Digital Health

UCSD Extension – Specialization Certificate

Data Science for Healthcare

L3: Statistics

Hobson Lane, UC San Diego
Instructor

Agenda

- What is statistics and probability?
- Descriptive statistics
  - Probability distributions
  - Distribution parameters
- Prescriptive statistics
  - Accuracy
  - Bayes Rule

Statistics

- **Describe** a set of numbers in a dataset
- Describe **probabilities** outside the dataset
- Probabilities used to **infer** properties about the world
- Probabilities used to **predict** the future
- Probabilities used to **prescribe** actions in the world

Probability

- How often an event will happen among a set of trials
- Inherently binary (coin flip)

		heads	tails
			
Probability of **Heads**	$P(H)$		
Probability of **True result**	$P(T)$	True	False
Probability of **binary 1**	$P(1)$	1	0
Probability of **red pill**	$P(R)$		

Binary Probabilities are Complementary

		heads	tails
Probability of **Heads**	$P(H)=1-P(T)$		
Probability of **True**	$P(T)=1-P(F)$	True	False
Probability of **1**	$P(1)=1-P(0)$	1	0
Probability of **red pill**	$P(R)=1-P(B)$		

Binary = Mutually Exclusive

- P(Event) = 1 – P(All other events)
- Binary event: only one single alternative event

Probability of **Tails** $P(T)=1-P(H)$

Probability of **False** $P(0)=1-P(1)$ False

Probability of **Zero** $P(F)=1-P(T)$ 0

Probability of **Blue** $P(B)=1-P(R)$

Binary Dice?

- **Machines** “think” and **learn** in binary logic
  - Is it a 1?
  - Is it a 2?
  - Is it a 3?
  - Is it a 4?
  - Is it a 5?
  - Is it a 6 (six pips up) ?
- **Presence** or absence of a number
- 1 bit for each number or **category**
- Works even for non numerical (pictograph) dice or **loaded dice**

0 0 0 0 0 1

0 0 0 1 0 0   1 0 0 0 0 0

One-hot encoding
(used for categorical variables)

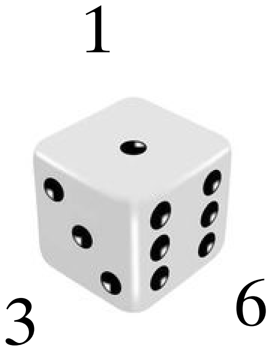
Ordinal Dice?

- But machines can work with numbers right?
- Ordinal encoding
- Sides assigned discrete numerical values
- Six mutually exclusive discrete values
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6

JUST SAY “**NO!**”

or “**YES!**”

but never “2, 3, 4, 5, 6”



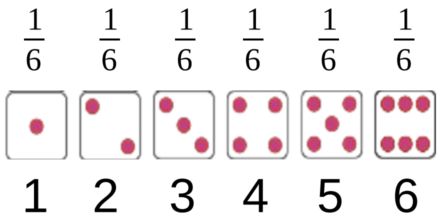
Ordinal values

Dice Probability

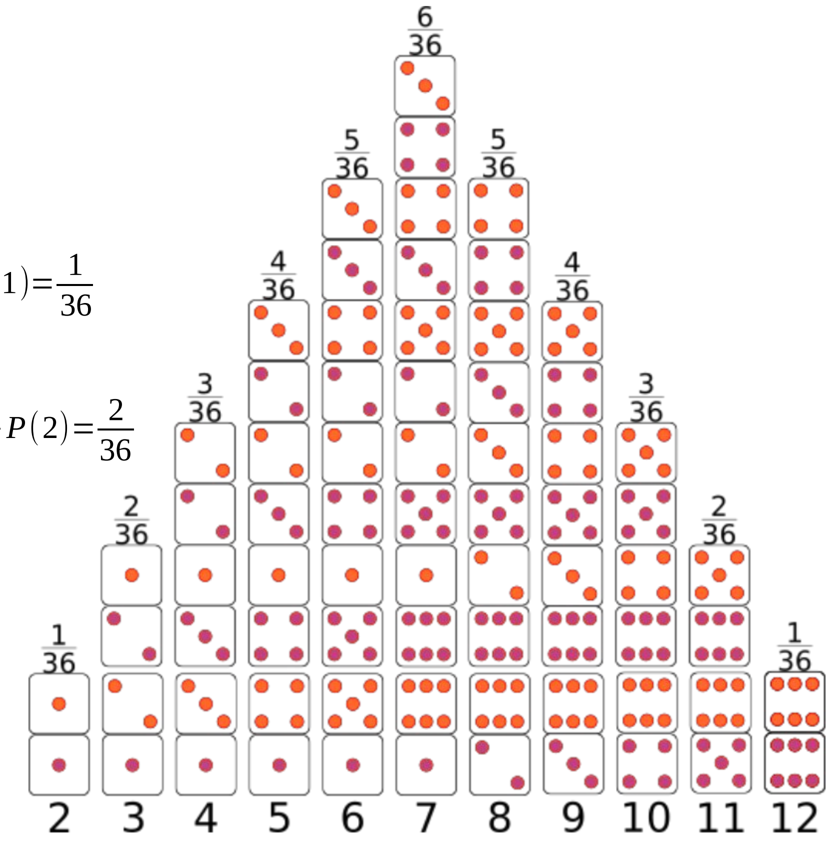
Die 1	Die 2	Sum	P
1	1	**2**	1/6
1	2	**3**	1/6
2	1	**3**	1/6
1	3	**4**	1/6

$$P(s=2)=P(1)\cdot P(1)=\frac{1}{36}$$

$$P(3)=2\cdot P(1)\cdot P(2)=\frac{2}{36}$$



One Die Roll



Sum of Two Die Rolls

Categorical Probability

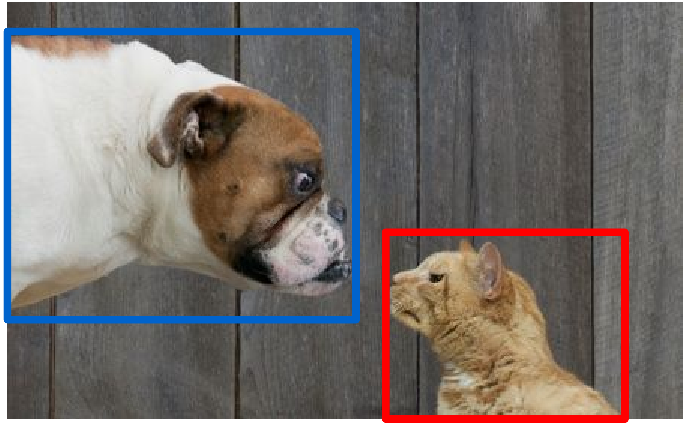
- Thought of as binary
- 1 bit for each category
- Often need “unknown” bit
- Probability = confidence
- 1 =

    P(Malignant)

    + P(Benign)

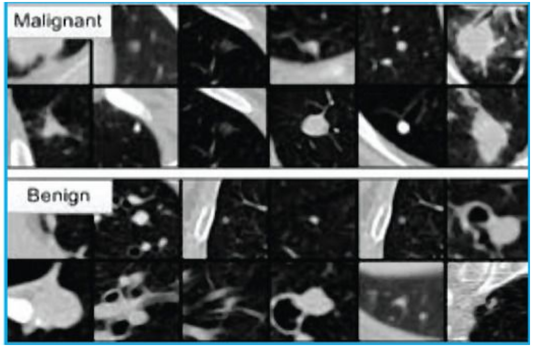
    + P(Healthy)

    + P(Unknown)



Dog? 0 or 1
Cat? 0 or 1
Fox? 0 or 1
Unk? 0 or 1

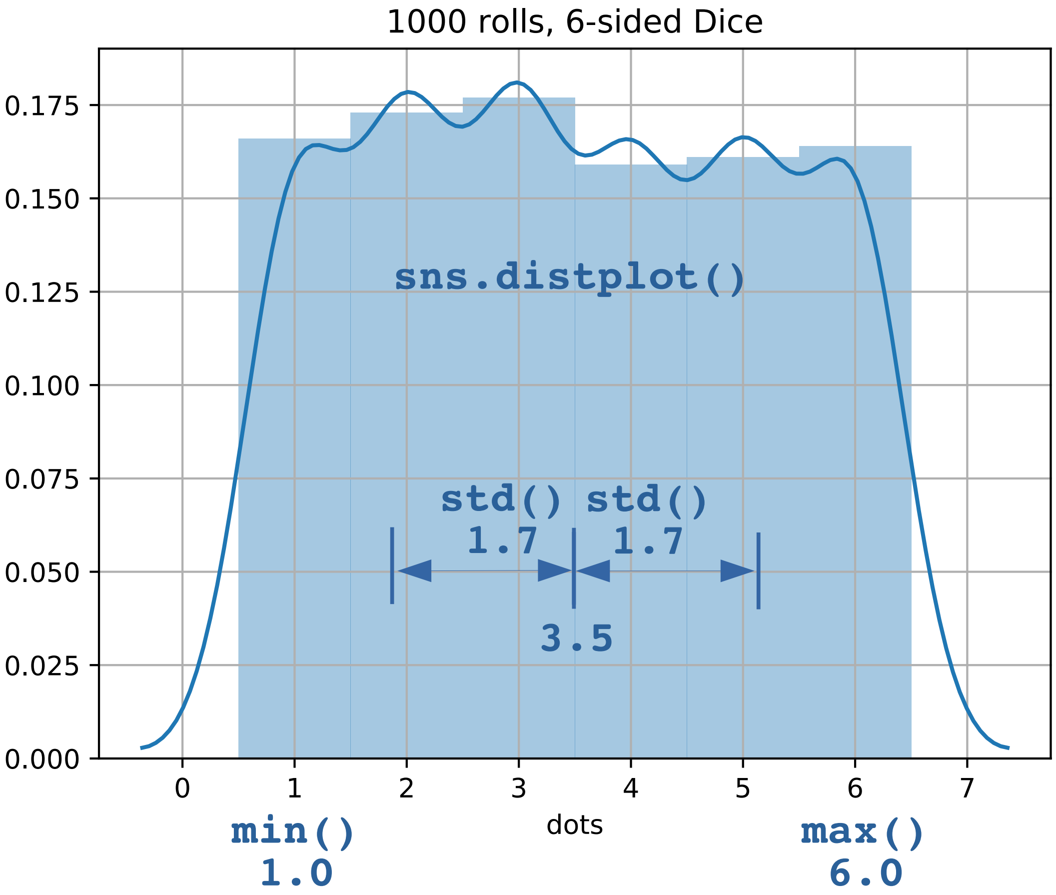
Dog? 0 or 1
Cat? 0 or 1
Fox? 0 or 1
Unk? 0 or 1



Malignant?
Benign?
Healthy?
Unknown? (none of the above)

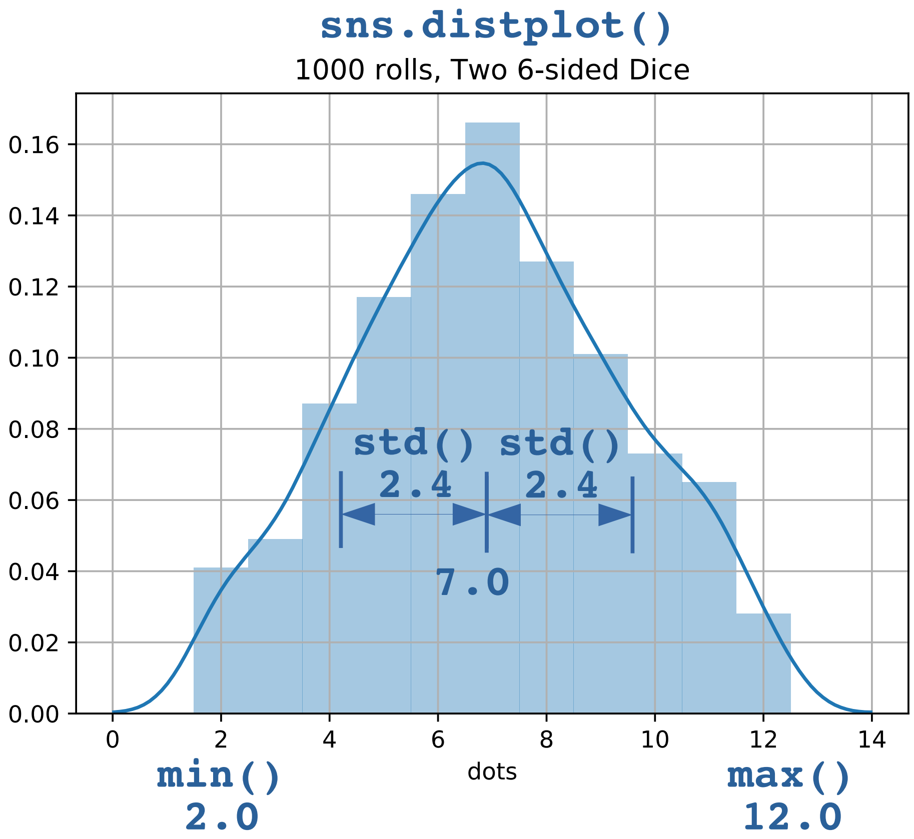
Die Statistics

- Min: 1.0
- Max: 6.0
- Mean: 3.47
- Median: 3.0
- Mode: 3.0
- Midpoint: 3.5
- Variance: 2.9
- Standard deviation: 1.7
- Histogram



2 Dice Statistics

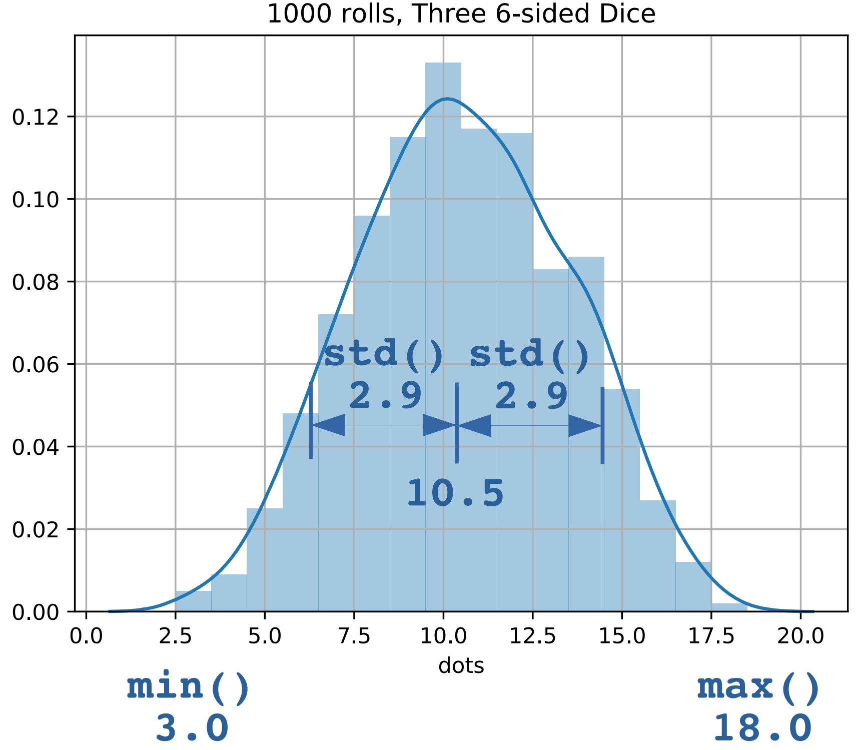
- Min: 2.0
- Max: 12.0
- Mean: 6.9
- Median: 7.0
- Mode: 7.0
- Midpoint: 7.0
- Variance: 6.1
- Standard deviation: 2.4
- Histogram



3 Dice Statistics

- Min: 2.0
- Max: 18
- Mean: 10.5
- Median: 10
- Mode: 10.0
- Midpoint: 10.5
- Variance: 8.6
- Standard deviation: 2.9
- Histogram

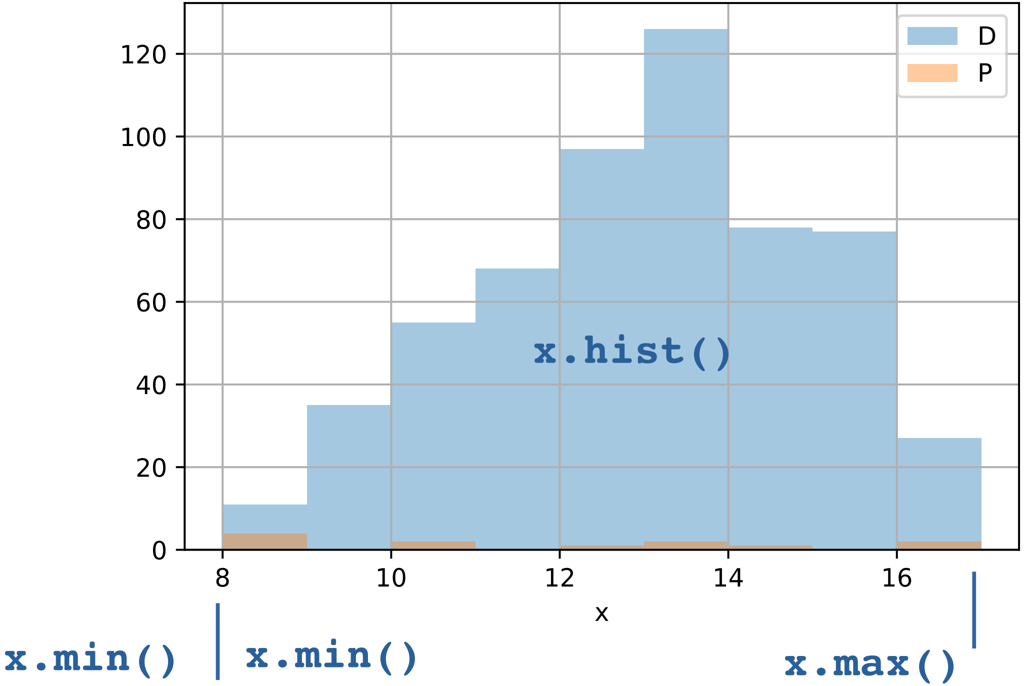
sns.distplot()



Mystery Dataset

	x
count	13.0
mean	12.5
std	3.4
min	8.7
25%	9.0
50%	12.6
75%	14.9
max	18.9

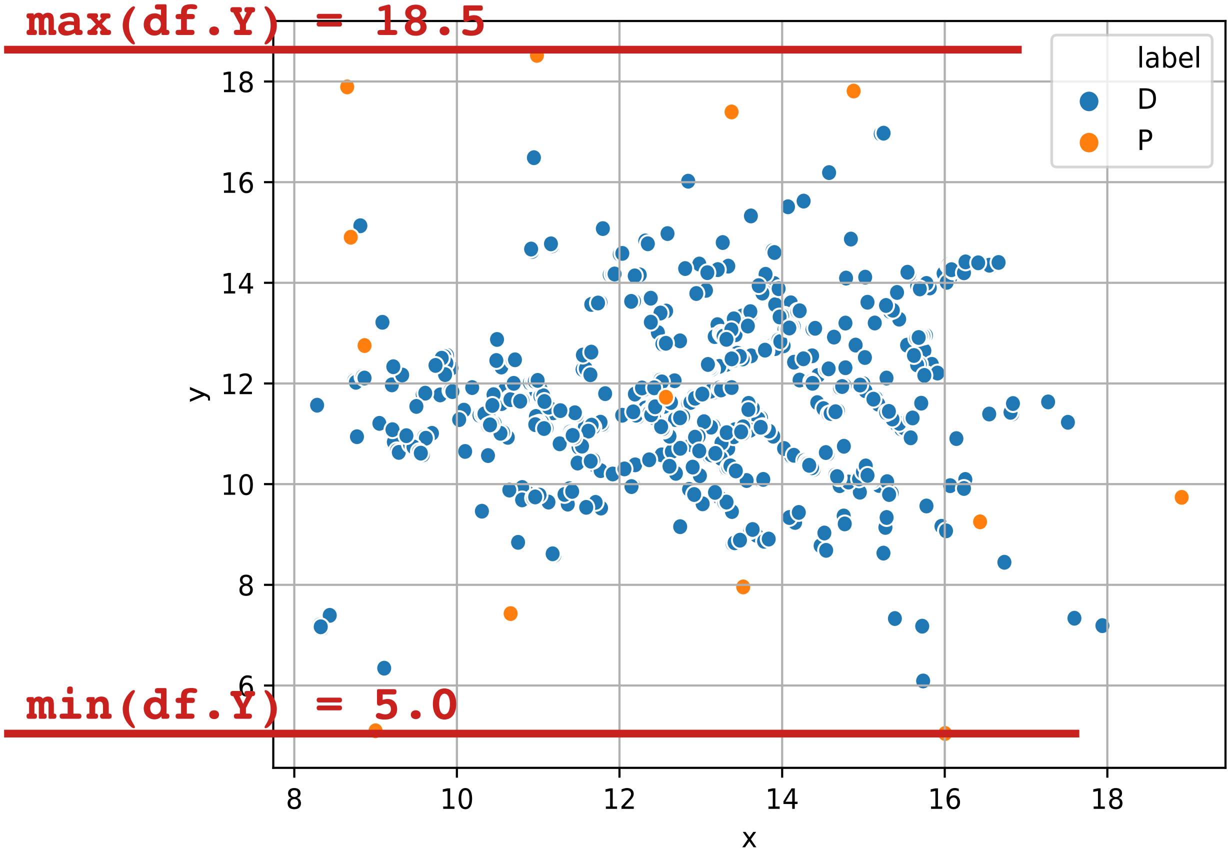
	x
count	578.0
mean	13.0
std	2.0
min	8.3
25%	11.6
50%	13.2
75%	14.5
max	17.9



df.describe()

	x	y
count	13.0	13.0
mean	12.5	12.0
std	3.4	5.0
min	8.7	5.0
25%	9.0	8.0
50%	12.6	11.7
75%	14.9	17.4
max	18.9	18.5

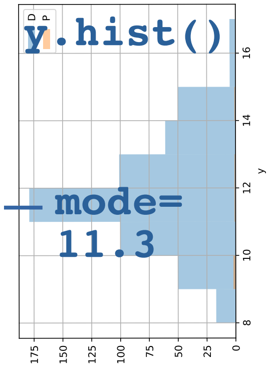
	x	y
count	578.0	578.0
mean	13.0	11.7
std	2.0	1.6
min	8.3	6.1
25%	11.6	10.7
50%	13.2	11.5
75%	14.5	12.8
max	17.9	17.0

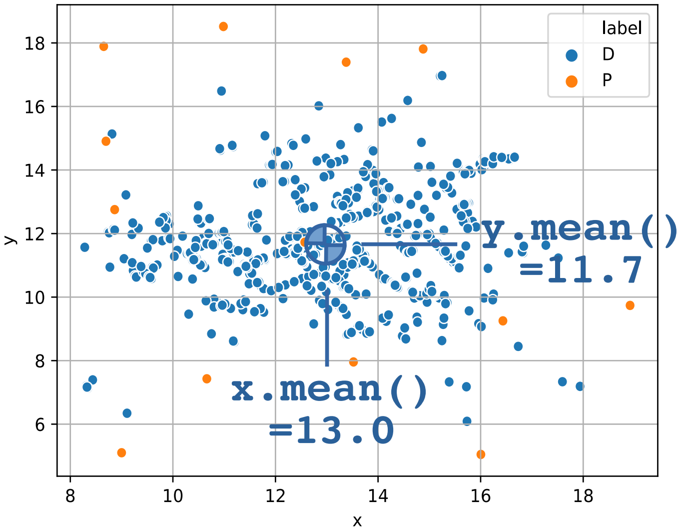


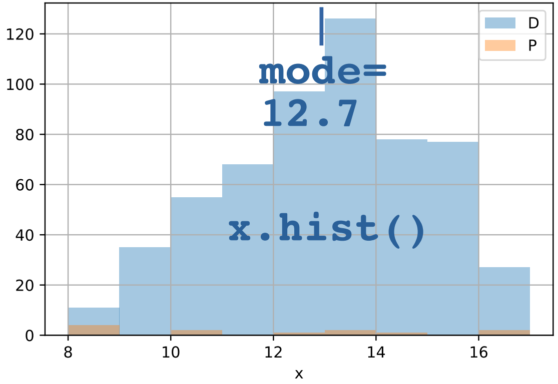
df.hist

	x	y
count	578.0	578.0
mean	13.0	11.7
std	2.0	1.6
min	8.3	6.1
25%	11.6	10.7
50%	13.2	11.5
75%	14.5	12.8
max	17.9	17.0

mode()
(12.7, 11.3)
median()
(13.2, 11.5)

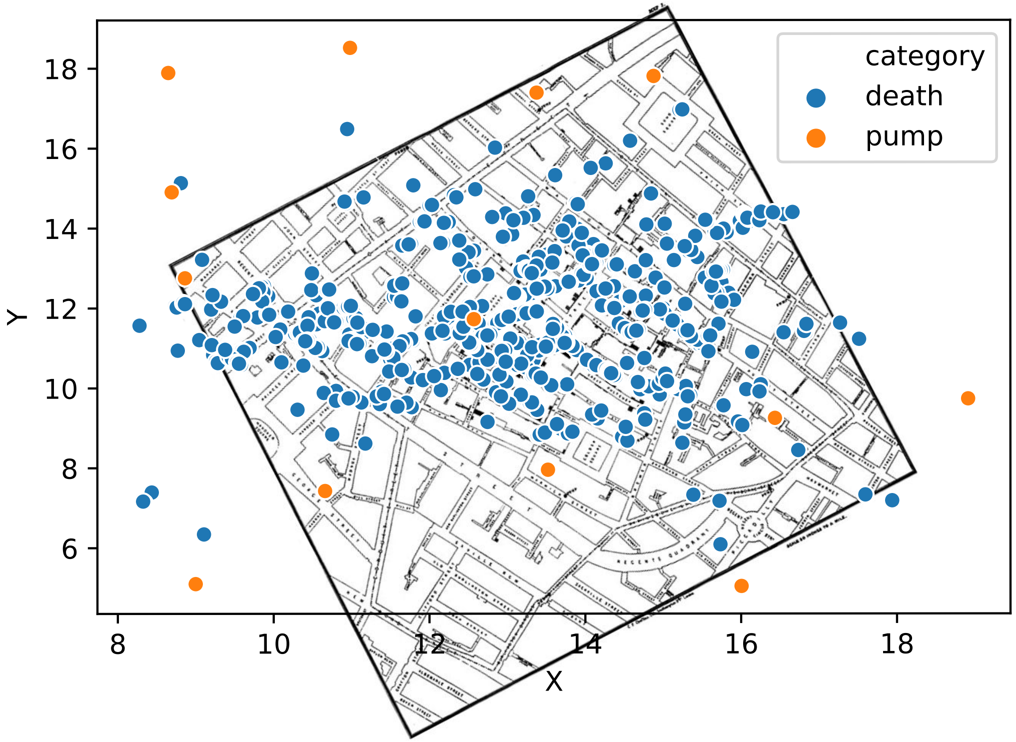






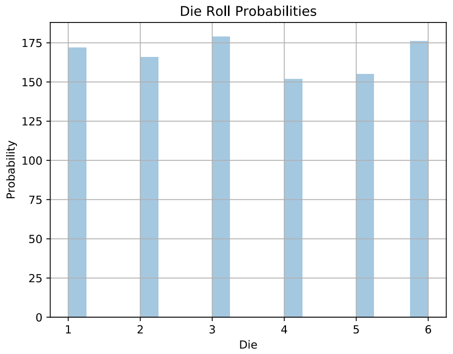
Coordinate Frame (Representation) Matters!

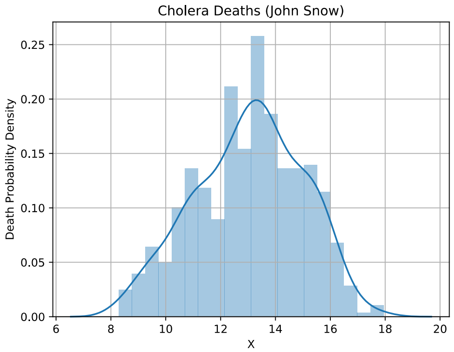
- Coordinate frame
  - Origin (offset)
  - Orientation
- Scale (units)
- Filtering
  - Clipping
  - Outliers
  - Smoothing/imputing

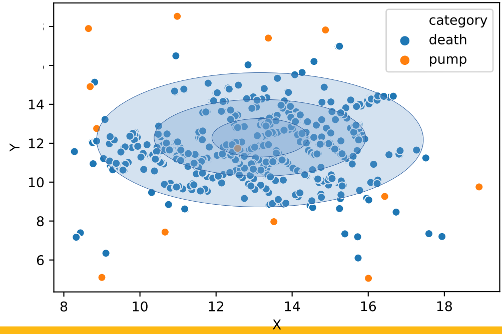


Kinds of Distributions

- Discrete probability:
  - Dice, gender, disease, death, recovery
  - Distribution (histogram)
- Continuous probability:
  - Height, weight, (x, y) position
  - Probability density (kernel density)
- Conditional probability:
  - Height based on weight (continuous condition)
  - Death based on distance (discrete & continuous condition)
  - 2D probability density

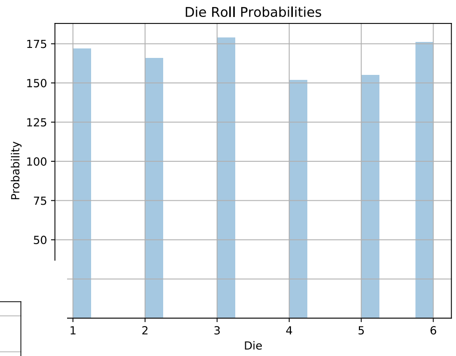


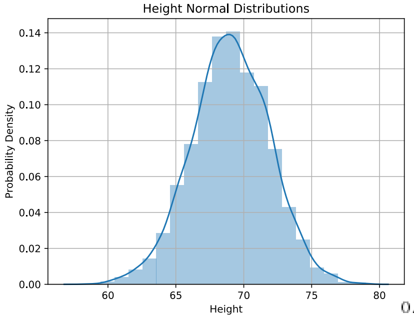


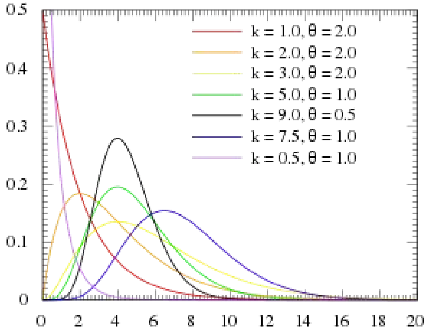


Common Distributions

- Uniform
  - Dice, coin flips, cards
- Normal (Gaussian)
  - Height, weight
  - Probability density (kernel density)
- Gamma (generalized Poisson)
  - Time between infections, epidemics
  - Asymmetric distribution



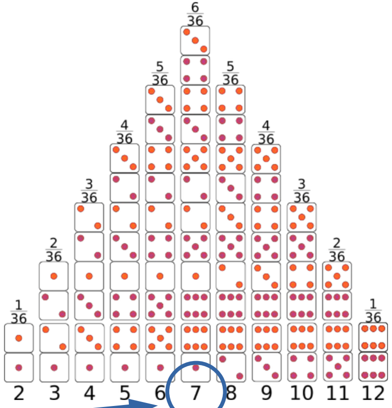




Arithmetic **Mean** (Average)

Tyco Brahe, 1587

mean $\mu = \frac{\sum_{i=0}^{N} x_i}{N}$

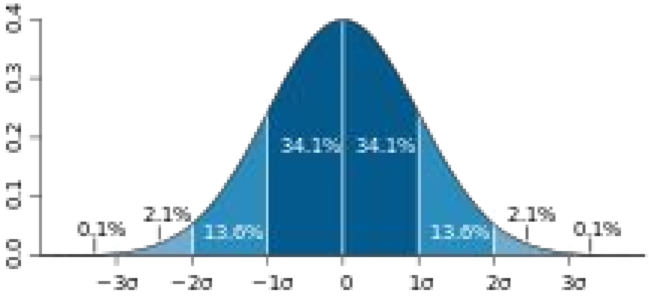


$$\mu = \frac{2+3+3+4+\ldots+11+11+12}{36} = 7$$

```
# x = [2, 3, 3, 4, 4, 4, 5, 5, 5, 5, … 9, 9, 9, 9, 10, 10, 10, 11, 11, 12]
# x = [die1 + die2 for (die1, die2) in product(range(1, 7), range(1, 7))]
# mu = sum(x) / len(x)

x = [sum(dice) for dice in product([1, 2, 3, 4, 5, 6], [1, 2, 3, 4, 5, 6])])
mu = np.mean(x)  # 7
```

Standard Deviation (and Variance)

Karl Pearson, 1894



variance: $$\sigma^2 = \sum_{i=0}^{N} \frac{(x_i - \mu)^2}{N}$$

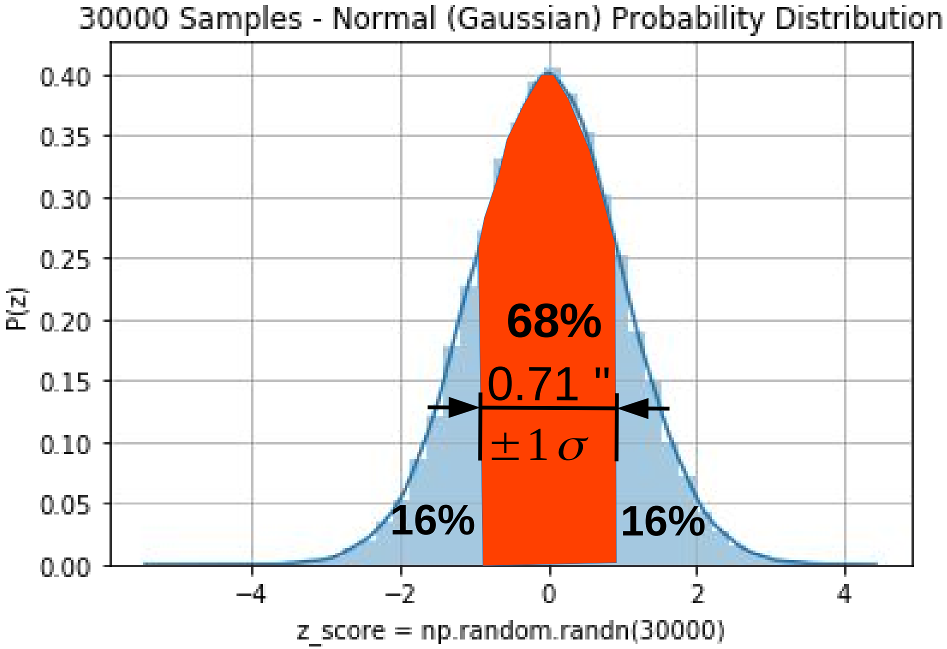
standard deviation: $$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=0}^{N} \frac{(x_i - \mu)^2}{N}}$$

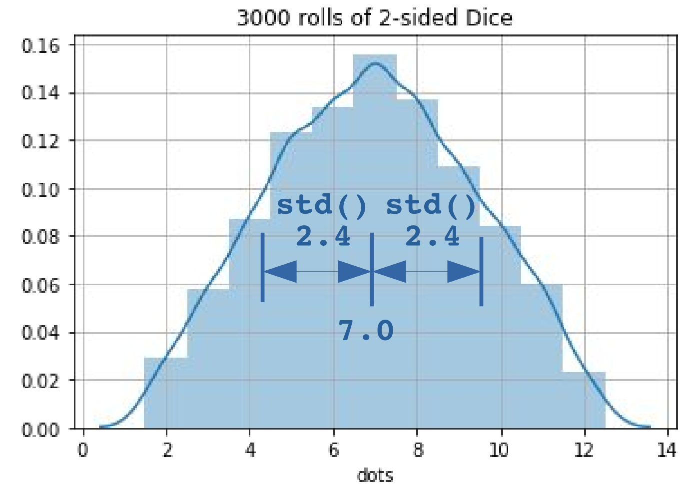
```
# sigma = np.sqrt(sum(x**2) / len(x))

sigma = np.std(x)  # 2.415...
```

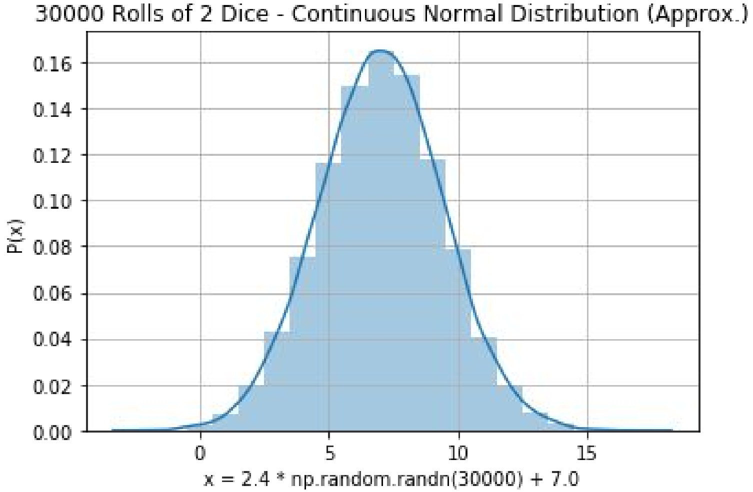
Normal (Gaussian) Distribution

68-95-99.7 rule

$1\sigma = 68\%$ $2\sigma = 95\%$ $3\sigma = 99.7\%$







Accuracy

- Continuous numerical predictions (regression):
  - How close to correct are your predictions?
- RMSE: Root Mean Square Error

$$RMSE(y,z)=\sqrt{\frac{\sum_{i=0}^{N}(y_i-z_i)^2}{N}}$$

Categorical Accuracy

- Categorical predictions (classification):
  - What proportion of your predictions were correct?

$$\frac{N_{correct}}{N_{total}} = \frac{N_{correct}}{N_{correct} + N_{incorrect}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Bayes Rule

Updated Probability = Likelihood Ratio * Prior Probability

$$P(D|T) = \frac{P(T|D)}{P(T)} \cdot P(D)$$

Likelihood Ratio

Bayes Rule Example

Prior	P(D)	Probability of getting breast cancer	1 in 700 per yr 1 in 70,000 (men)
True Positive Rate (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
Positive Rate	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

Real Numbers

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

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

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

What is statistics?

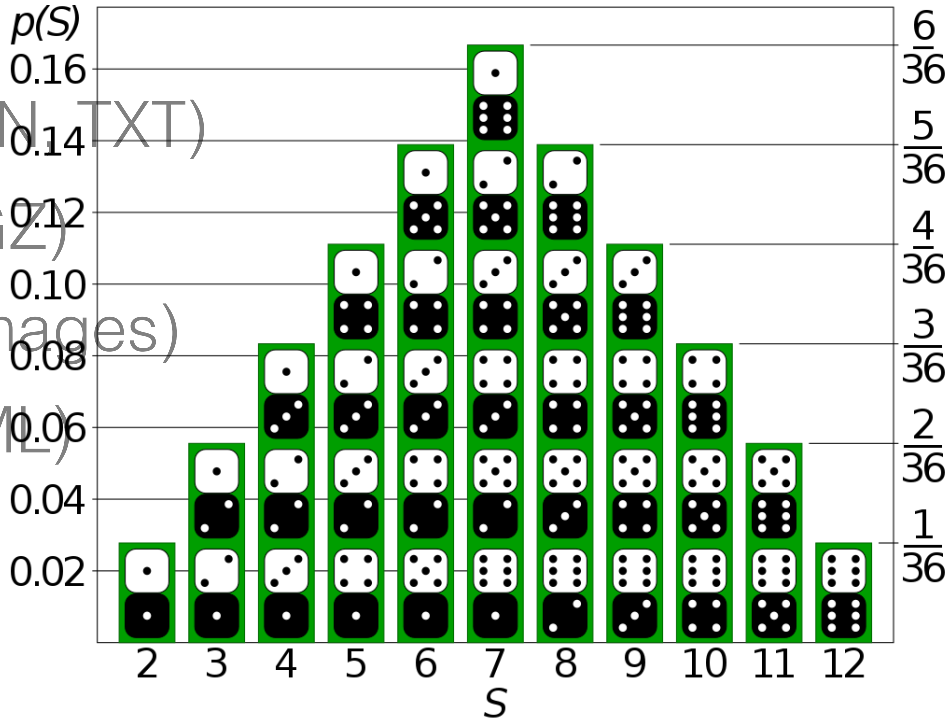
How is statistics used in healthcare?

Probability

Conditional Probability

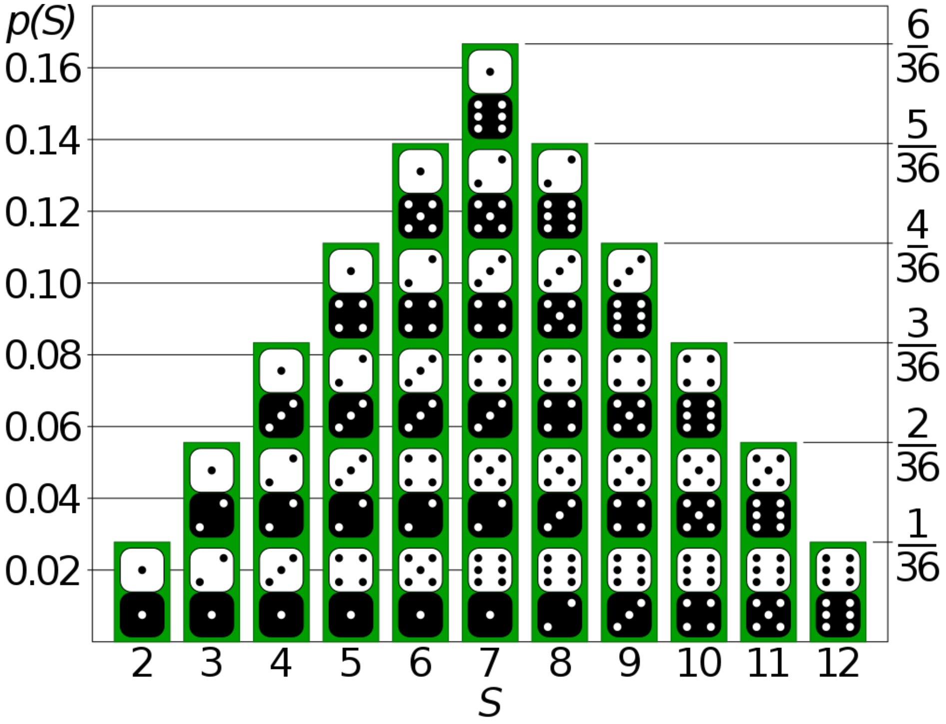
Probability Distribution

- 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

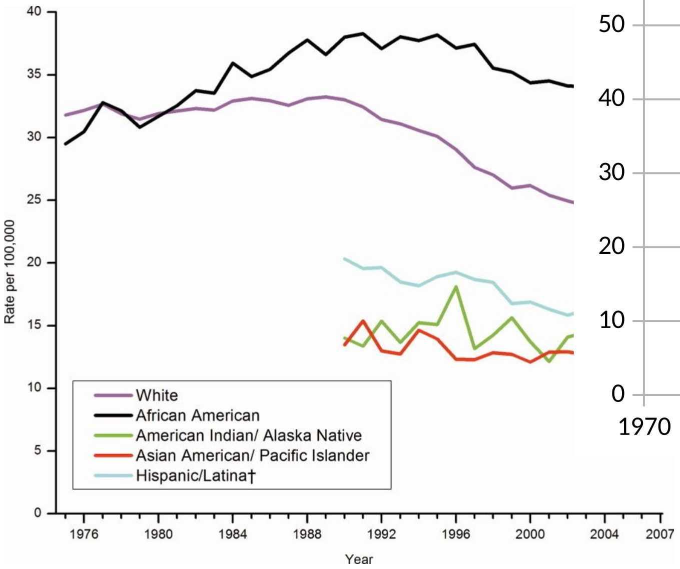
2 days in advance
55% accuracy for acute problems
90% accuracy for serious issues

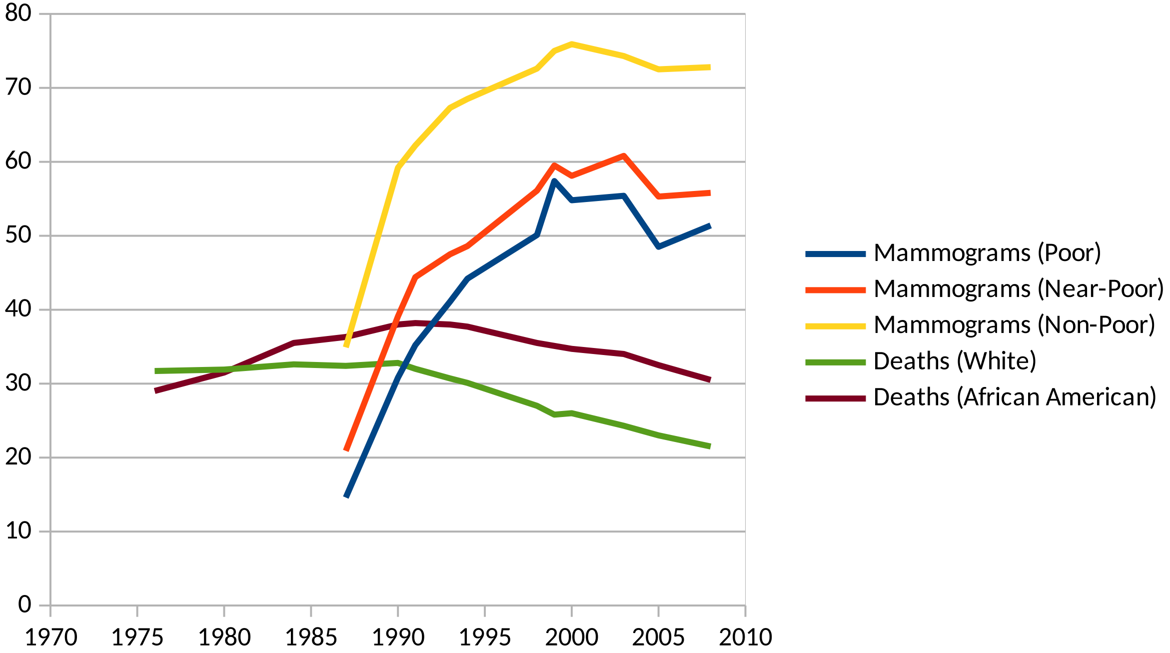
Dataset:
100% UK citizens
100% military
90% male

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

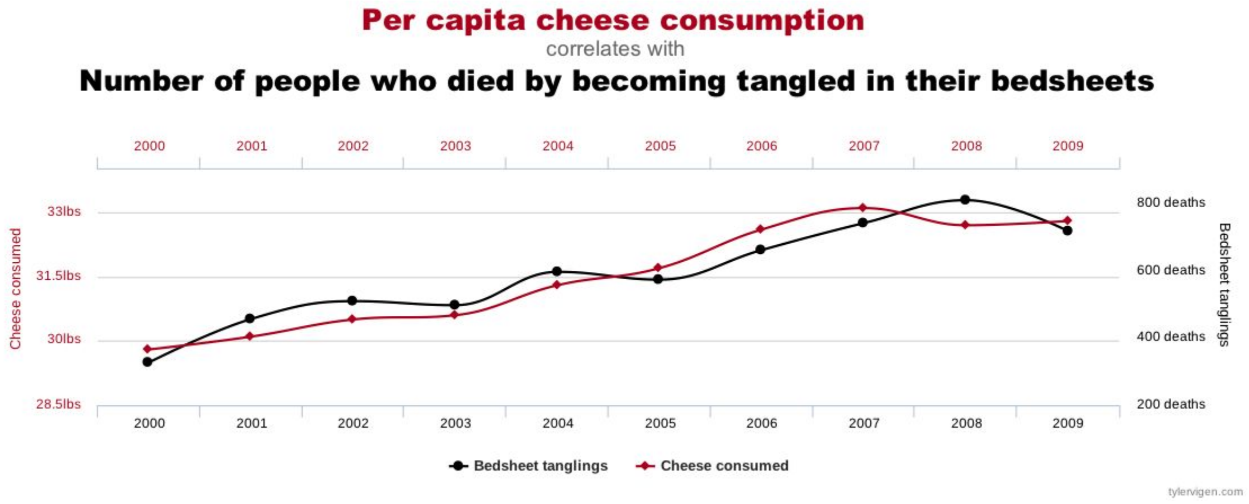




Breast Cancer Rates 2011: bit.ly/ucsdbreast

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

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

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.