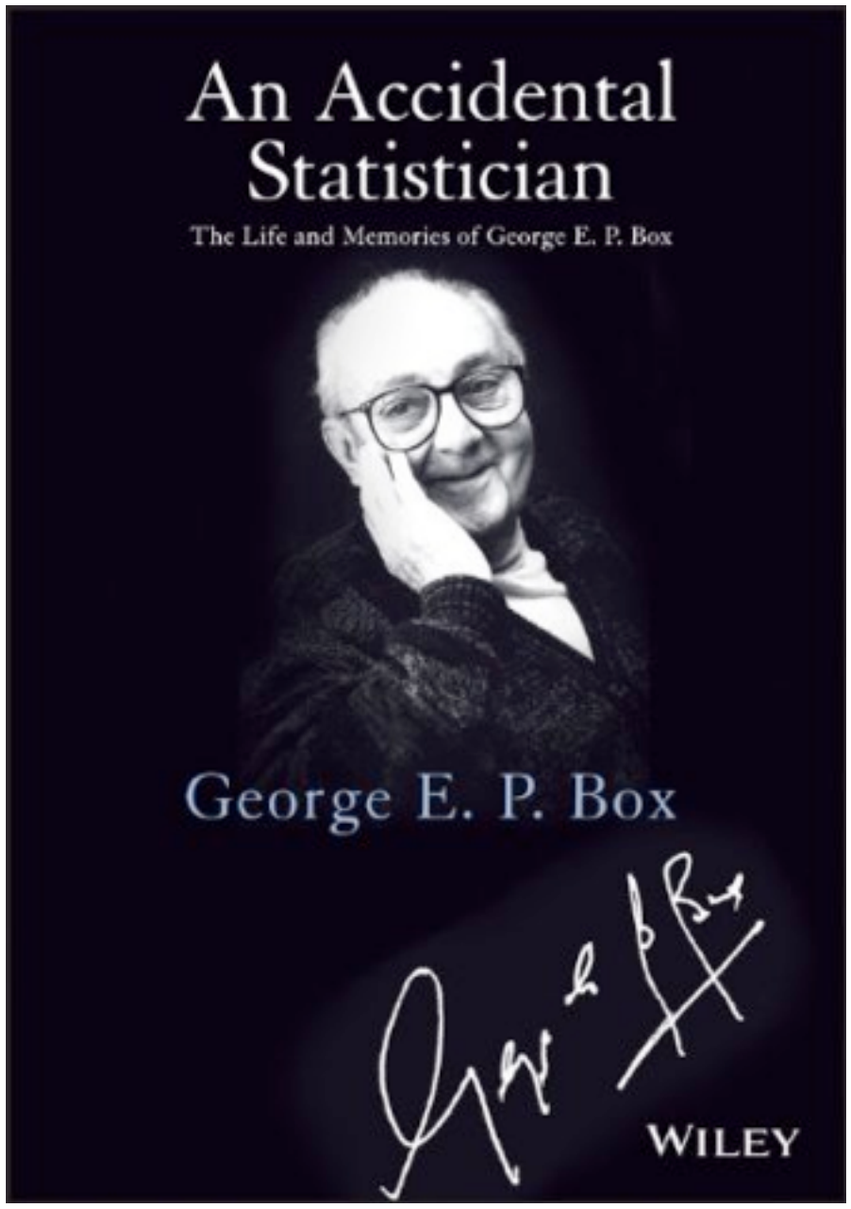
Parametric vs. Nonparametric

- 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!



What *good* is a statistical model?

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

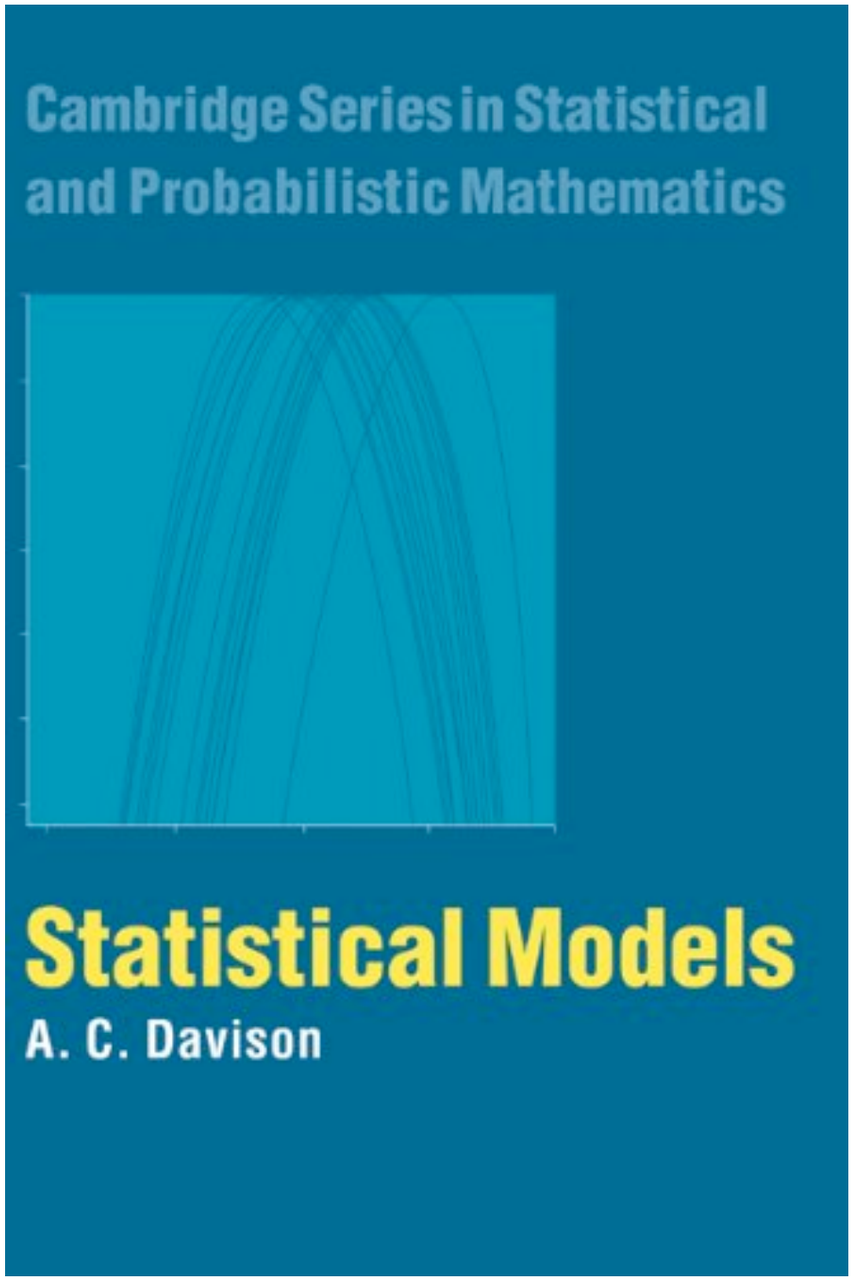


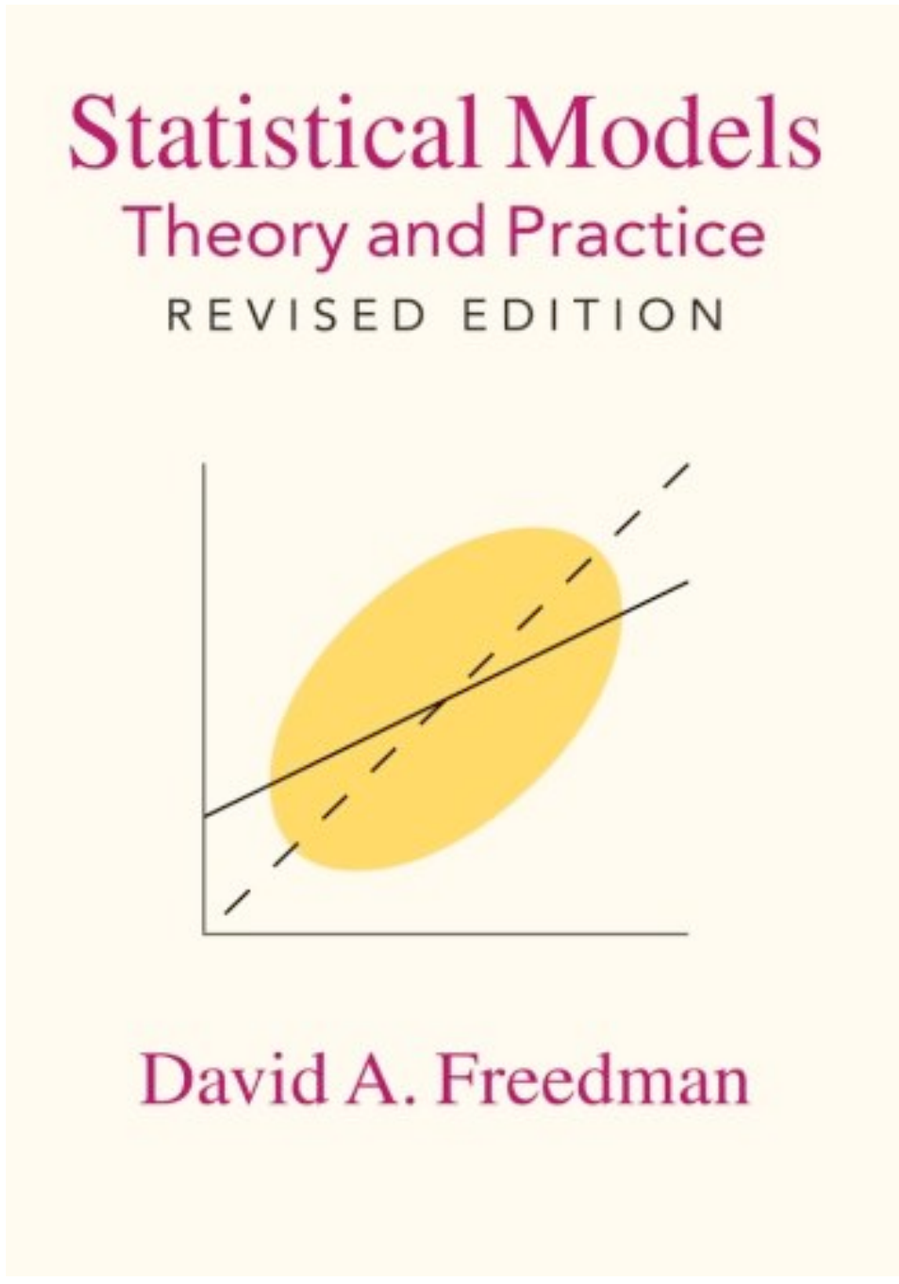
Jorge Luis Borges,
“On Exactitude in Science”

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

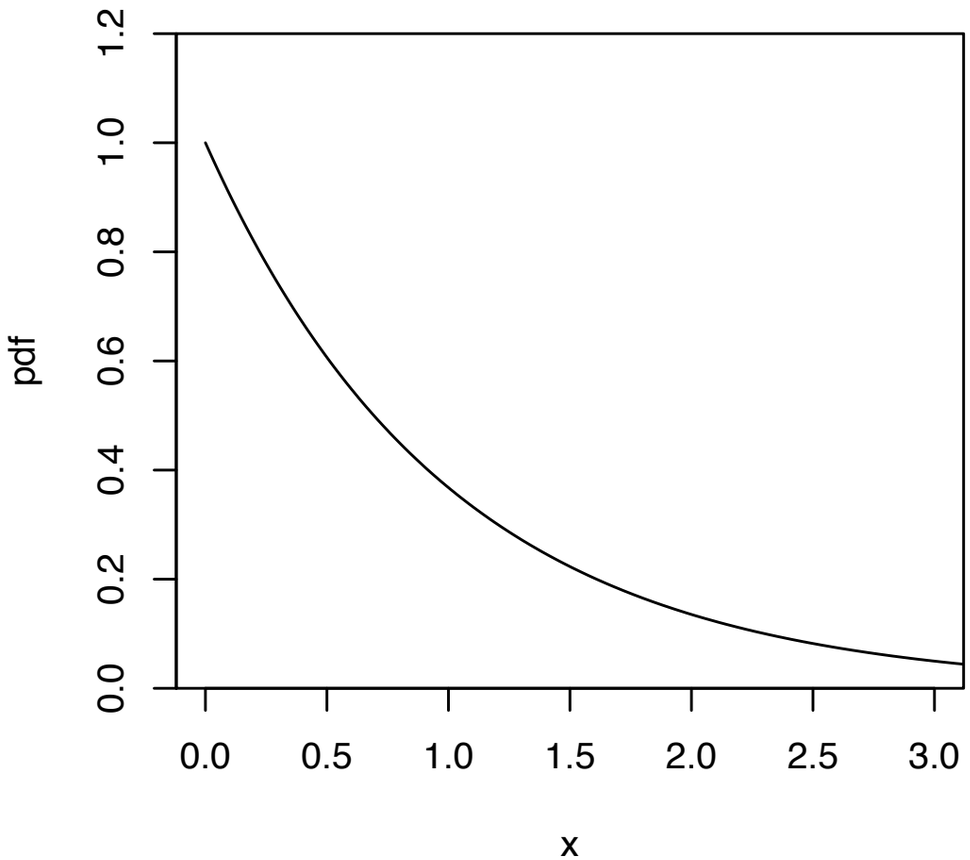
Statistical Models: Two Books

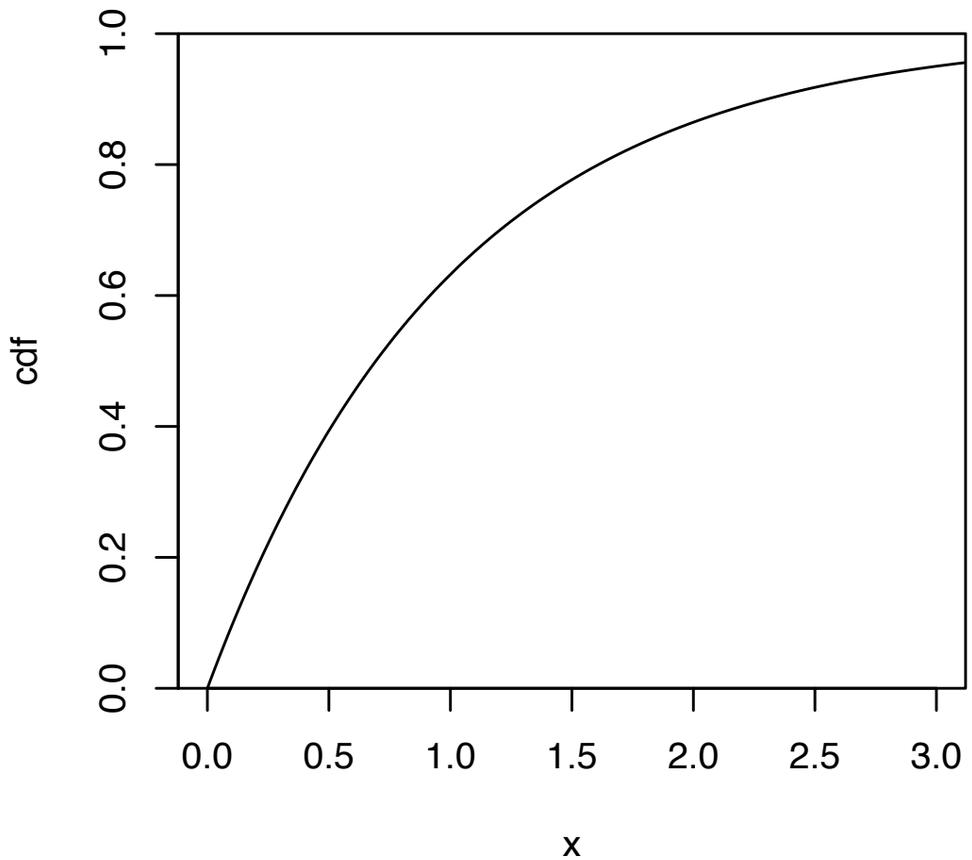




Parametric Model Example: Exponential Distribution

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

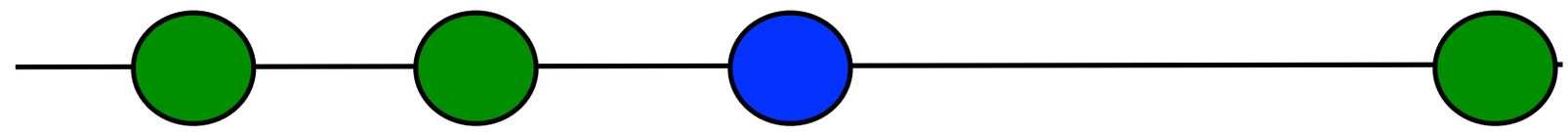




Remember the memoryless property!

Length-Biasing Paradox

What is the waiting time for a bus?

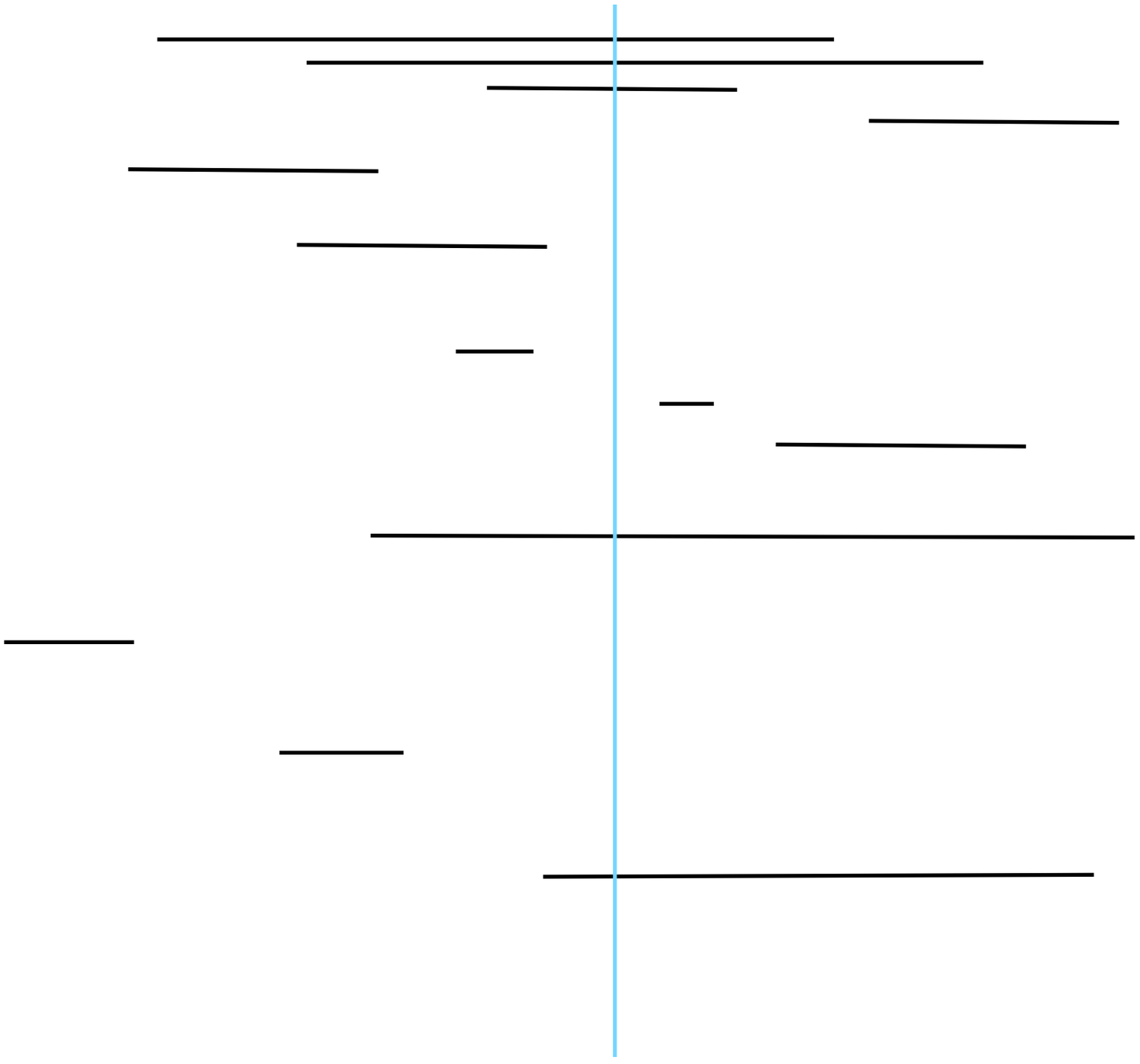


timeline

For i.i.d. Exponential arrivals, your average waiting time is the same as the average time between buses!

Length-Biasing Paradox

How would you measure the average prison sentence?



Exponential Distribution

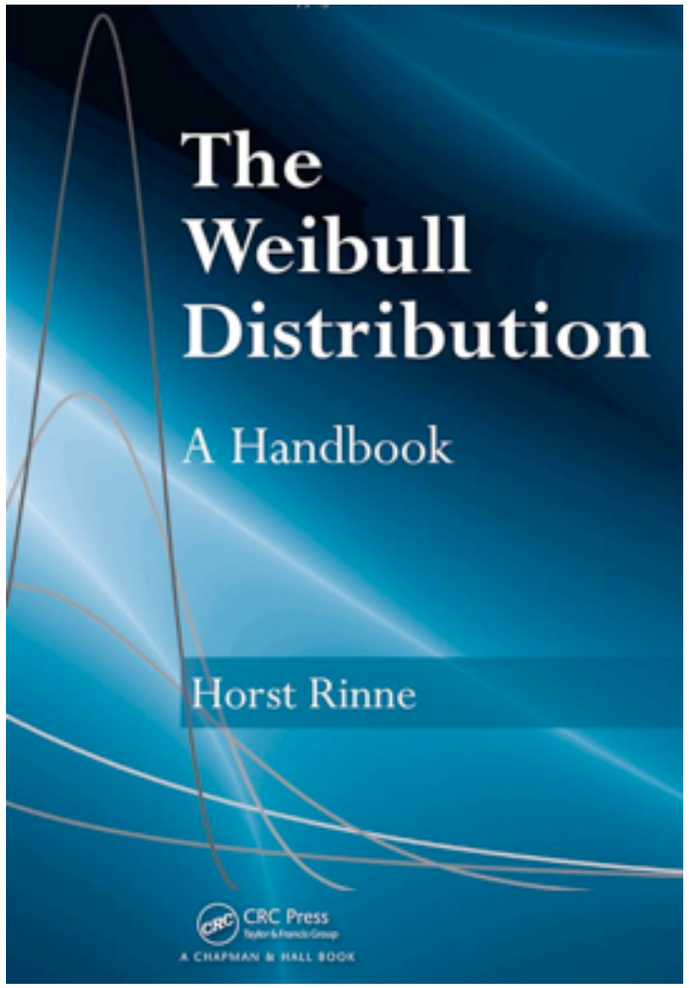
$$f(x) = \lambda e^{-\lambda x}, x > 0$$

- Exponential is *characterized* by memoryless property
- and *characterized* by having constant hazard function
- 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

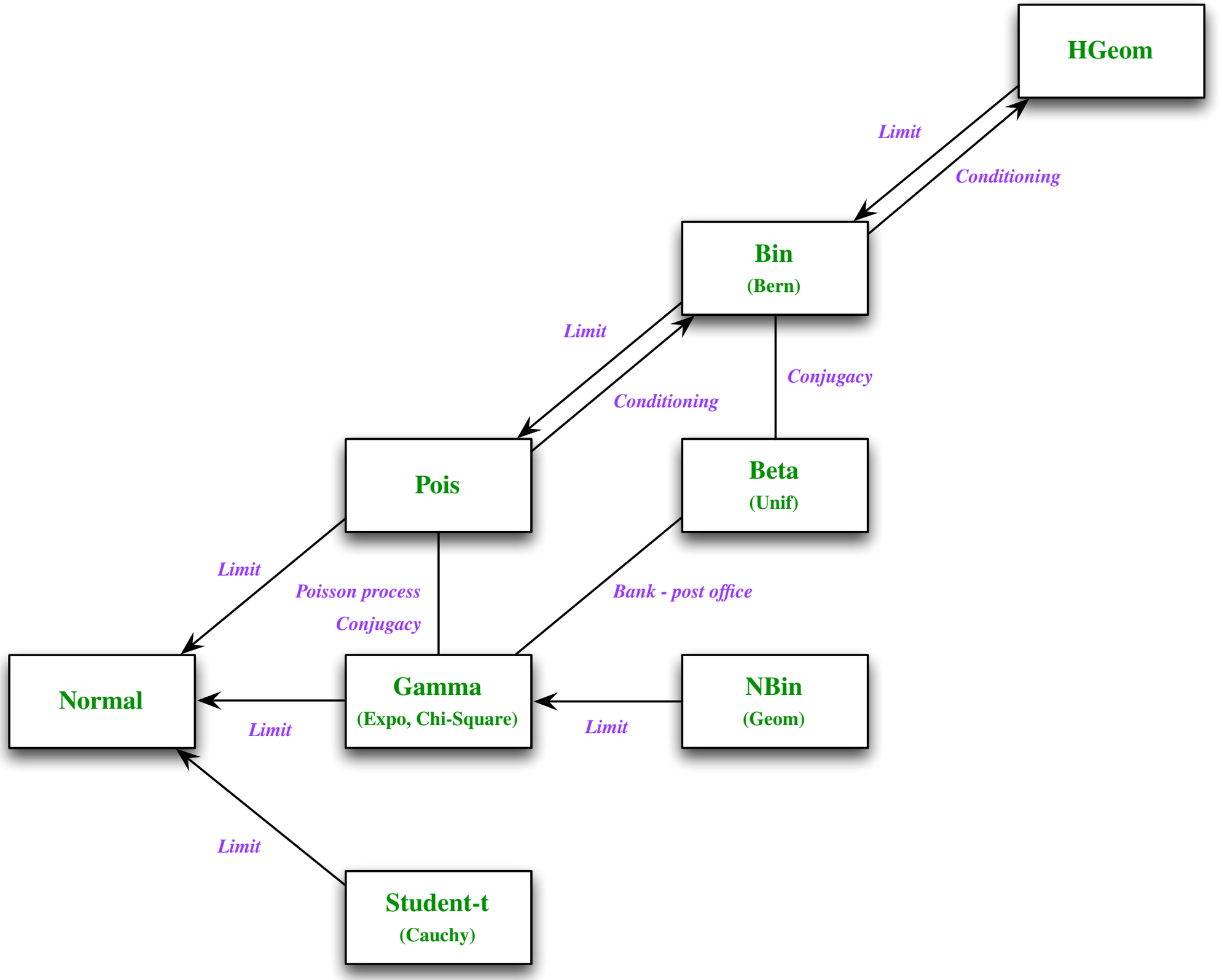
- Exponential has constant hazard function
- Weibull generalizes this to a hazard that is t to a power
- much more flexible and realistic than Exponential
- *representation*: a Weibull is an Expo to a power



The Evil Cauchy Distribution

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

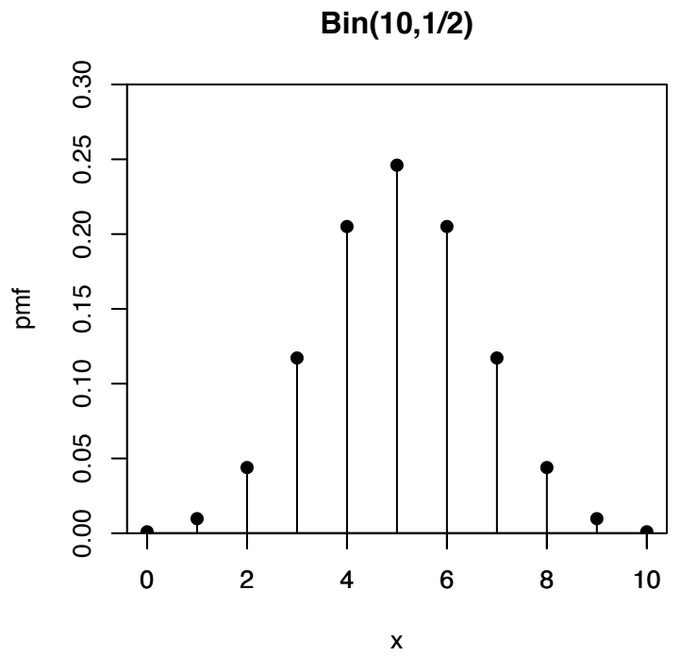
Family Tree of Parametric Distributions

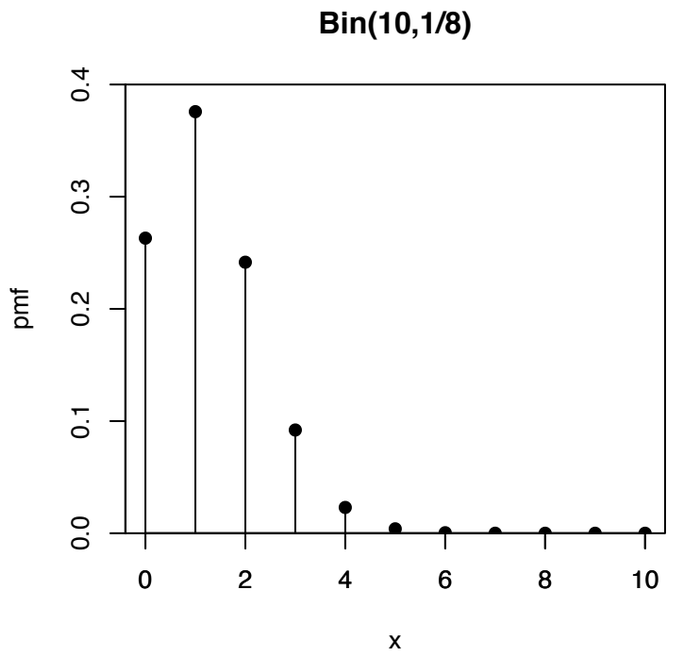


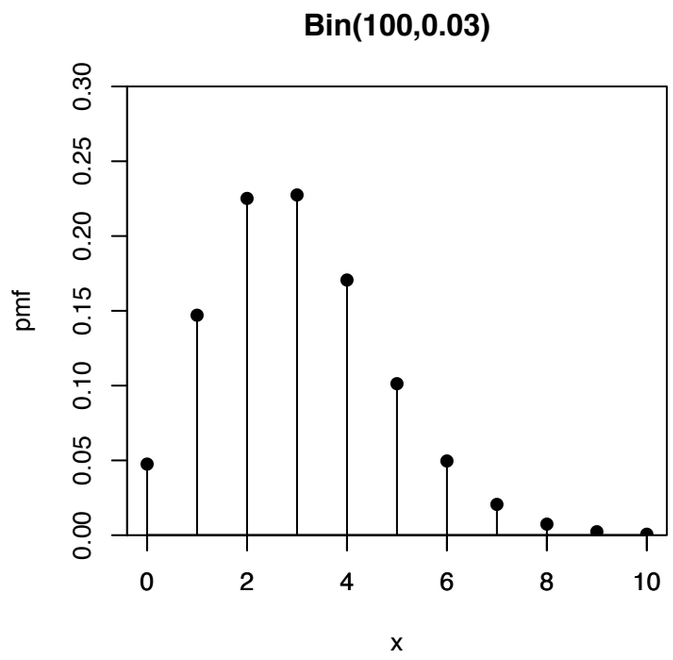
Blitzstein-Hwang, Introduction to Probability

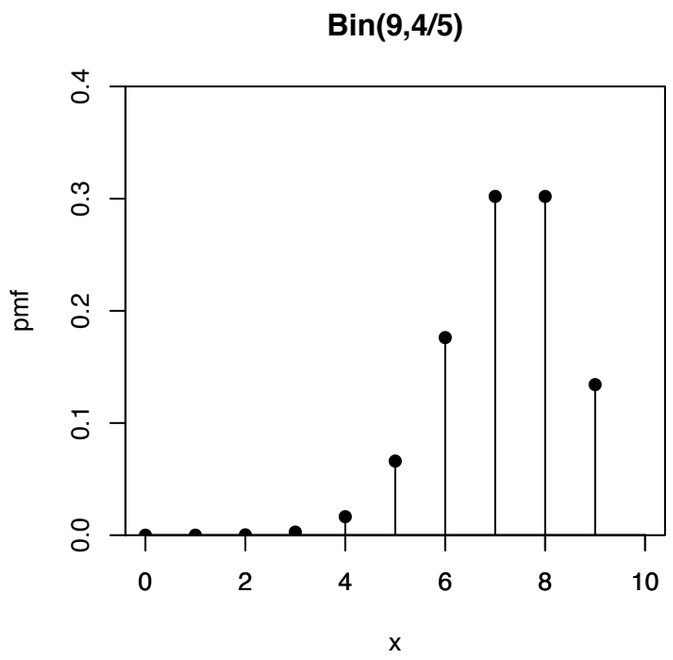
Binomial Distribution

story: X~Bin(n,p) is the number of successes in n independent Bernoulli(p) trials.









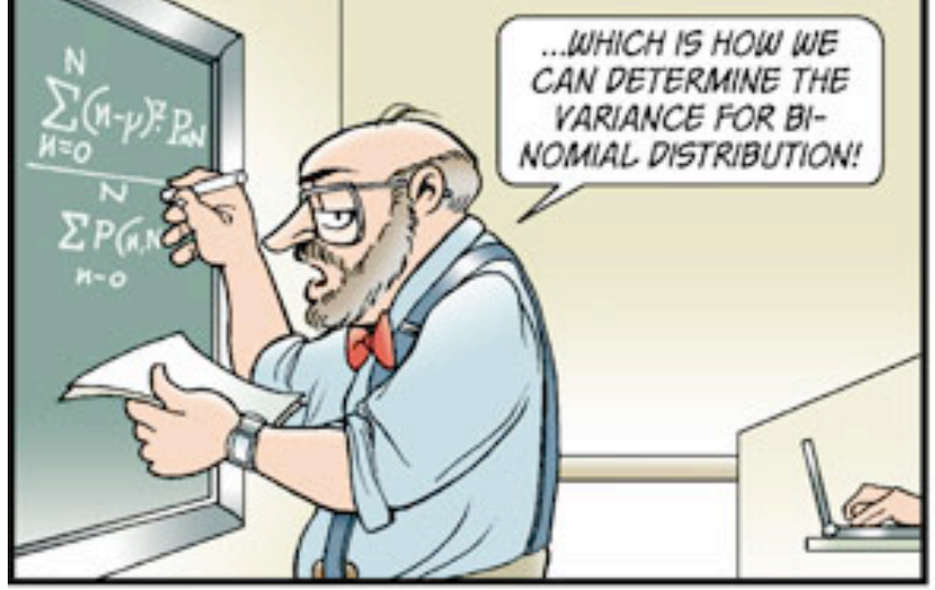
Binomial Distribution

story: X~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

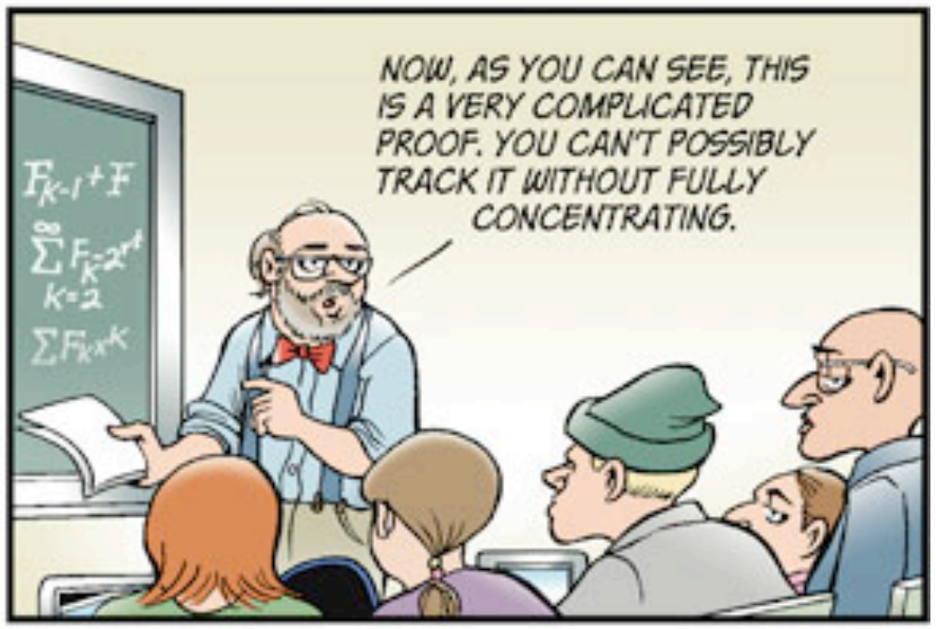
mean is np (by story and linearity of expectation: E(X+Y)=E(X)+E(Y))

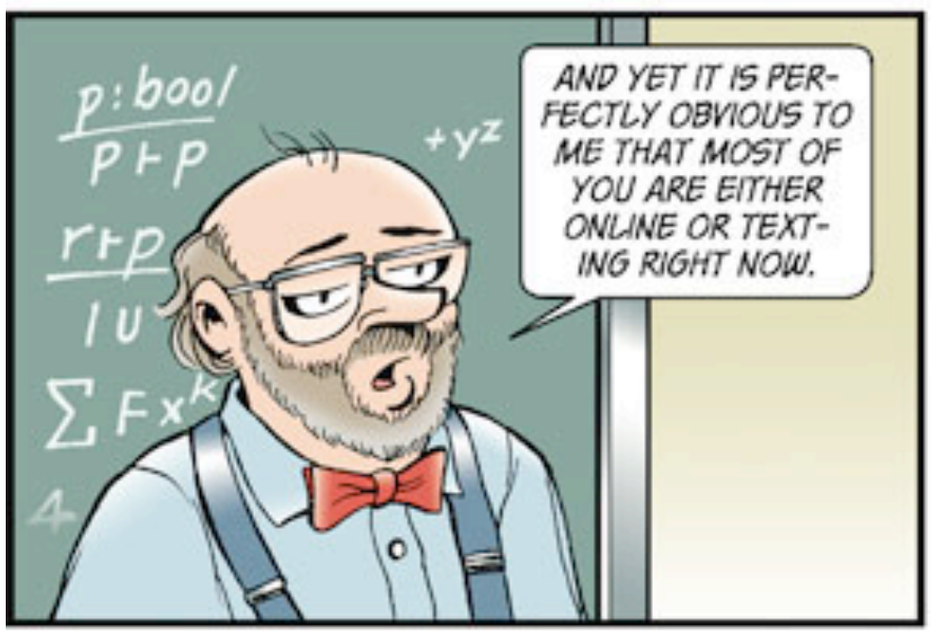
variance is np(1-p) (by story and the fact that Var(X+Y)=Var(X)+Var(Y) if X,Y are uncorrelated)

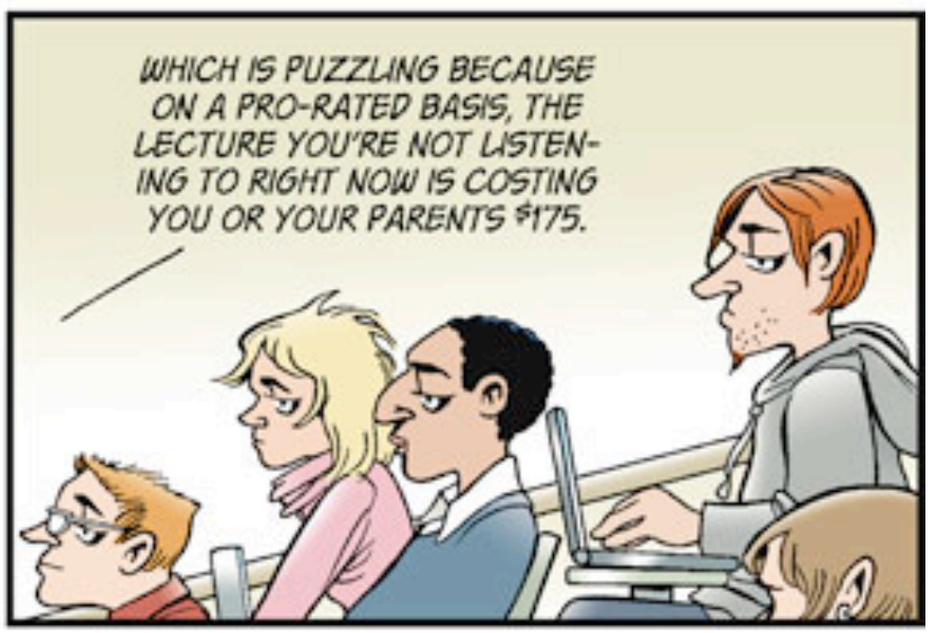

...WHICH IS HOW WE CAN DETERMINE THE VARIANCE FOR BI-NOMIAL DISTRIBUTION!


PRETTY GREAT, RIGHT? ANY-ONE SEE **THAT** COMING?


THE VALIDITY OF PROPOSITION 3 FOR L≥4 IS THUS CONFIRMED IN A MOST ELEGANT FASHION!


NOW, AS YOU CAN SEE, THIS IS A VERY COMPLICATED PROOF. YOU CAN'T POSSIBLY TRACK IT WITHOUT FULLY CONCENTRATING.

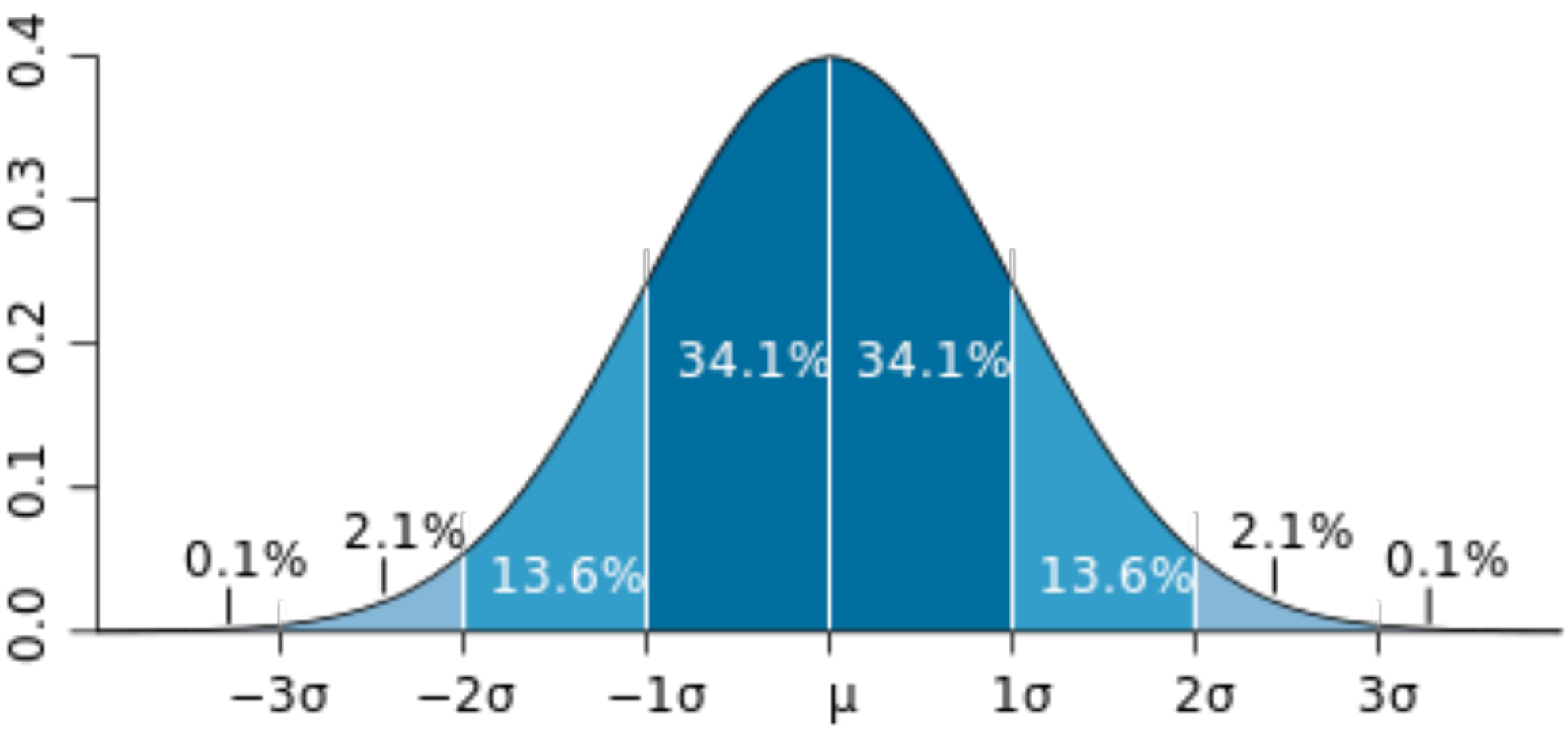

AND YET IT IS PER-FECTLY OBVIOUS TO ME THAT MOST OF YOU ARE EITHER ONLINE OR TEXT-ING RIGHT NOW.


WHICH IS PUZZLING BECAUSE ON A PRO-RATED BASIS, THE LECTURE YOU'RE NOT LISTEN-ING TO RIGHT NOW IS COSTING YOU OR YOUR PARENTS $175.

(Doonesbury)

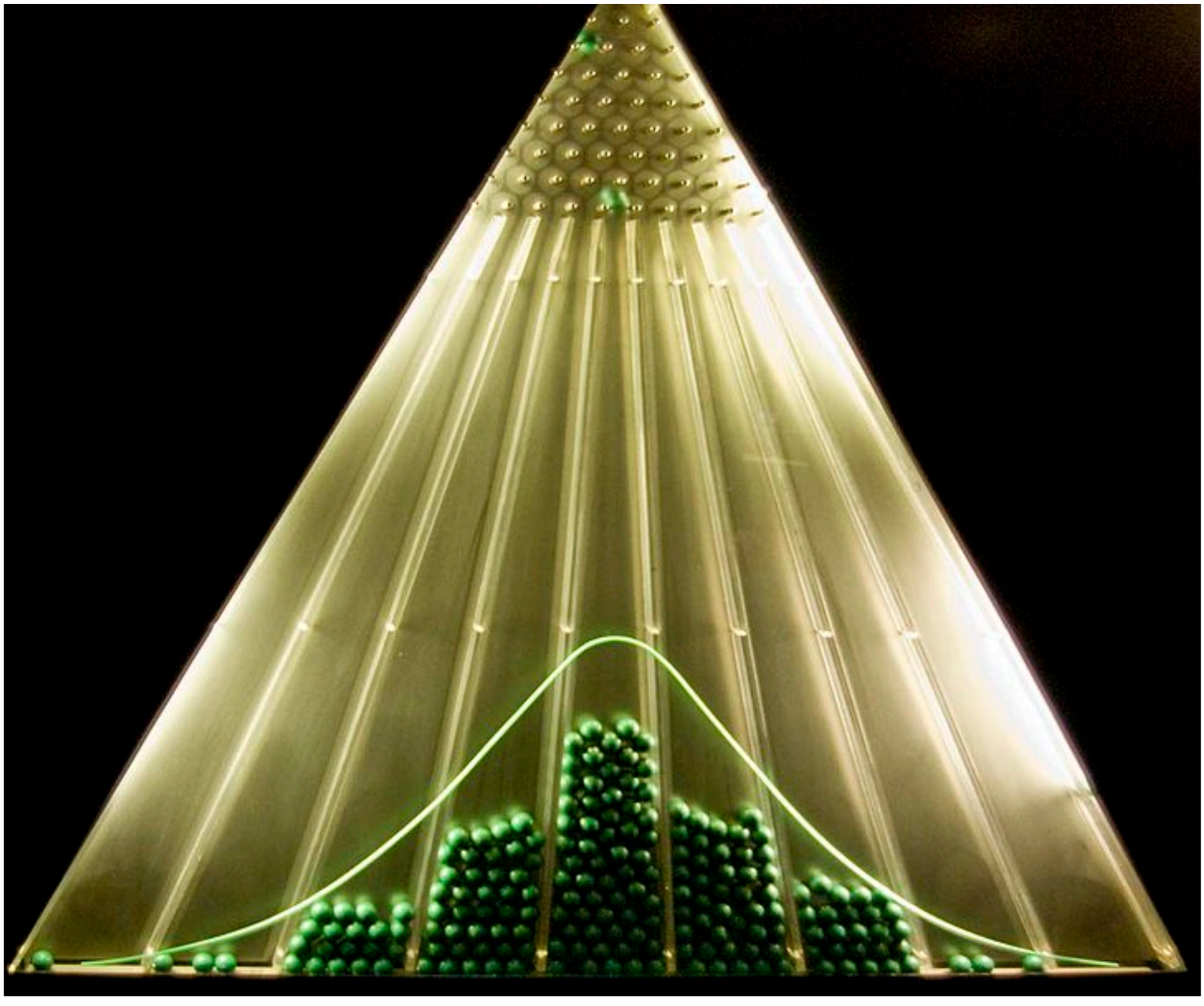
Normal (Gaussian) Distribution

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



Wikipedia

Normal Approximation to Binomial



Wikipedia

Bootstrap

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