L2: Statistics

Hobson Lane, UC San Diego
Instructor

UC San Diego EXTENSION

UC San Diego SCHOOL OF MEDICINE

Syllabus

Lesson	Title	Topics
1	Data Science for Digital Health	applications, terminology, HIPPA
2	Spreadsheet Data Science	ETL, exploration & visualization
3	Statistics, Privacy, Ethics	causality, correlation, MLE
4	Clinical Data Science & ML	PII, prescriptive vs descriptive
5	Deep Learning & AI	neural nets, radiology, CV
6	Hospital Performance Modeling	time series, unintended conseq.
7	Population Health & Epi	GIS, spatio-temporal modeling
8	Healthcare Public Policy	scoping review, gap analysis of diabetes
9	Natural Language Processing	IA, summarization, text mining
10	Bioinformatics	DNA, RNA, proteins, algorithms
Project	Train a healthcare ML model	find/download data, ETL,

Agenda

What is statistics?

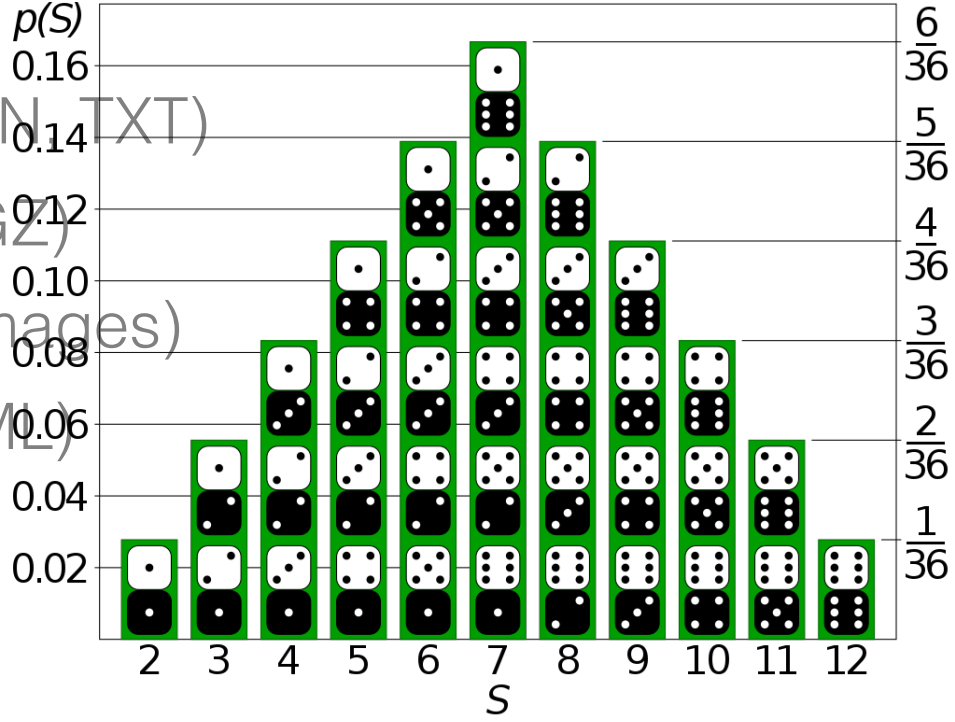
How is statistics used in healthcare?

Probability

Conditional Probability

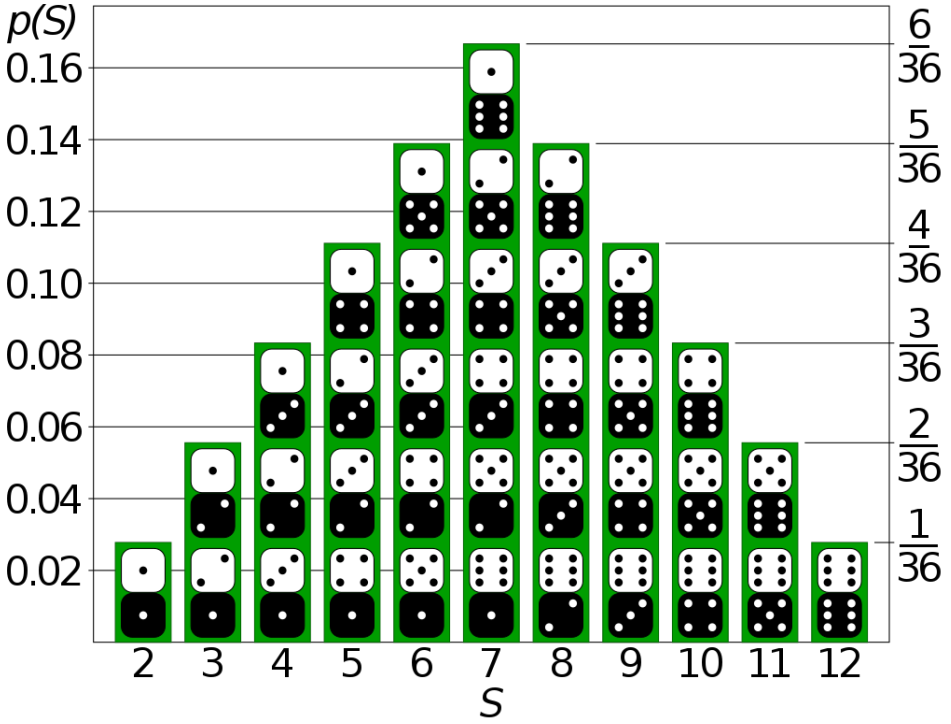
Probability Distribution

- Text files (CSV, TSV, JSON, TXT)
- Compressed files (ZIP, GZ)
- Binary files (XLS, PDF, Images)
- Web pages (links to HTML)
- Databases



Probability Distribution

PMF: Probability **Mass** Function (Discrete PDF)



PDF: Probability **Density** Function

Continuous Probability Distribution

Ethics and Accuracy

DeepMind (London)

Clinical records can predict Kidney failure

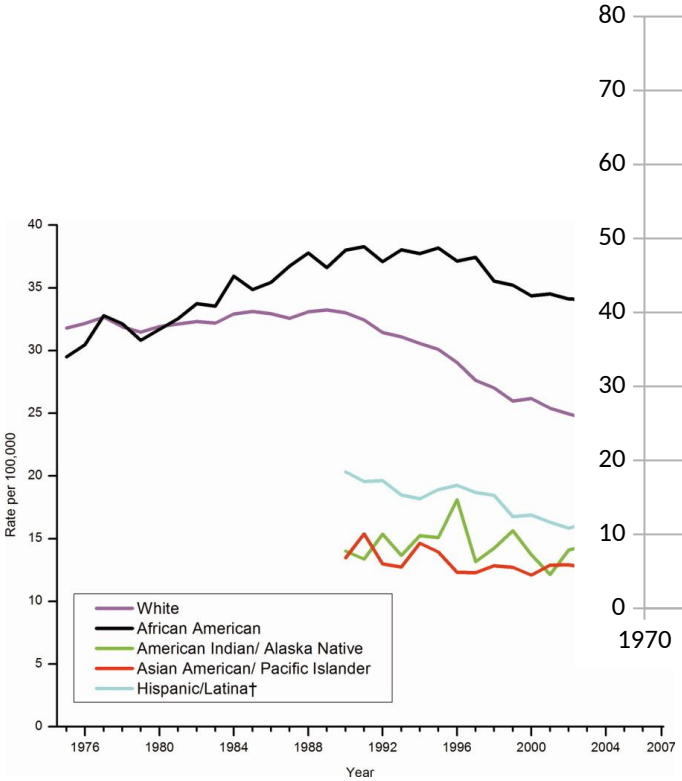
2 days in advance
55% accuracy for acute problems
90% accuracy for serious issues

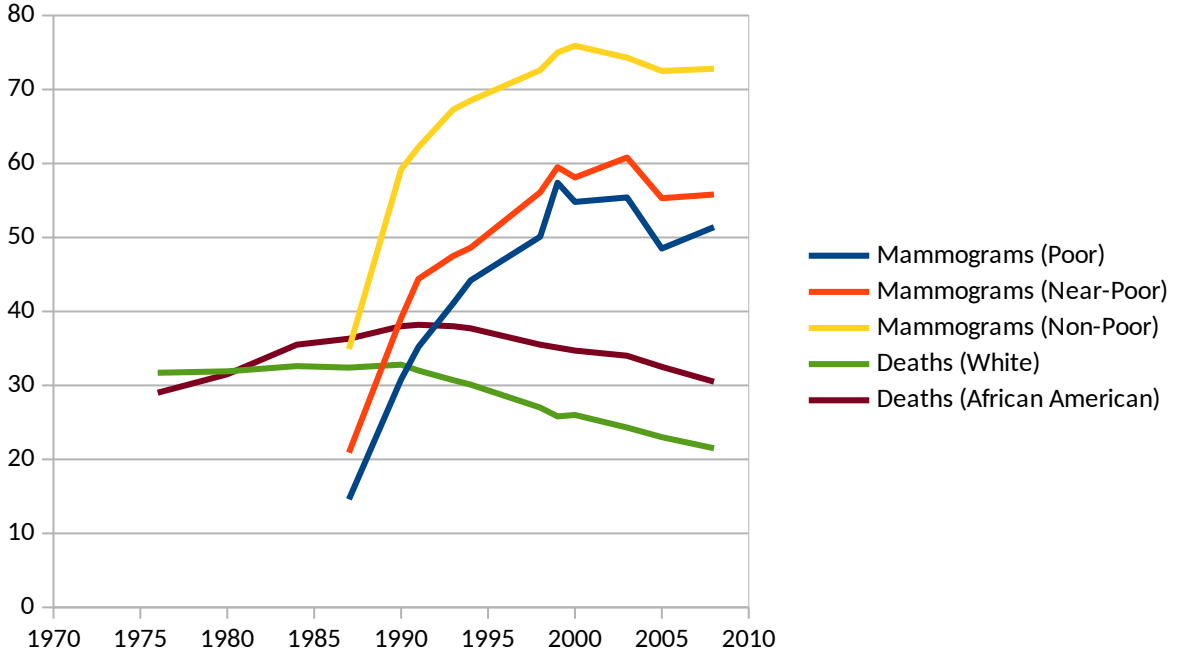
Dataset:
100% UK citizens
100% military
90% male

Berkson's Paradox

	General Population			Hospitalization past 6 mo		
	Bone Disease	No Bone Disease	% Bone Disease	Bone Disease	No Bone Disease	% Bone Disease
Lung disease	17	207	7.6%	5	15	25.0%
No lung disease	184	2,376	7.2%	18	219	7.6%

Correlation enables prediction

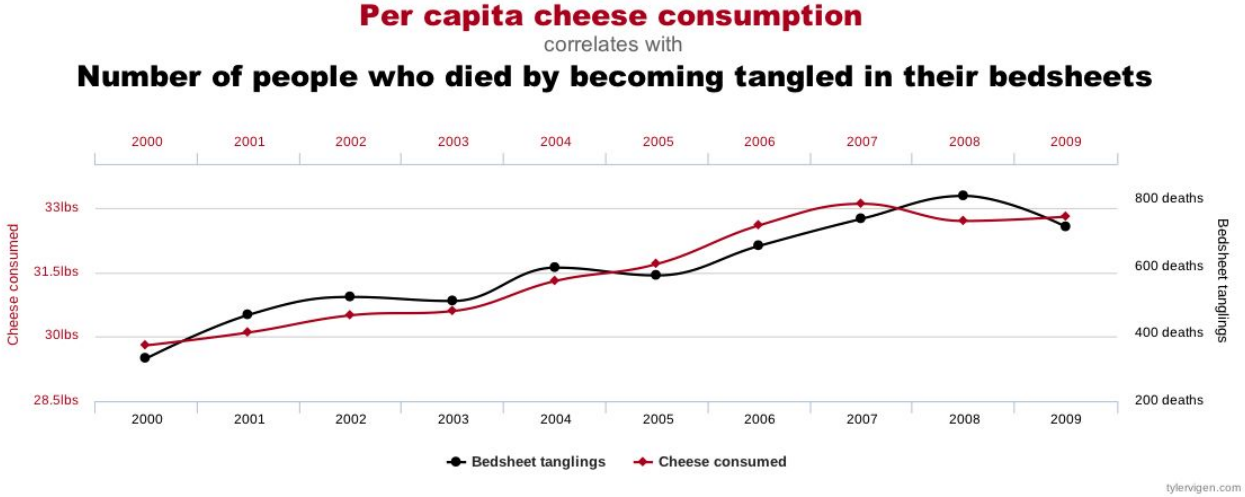




Breast Cancer Rates 2011: bit.ly/ucsdbreast

Correlation is not enough

- Computers are good at finding patterns
- But often those patterns are “spurious correlation”



Bayes Rule

Updated Probability = Likelihood Ratio X Prior Probability

$$P(D \vee T) = \frac{P(T \vee D)}{P(T)} \times P(D)$$

Bayes Rule Example

Prior	P(D)	Probability of getting breast cancer	1 in 700 per yr 1 in 70,000 (men)
Sensitivity	P(T|D)	Probability of mammogram detecting cancer	.73
False Positive Rate (False Alarm)	P(T|~D)	Probability of positive mammogram w/o cancer	.12
	P(T) = P(D) * P(T|D) + P(~D) * P(T|~D)	Probability of a positive mammogram among all women	.73 * 1 / 700 + .27 * 699 / 700 = .121

Mammograms can cause harm!

P(D)	1/700
P(T|D)	.73
P(T)	.121

ACP: biannually after age **50+**

previously: annual exams at 40+

$$P(D \vee T) = \frac{P(T \vee D)}{P(T)} \times P(D)$$

$$P(D \vee T) = \frac{.73}{.121} \times \frac{1}{700} = .0086 \approx 1\%$$

Assignments

Quiz

1. Why is understanding Baye's Rule so important?

Homework: Create diabetes MLE

1. Download diabetes dataset:
   http://totalgood.org/midata/...
2.

Project

1. Use numpy.random.randint() to simulated rolling a pair of dice.
2.