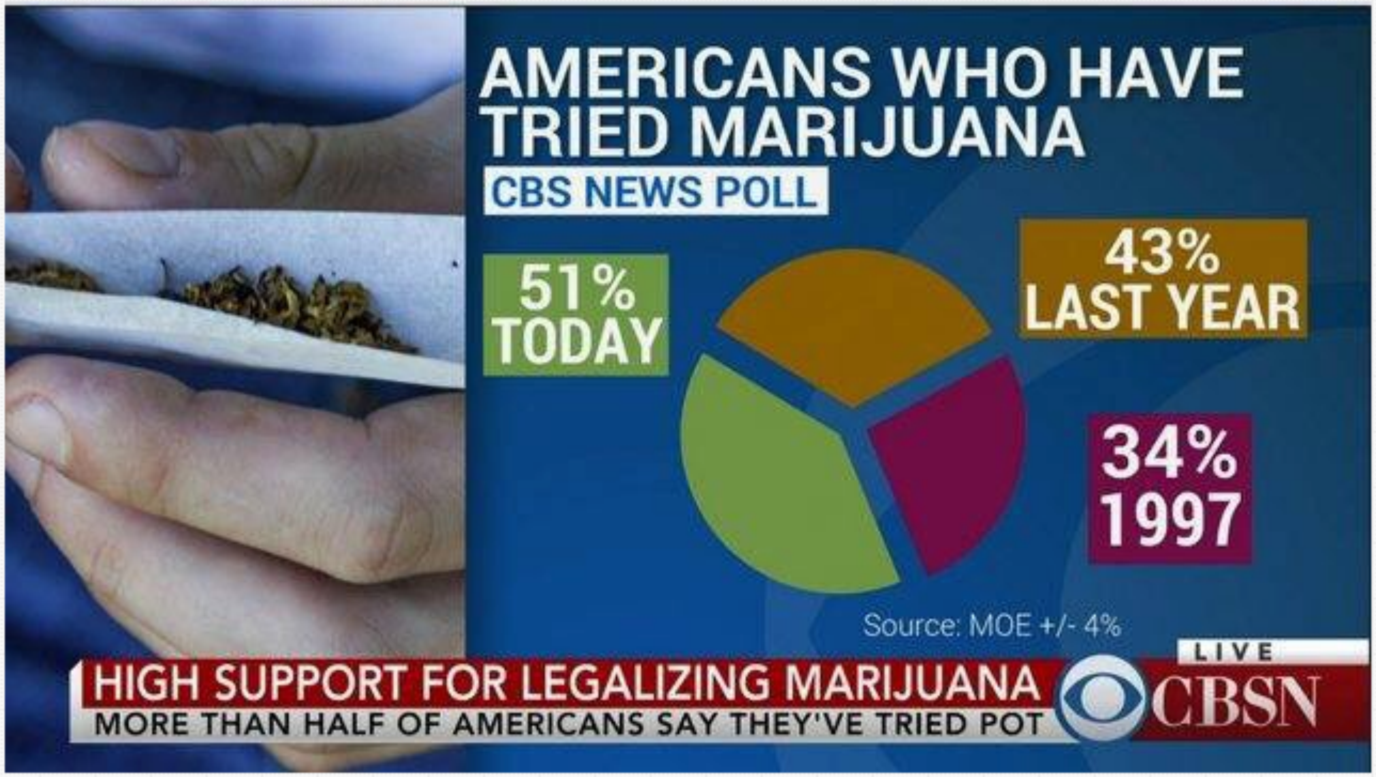
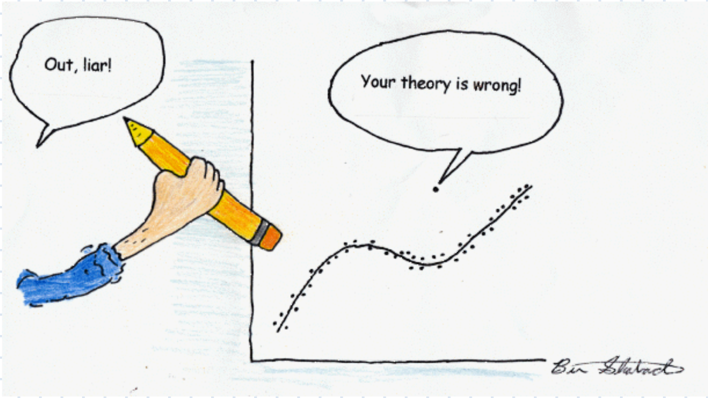
Basically, cleaning your data & then a little probability but first…bad data visualization

Getting High on Bad Data Visualization



Data Cleaning

- Always necessary with a new dataset
- Assume your data has errors
- Tables & summary statistics & graphs can help
- Outliers – extreme values
- Always keep a copy of the original data



Data Cleaning

- Data cleaning tasks
  - Examine your data
  - Fill in missing values (many ways)
  - Identify outliers and smooth out noisy data*
  - Correct inconsistent data
    - Noisy
    - Redundant
    - Biased

*2005 Leff, Allen et al, Inside Employment Interventions: Effects of Job Development and Job Support on Competitive Employment of Persons with Severe Mental Illness, Psychiatric Services, 56(10): 1237-1245.

Outliers – what do we do?

- Is the value physically possible?
- Yes, keep them but…
- Do sensitivity analyses (what’s that?)
- Which summary measures are they likely to affect?
- Trimmed means, robust statistics, transformations

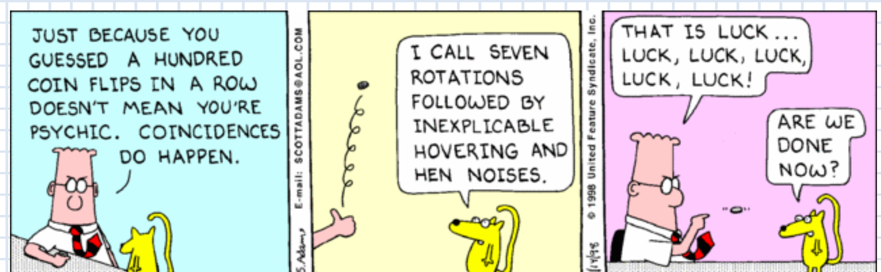
Probability! Playing Rock Paper Scissors http://www.essentially.net/rsp/

It’s all about uncertainty!

- Probability is the numerical measure of the likelihood that an event will occur
- Value is between 0 and 1
- Sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1

Defining an Event

- The result of an experiment or observation
- The event occurs or does not occur
- We all apply probability to events and we want to estimate the relative frequency that an event occurs over a large number of identical trials
- An event can be a disease occurrence or an abnormal lab value



Probability of an event

- Probability of an event is the relative frequency of its occurrence in a large number of trials repeated under the same conditions
  - Always lies between 0 and 1 (inclusive)
  - Denoted P(A) or P(X) etc…
  - Examples?
    - Efficacy trials?
    - Vaccine trials?

Probability: some summary definitions

- Complement of the probability of event A [P(A)] is = 1-P(A)
- Mutually exclusive events:

  For 2 events P(A) and P(B)

  If A and B are *mutually exclusive* then they can never happen together

  P(A) and P(B) happening together = 0

Conditional Probability or Bayes Theorem

- The probability that an event B will occur given that event A has occurred
  - The probability of B given A = P(B|A)
  - The probability that it rains given you brought your umbrella
  - P(Rain|Have umbrella)
- P(B|A) = P(A and B)/P(A)
  - The relative size of the probability of the the events happening together compared to the relative size of P(A)

Conditional Probability Example

- Very often used in diagnostic tests
- P(HIV|given a positive test) OR P(Positive test|HIV)
  - In diagnostic testing these are called positive predictive value and sensitivity
- What are each of these telling us about a test?
- P(not HIV|negative test) OR P(Negative test|not HIV)
  - Negative predictive value and specificity

Conditional Probability Example:

The dangers of screening everyone

- P(Disease) = 0.001 or 100 of 100,000
- P(No Disease) = 1 – P(Disease) = 0.999
- P(Positive+|Disease) = 0.99 or 99 or 100
- P(Negative-|No Disease) = 0.99 or 98,901 of 99,900

	Disease	No Disease	
Test Positive	99	999	1098
Test Negative	1	98901	98902
	100	99900	100000

P(Disease) =	100/100000 =	0.001	
P(No Disease) =	99900/100000 =	0.999	
P(Test+|Disease) =	99/100 =	0.99	
P(Test-|No Disease) =	98901/99900 =	0.99	
P(No Disease|Test+) =	999/1098 =	0.91	

Probability example using a table

2014 State of Michigan Live Births

- Probability that a mother's age is 20-24 = .237
- Probability that a mothers age was <=24 =
  - .001 + .061 +.237 = .299
  - Mutually exclusive!

- *Conditional probability*: Given that a mother's age is at least 30, what is the probability that she at least 40?

- P(Mother's age >=30 ***and*** Mother's age >=40) / P(Mother's age >=30)

  = P(Mother's age >=40)/P(Mother's age >=30) ***(why?)***
  = (.024) / ((30165 + 12475 + 2757)/114460)
  = **0.061**

Maternal age group	Totals	% of total
<15	69	0.1%
15 - 19	6,968	6.1%
20 - 24	27,134	23.7%
25 - 29	34,884	30.4%
30 - 34	30,165	26.4%
35 - 39	12,475	10.9%
40+	2,757	2.4%
Total	114,460	100%

Examples of conditional probabilities

- Sensitivity & Specificity
- Relative risk

  P(disease | exposure) / P(disease | no exposure)
- Odds ratios

  (P(exposure | disease) / (1 – P(exposure | disease)) / (P(exposure | no disease) / (1-P(exposure | no disease))

…we'll see these in later videos and use them in models as well as separate calculations

Independence

- If an event, B, does not depend on the occurrence of event A, B and A are independent. But they are not necessarily mutually exclusive.
- The definition: $P(B|A)=P(B)$
- The P(A) and P(B) together = $P(A)*P(B)$
- Examples:
  - Probability of becoming infected with malaria given that you have 2 sisters = Probability of becoming infected with malaria
  - Probability of a (fair) coin toss landing on heads given that the previous coin toss was a heads
  - Probability you are a banker and having green eyes = P(Banker)*P(GreenEyes)

Using Probability: Diagnostic Tests

- True positives – the test is positive given the person has the disease
  - The probability of this is $P(T^+|D^+)$ = Sensitivity
- False positives – the test is positive although the person does not have the disease
- True negatives – the test is negative given the person does not have the disease
  - The probability of this is $P(T^-|D^-)$ = Specificity
- False negatives – the test is negative even though the person has the disease

Applying probability to diagnostic test results

- Suppose you have a panel of diagnostic tests that give false positive results 2% of the time. If you test a disease-free patient, what is the probability of the correct negative result? This is the _______________
- Suppose you test a disease-free patient twice with this test and assuming the tests are independent. What are the probabilities:
- Negative Negative
- Negative Positive
- Positive Negative
- Positive Positive
- What is the P(1 or more tests are positive)?

Applying probability to diagnostic test results

- Suppose you test a disease-free patient twice with this test. What are the probabilities:
- Negative Negative = 0.98 * 0.98 = 0.9604
- Negative Positive = 0.98 * 0.02 = 0.0196
- Positive Negative = 0.02 * 0.98 = 0.0196
- Positive Positive = 0.02 * 0.02 = 0.0004
- What is the P(1 or more tests are positive)?

  Which of the above tests have at least 1 positive?

  = 0.0196 + 0.0196 + 0.0004 = 0.0396

Applying probability to diagnostic test results

- What is the P(1 or more tests are positive)?
- This can also be written as 1 – (Neg Neg)

  0.0396 = 1 – 0.9604

- Then the probability of at least 1 false positive if 5 tests are run = $1 - (0.98)^5$ = 0.096

- Suppose the false positive probability was = 0.05 then = $1 - (0.95)^5$ = 0.226

Back to Bayes' theorem

- Suppose there is a super rapid HIV antibody test
  - Sensitivity = $P(T+|HIV+) = 0.96$
  - Specificity = $P(T-|HIV-) = 0.99$

- You want to know the probability that someone with a positive test is truly HIV+
  - $P(HIV+|T+)$
  - Positive Predictive Value (PPV)
  - You also know the $P(HIV+) = 0.02$ in this region

Bayes' theorem

$P(HIV+|T^+) = P(T^+|HIV+)*P(HIV+) / P(T^+)$

- $P(T^+|HIV+) = 0.96$ (sensitivity)
- $P(HIV+)$ in region of study is = 0.02
- $P(T^+)$ = the overall chances of having a positive test

$P(T^+) = P(T+|HIV+)\ P(HIV+) + P(T+|HIV-)\ P(HIV-)$

$= 0.96*0.02 + 0.01*0.98 = 0.029$

So this gives us:

$P(HIV+|T^+) = 0.96*0.02/0.029 = 0.662$

How this helps you: prior to posterior probability

- The prevalence of HIV+ was assumed to be = 2%
- Prior to testing the probability that a randomly selected person is infected with HIV = 0.02
  - This is the Prior probabiity
- The probability that someone who tests positive is actually infected with HIV = 0.662
  - This is called the Posterior probability incorporating information gained from doing the test.

Next: Merging statistical summaries with probability = Random Variables