Data Mining & Machine Learning

CS37300
Purdue University

August 30, 2017

Announcements

- Due to labor day 9/4, my office hours will be Tue 9/5 10-11am

- Updated slides of Lectures 2 and 3

Probability and statistics (cont)

Common distributions

- Bernoulli
- Binomial
- Multinomial
- Poisson
- Normal

Bernoulli

- Binary variable (0/1) that takes the value of 1 with probability p
    - E.g., Outcome of a fair coin toss is Bernoulli with *p=0.5*

$$P(x) = p^x(1-p)^{1-x}$$

$$E[X] = 1(p) + 0(1-p) = p$$

$$\begin{aligned} Var(X) &= E[X]^2 - (E[X])^2 \\ &= 1^2(p) + 0^2(1-p) - p^2 \\ &= p(1-p) \end{aligned}$$

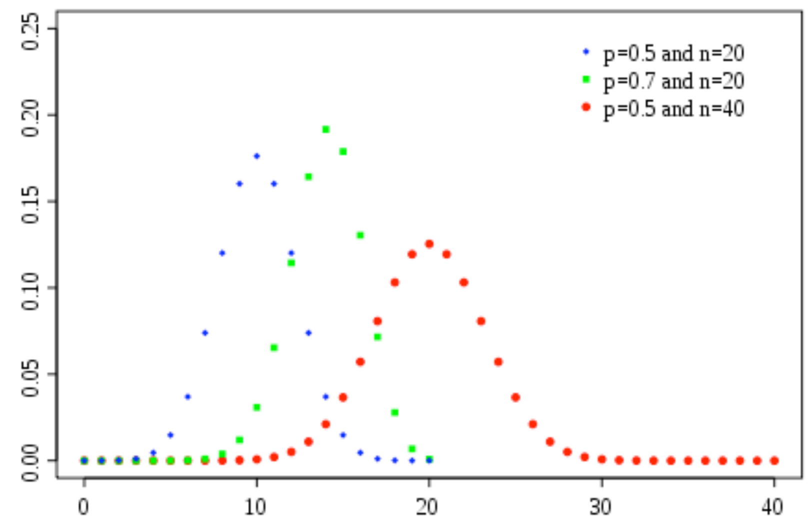
Binomial

- Describes the number of successful outcomes in *n* independent Bernoulli(p) trials
  - E.g., Number of heads in a sequence of 10 tosses of a fair coin is Binomial with *n=10* and *p=0.5*

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$E[X] = np$$

$$Var[X] = np(1-p)$$



Multinomial

- Generalization of binomial to *k* possible outcomes; outcome *i* has probability p_i of occurring
    - E.g., Number of {outs, singles, doubles, triples, homeruns} in a sequence of 10 times at bat is Multinomial
- Let X_i denote the number of times the *i*-th outcome occurs in *n* trials:

$$P(x_1, \ldots x_k) = \binom{n}{x_1, \ldots x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

$$E[X_i] = np_i$$

$$Var(X_i) = np_i(1 - p_i)$$

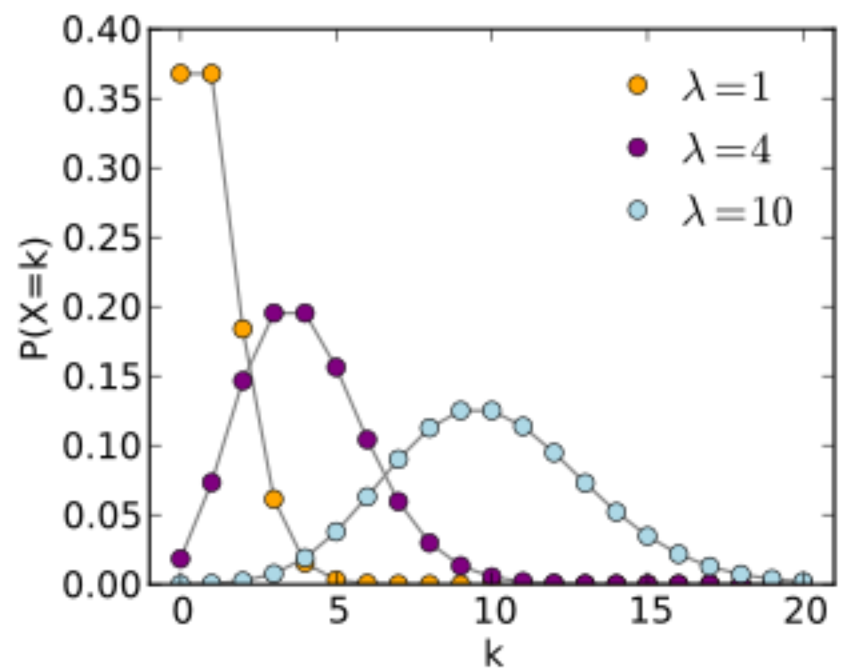
Poisson

- Describes the number of successful outcomes occurring in a fixed interval of time (or space) if the “successes” occur *independently* with a known average rate
  - E.g., Number of emergency calls to a service center per hour, when the average rate per hour is $\lambda=10$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$E[X] = \lambda$$

$$Var[X] = \lambda$$



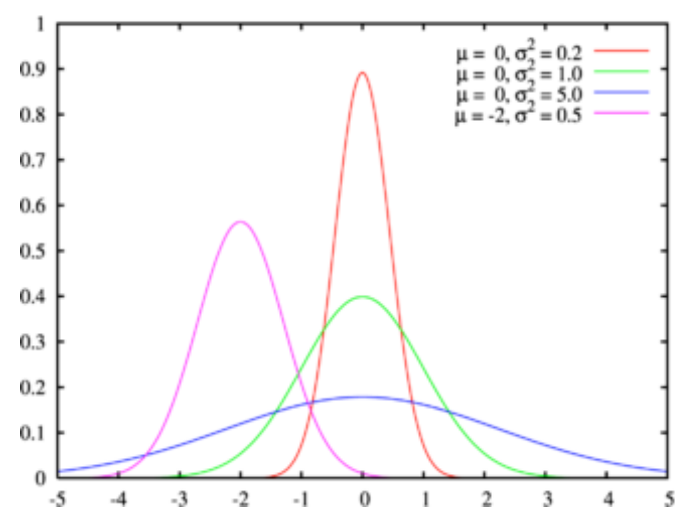
Normal (Gaussian)

- Important distribution gives well-known bell shape
- Central limit theorem:
    - Distribution of the mean of n samples becomes normally distributed as n ↑, regardless of the distribution of the underlying population

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$



Multivariate RV

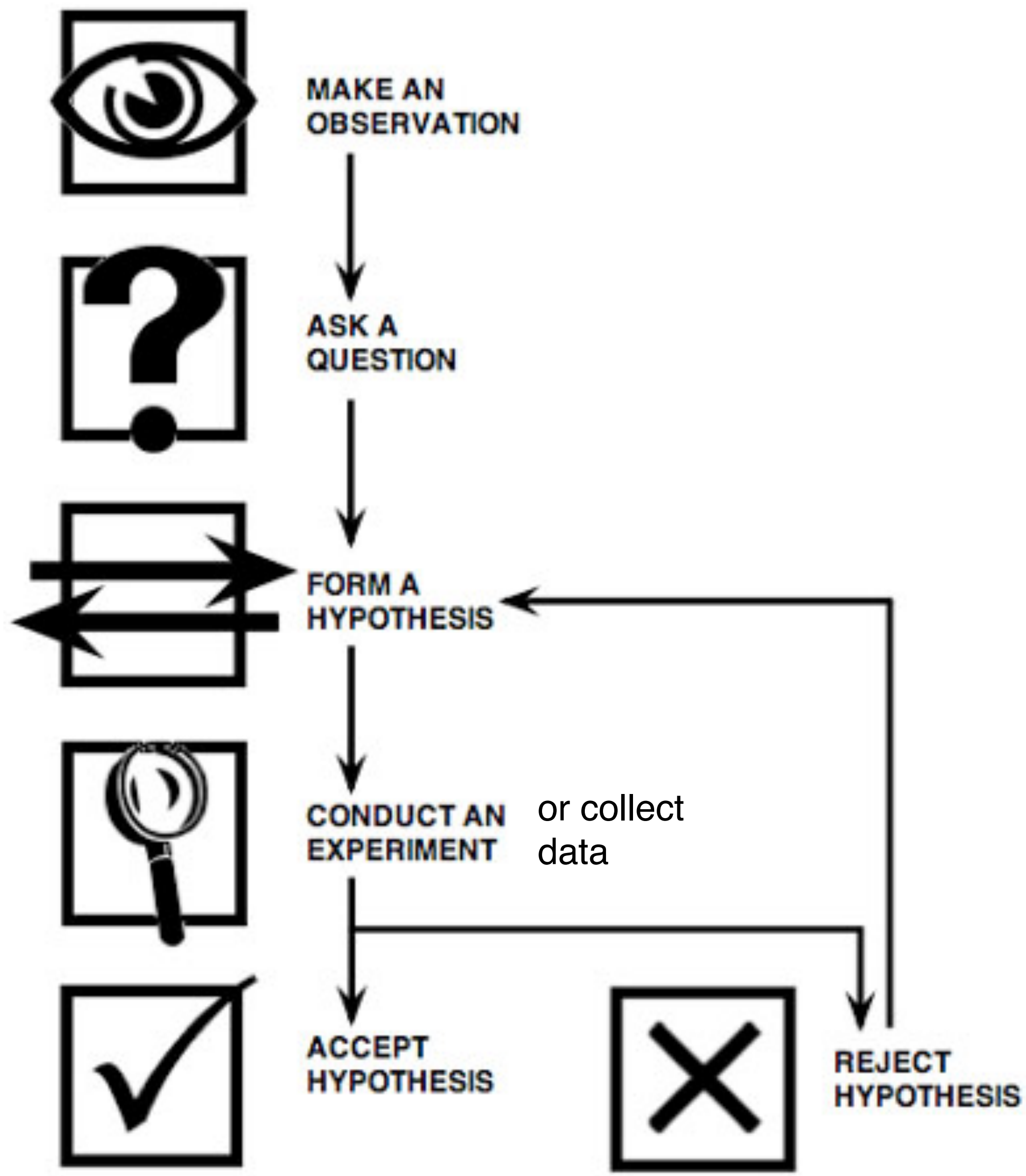
- A multivariate random variable ***X*** is a set X_1, X_2, ..., X_p of random variables
- **Joint** density function: $P(\mathbf{x})=P(x_1, x_2, ..., x_p)$
- **Marginal** density function: the density of any subset of the complete set of variables, e.g.,:

$$P(x_1) = \sum_{x_2, x_3} p(x_1, x_2, x_3)$$

- **Conditional** density function: the density of a subset conditioned on particular values of the others, e.g.,:

$$P(x_1 | x_2, x_3) = \frac{p(x_1, x_2, x_3)}{p(x_2, x_3)}$$

Primer on hypotheses


MAKE AN OBSERVATION
ASK A QUESTION
FORM A HYPOTHESIS
CONDUCT AN EXPERIMENT
or collect data
ACCEPT HYPOTHESIS
REJECT HYPOTHESIS

What is a hypothesis?

- **Hypotheses** are tentative statements of the expected relationships between two or more variables
  - **Inductive** hypotheses are formed through inductively reasoning from many specific observations to tentative explanations (*bottom-up*)
  - **Deductive** hypotheses are formed through deductively reasoning implications of theory (*top-down*)

- Reasons for using hypotheses
  - Provides focus and directs research investigation
  - Allows the investigator to confirm or not confirm relationships
  - Provides a useful framework for organizing and summarizing results and conclusions

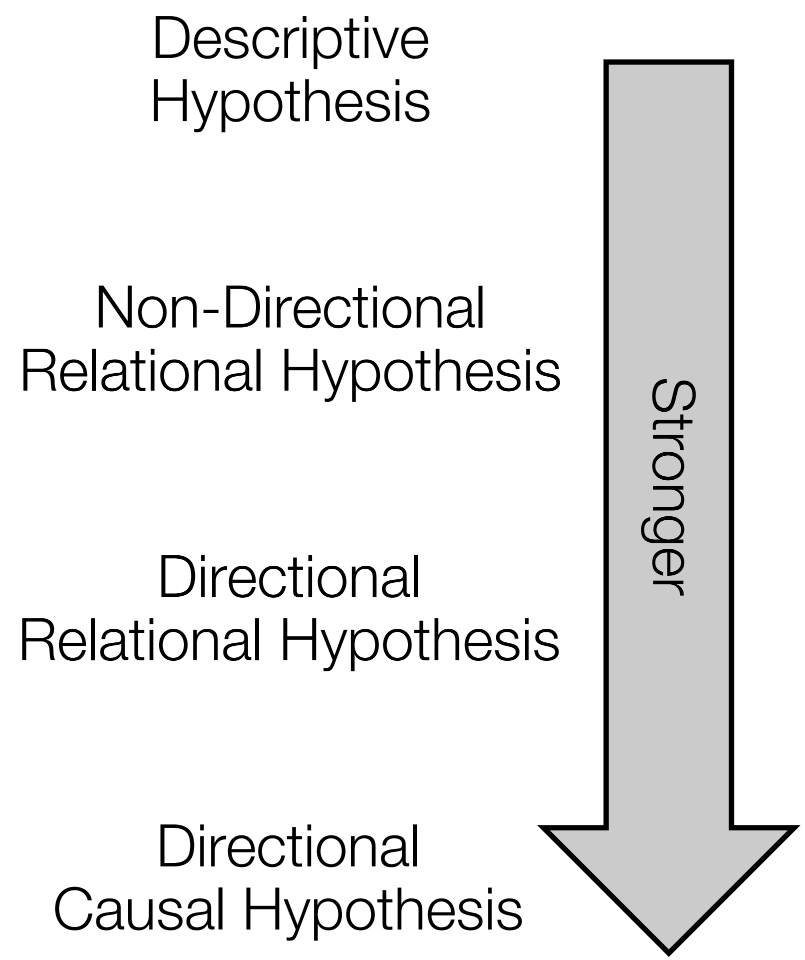
Types of hypotheses

Broad categories

- **Descriptive**: propositions that describe a characteristic of an object
- **Relational**: propositions that describe the relationship between 2+ variables
- **Causal**: propositions that describe the effect of one variable on another

Specific characteristics

- **Non-directional**: an differential outcome is anticipated but the specific nature of it is not known (e.g., yelling at your boss will change your salary)
- **Directional**: a specific outcome is anticipated (e.g., CS graduates have the highest average starting salary of all Purdue graduates)



From claims to testable hypotheses

- Over the years, Democrats (DEM) have argued that average donation to their candidates are smaller than that of Republican (GOP) candidates.
  Data: GOP 2012 donations https://goo.gl/e61m9t
  DEM 2012 donations https://goo.gl/By3qRc
  Rows: donations; columns: candidate, amount(USD), state

- **Step 1**: Express data as random variables (joinly). E.g.:

$$(X, Y) \equiv \text{ (political party of candidate, donation value to candidate)}$$

- **Step 2**: Restate claim as a hypothesis about the relationship between the random variables, e.g.,

  - Hypothesis: $\mathrm{E}[Y \mid X = \mathrm{DEM}] < \mathrm{E}[Y \mid X = \mathrm{GOP}]$

- **Step 3**: Determine type of hypothesis (and consider whether you can make it stronger), e.g., for X $\in$ {GOP,DEM}

  - Directional-relational: X=DEM is associated with smaller Y

From claims to testable hypotheses

- Over the years Democrats (DEM) have argued that average individual donations to their candidates are smaller than that of Republican (GOP) candidates.
- **Types of hypotheses**:
  - *Descriptive*: Donations values vary (i.e., Y varies).
  - *Non-directional relational*: Y varies based on party affiliation (i.e., X and Y are associated)
  - *Directional-relational*: Democrats get smaller donations (i.e., X=DEM is associated with smaller Y)
  - *Causal-relational*: Democrats get smaller donations because they have stronger candidates in poor districts (i.e., X=DEM is associated with smaller Y, but if you control for average income in district, this effect may disappear)

Using Data to Test Hypotheses

IMPORTANT: Random variable definition is tailored to task

- Data as a table: Rows: donations; columns: candidate, amount(USD), state
- Example:

  …
  GOP_donations_2012.csv data (only GOP candidates):
  D= […,[John McCain, 2500, AZ],[John McCain, 500, AZ],…]

  Data ARE samples of our random variables:
  - What is a (X,Y) sample?
    - (GOP, Amount)
    - We are ignoring the candidate (John McCain)
    - What is the GOP average donation?

$$E[Y|X = x] = \sum_{y=1}^{\infty} yP[Y = y|X = x]$$

Probability computed as % of donations
with value y in the file GOP_donations_2012.csv ?

IMPORTANT: Random variable definition is tailored to task

- Our random original variables:

$$(X, Y) \equiv \text{(political party of candidate, donation value to candidate)}$$

- Over the years, Democrats (DEM) have argued that **average donation to the DEM candidates** are smaller than that of Republican (GOP) candidates.

Means the average
PER CANDIDATE or
just the GOP average donation?

- Some people may disagree with the definition, arguing that the claim is an average PER CANDIDATE

We need to expand the random variables

Expanded Random Variables

- Over the years, Democrats (DEM) have argued that **average donation** to their candidates are smaller than that of Republican (GOP) candidates.

- Redefining random variables to get the PER CANDIDATE average:

$$(Z, X, Y) \equiv \text{ (candidate, party of candidate, donation value to candidate)}$$

  - Gives a sample: (John McCain, GOP, 500)
  - Per candidate average

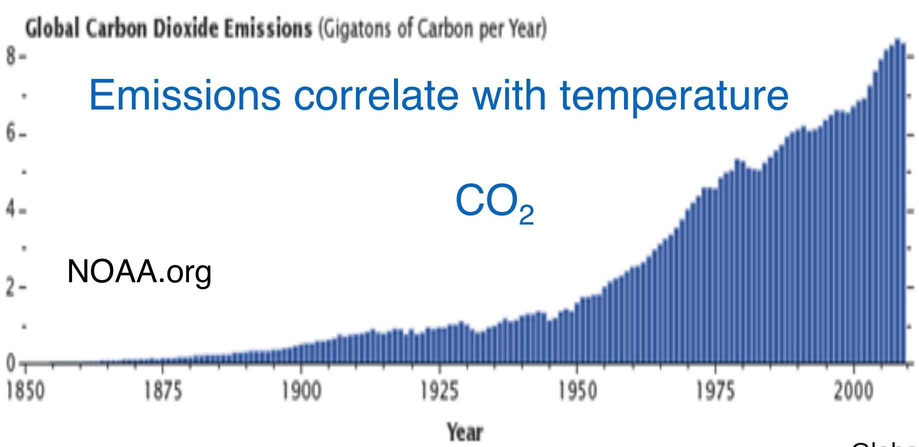
$$\text{Average} = \sum_{z \in \text{candidates}} \frac{E[Y|X = x, Z = z]}{(\text{no. candidates at party } x)} = \sum_{y=1}^{\infty} \sum_{\forall z \in \text{candidates}} y \frac{P[Y = y|X = x, Z = z]}{(\text{no. candidates at party } x)}$$

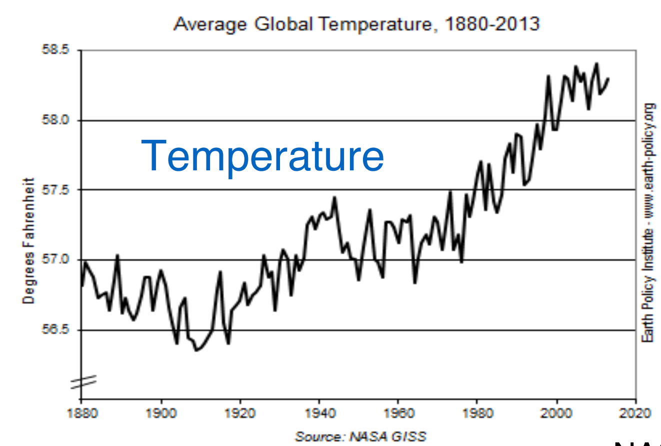
Probability computed as % of donations with value y of candidate z in the file GOP_donations_2012.csv ?

Quiz Answers

- CLAIM 1: The temperature of the planet is rising and the increase is due to human activities such as fossil fuel use and deforestation.

- Which kind of data could support such claim?

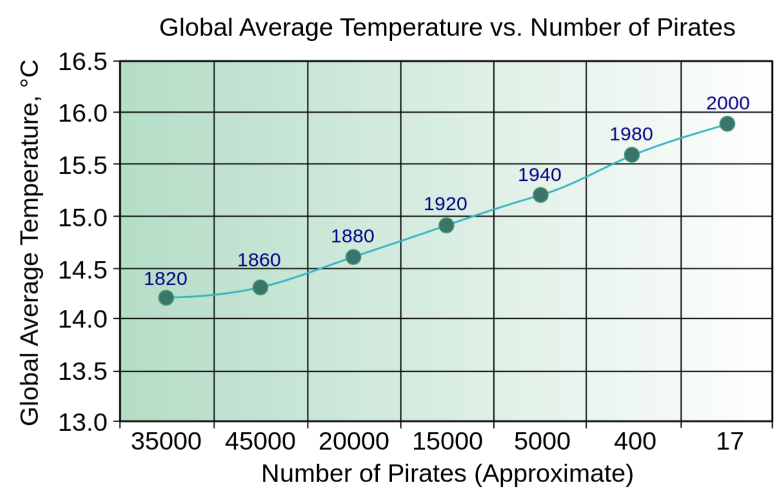




NASA

Not enough

Why?



Predictive Models Offer Stronger Hypotheses

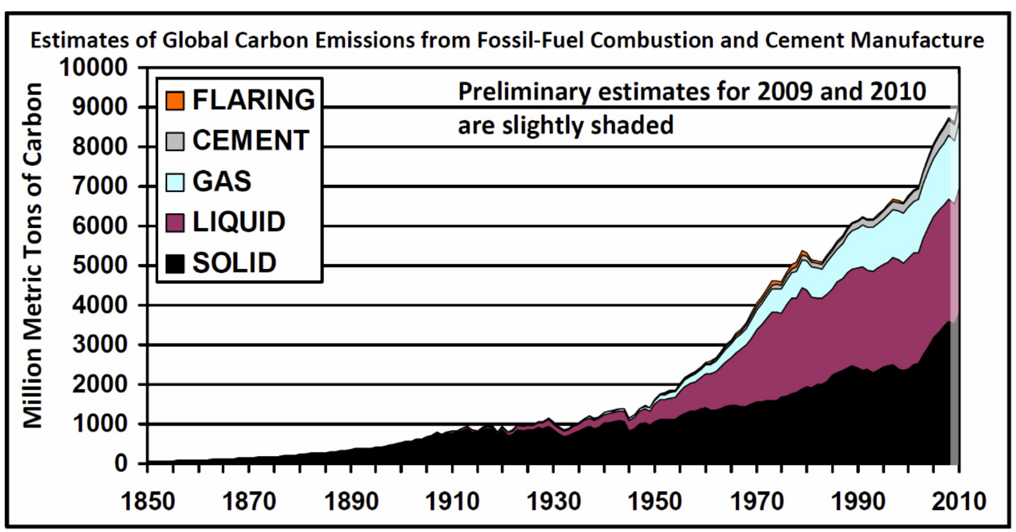
- CLAIM 1: The temperature of the planet is rising and the increase is due to human activities such as fossil fuel use and deforestation.
  - A good hypothesis often comes from a predictive model

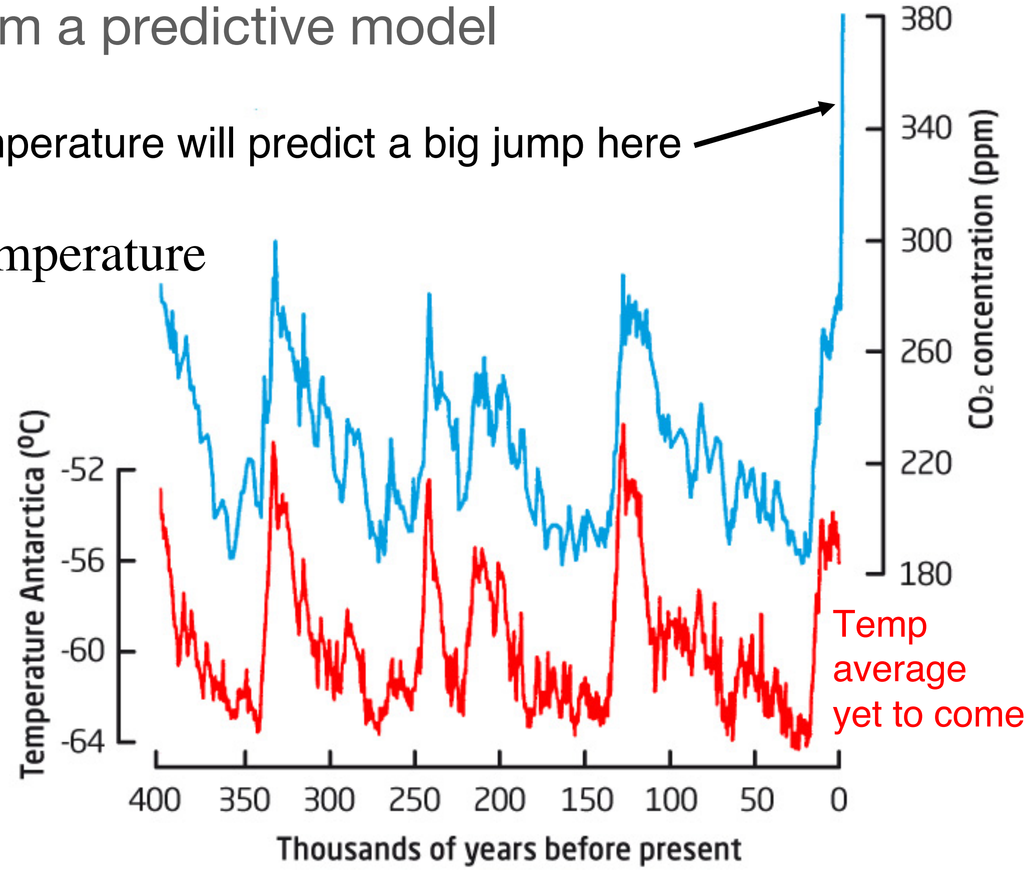
Model trained on CO2 data to predict Antartica's temperature will predict a big jump here

$$f(x_{CO2}) = \text{Temperature}$$

But is CO_2 jump related to human activity?

$$g(x_{fossil\text{-}fuel}) = \text{Atmospheric } CO_2 - \text{Natural } CO_2$$





Directional Relational Hypothesis

Examples of Student Answers

- Considering developing and developed countries, compare change in human-generated greenhouse gas emissions over a time period between the two countries

$$g(x_{fossil\text{-}fuel}) = \text{Atmospheric } CO_2 - \text{Natural } CO_2$$

- "…use past data to create a model, such as a graph that graphs the amount of CO2 released into the air per year… since carbon dioxide increase temperature…"

$$f(x_{CO2}) = \text{Temperature}$$

- And some outstanding answers… a great one that cited NASA and Scientific American.

Aspirin is effective in reducing cancer risk

- Here, we are looking at causal effects…
- Data
  - A person represented by random variable $X \in$ {{Age}, {Sick, Not Sick}, …}
    - Recruit people: x_{john} , x_{mary} , x_{eve} , x_{adam} , …
- Hypothesis
  - Force ½ (randomly chosen) of the people to take aspirin : $Y_{X,aspirin} \in \{1 \text{ - Cancer in 1yr}, 0 \text{ - No Cancer in 1yr}\}$
  - Force remaining ½ to NOT take aspirin: $Y_{X,no\_aspirin}$
    - Hypothesis: $E[Y_{X, ,no\_aspirin}] > E[Y_{X, ,aspirin}]$

Directional
Causal Hypothesis

Q3: Fathers who perform an equal share of household chores are more likely to have daughters who aspire to less traditionally feminine occupations

- Simplest Possible Data:
  - A random child represented by three random variables (X,F)
    - $X \in$ {Boy, Girl}, $F \in$ {Father Helps, Does not Help}
    - Actually recruit very young children (1/2 boys, 1/2 girls): x_{john} , x_{mary} ,…
    - Observe father f_{john}, f_{mary} , …
- Hypothesis
  - Check the aspiration of the child $Y_F \in \{0 = \text{Traditional}, 1 = \text{Untraditional}\}$
  - Hypothesis:
    $E[Y_{\text{Father Helps}} \mid x = \text{Girl}] - E[Y_{\text{Father Helps}} \mid x = \text{Boy}] >$
    $E[Y_{\text{Does not Help}} \mid x = \text{Girl}] - E[Y_{\text{Does not Help}} \mid x = \text{Boy}]$

Directional
Relational Hypothesis

Warning: Careful with Observation Biases

Warning: Observation Biases are Prevalent

- Your experience with buses at peak hours:
  - Bus at 99% capacity at peak hours
  - You wait on average 17 minutes and 9 seconds for it to arrive
- Transportation admin:
  - *buses at peak hour are at 60% capacity*
  - *average bus inter-arrival time is 10 minutes*

 ?

Inspection Paradox

- 40 minutes / 4 buses = 10 min inter-arrival time
- How long do you wait?
  - Assume you arrive uniformly during these 40 minutes
  - What is the probability you will arrive within the 37 minute interval?
    - P[Arrive at 37 min interval] = 37/40
    - What is the average waiting time if you arrive at the 37 min interval?
      - E[Wait | Arrive at 37 min interval] = 37/2 = 18.5

$$E[\text{Wait}] = \sum_i E[\text{Wait}|\text{Interval } i] P[\text{Interval } i] = \frac{37}{40} \times \frac{37}{2} + 3 \times \frac{1}{40} \times \frac{1}{2} = 17.15$$

Observation Biases in Data

- Over the years, Democrats (DEM) have argued that **average donation** to their candidates are smaller than that of Republican (GOP) candidates.
    - Per candidate average

$$\text{Average} = \sum_{z \in \text{candidates}} \frac{E[Y|X=x, Z=z]}{(\text{no. candidates at party } x)} = \sum_{y=1}^{\infty} \sum_{\forall z \in \text{candidates}} y \frac{P[Y=y|X=x, Z=z]}{(\text{no. candidates at party } x)}$$

Probability computed as % of donations with value y of candidate z in the file GOP_donations_2012.csv ?

What if there are GOP candidates with NO donations?
Do we have the right data to compute this probability?

End of Observation Bias Warning

Testing a Hypothesis

- population A

- population B

How much does it cost?

- population A

- When writing use the format:

  $1000 (no decimal points)

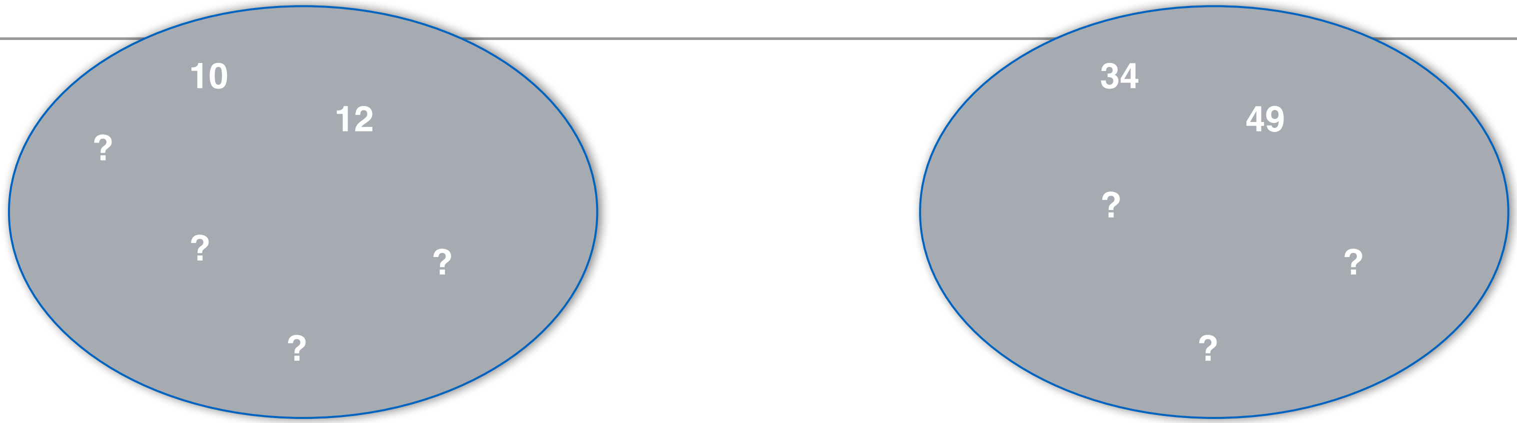
How much does it cost?

- population B

- When writing use the format:

  $10 (no decimal points)

Testing Hypotheses over Two Populations



Average μ_1 Average μ_2

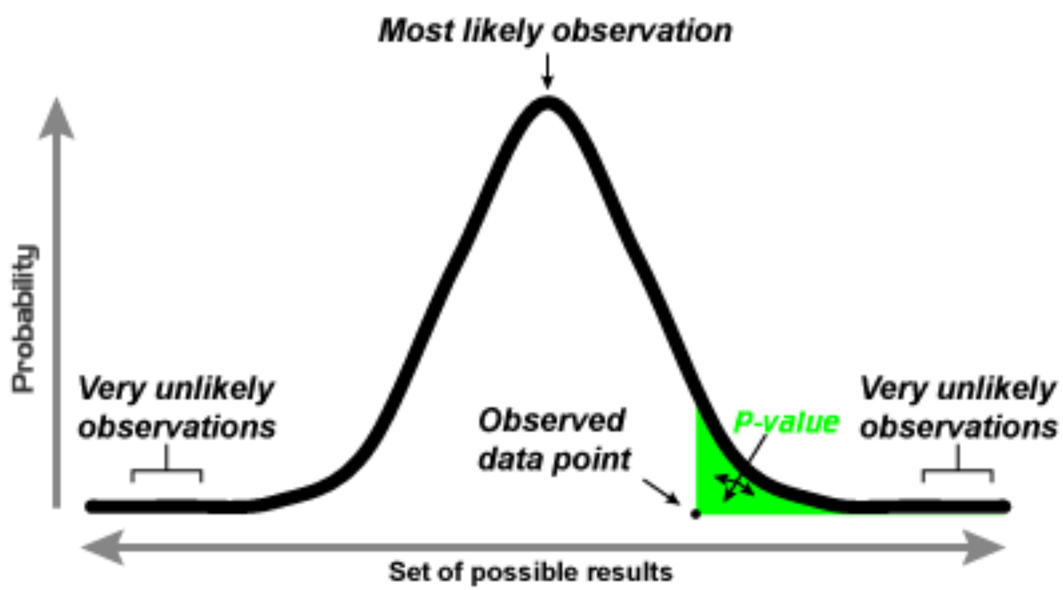
Hypothesis (Anchoring): Unrelated number biases wine price assessment
Experiment: Expose pop A to $1000 and pop B to $10
Population size: ~7 in each group

Answers from folks that saw $1000 as the amount
[50,10,37,650,400,80,130] … average $193

Answers from folks that saw $10 as the amount format
[20,30,60,10,100,40] … average $43

How can we test if hypothesis is true?

Standard Statistical Hypothesis Testing



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

- Traditional Hypothesis testing relies on p-values
- Roughly, the probability that we should see something a difference this extreme
  - \$193 - \$43 = \$150

Source:

ASA News

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

Simulating Alternative Universes

- Rather than p-values, we will see a computational approach

- We will “simulate” redoing the experiment hundreds of times…
  - But we will not actually redo the experiments… (see iPython Notebook)

Decision making

Psychological heuristics and biases

- Tversky & Kahneman, psychologists, propose that people often do not follow rules of probability when making decisions
- Instead, decision making may be based on heuristics
  - Lowers cognitive load but may lead to systematic errors and biases

- Examples:
  - Availability heuristic
  - Representativeness heuristic
  - Confirmation bias
  - Conjunction fallacy (we will not cover this)
  - Numerosity heuristic (we will not cover this)

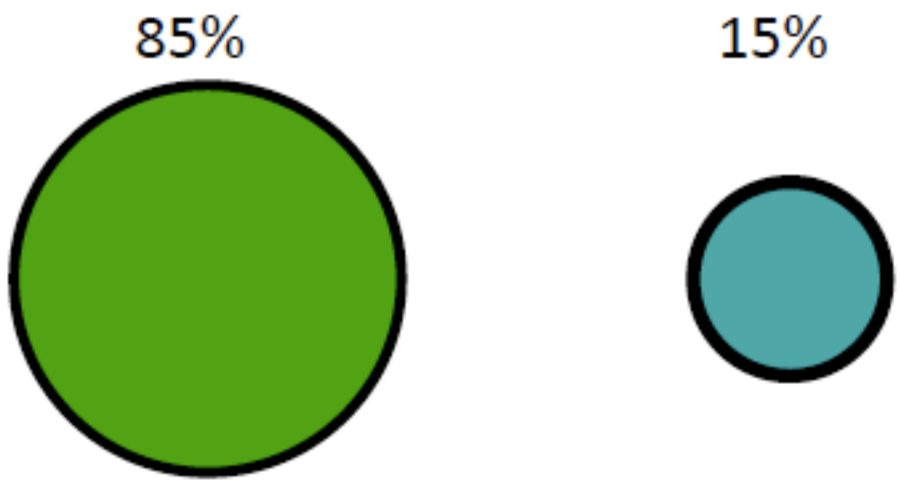
Neglecting base rates

- Taxi-cab problem *(Tversky & Kahneman '72)*
    - 85% of the cabs are Green
    - 15% of the cabs are Blue
    - An accident eyewitness reports a Blue cab
    - But she is wrong 20% of the time.
- What is the probability that the cab is Blue?
    - Participants tend to overestimate probability, most answer 80%
    - They ignore baseline prior probability of blue cabs.

More on neglecting base rates

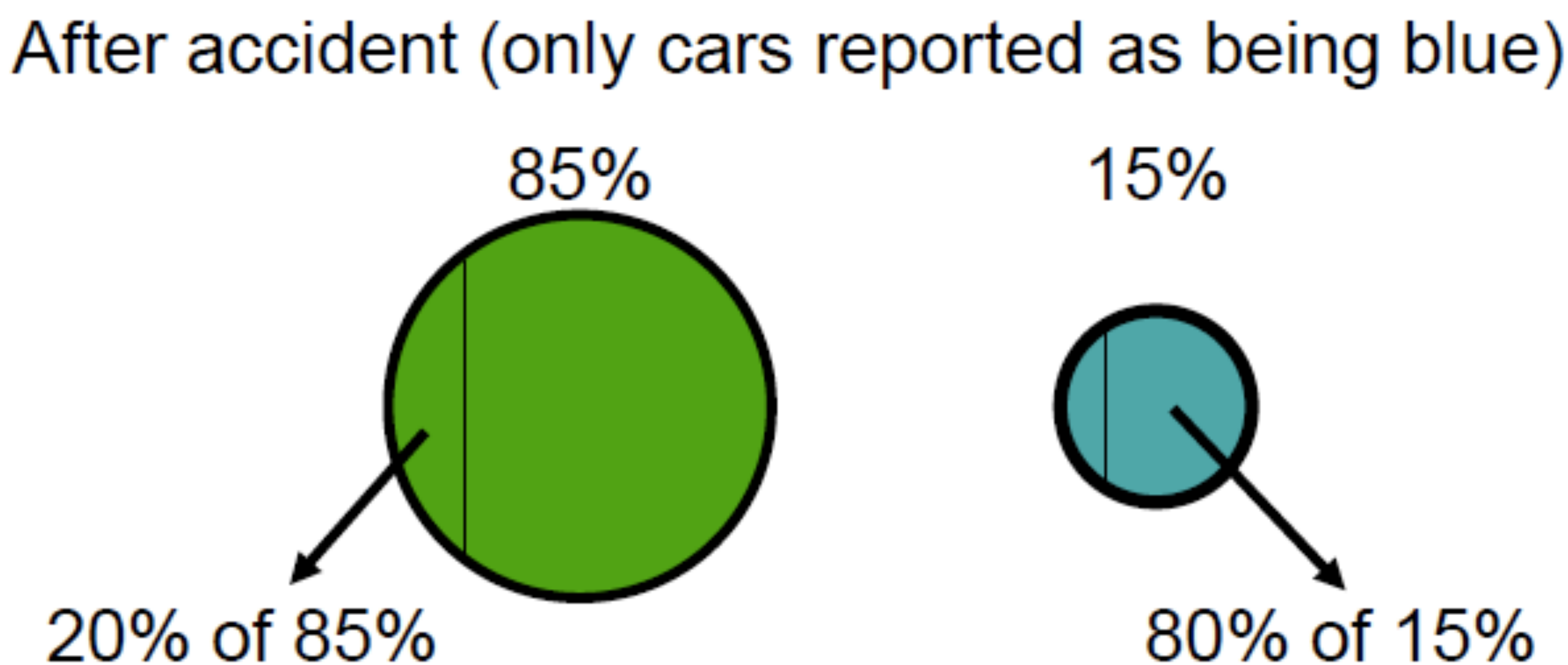
A priori (beforehand)

$$P(green) = 0.85$$
$$P(blue) = 0.15$$



$$P(seeBlue|blue) = 0.80$$
$$P(seeBlue|green) = 0.20$$

More on neglecting base rates



More on neglecting base rates

- How to compute probability

$$
\begin{aligned}
P(blue|seeBlue) &= \frac{P(blue \cap seeBlue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue|blue)P(blue) + P(seeBlue|green)P(green)} \\
&= \frac{0.80 \cdot 0.15}{(0.80 \cdot 0.15) + (0.20 \cdot 0.85)} \\
&= 0.41
\end{aligned}
$$

Most people answered 80%

Arthritis study *(Redelmeier & Tversky ’96)*

- Common belief:
  - **Arthritis pain is associated with changes in weather**
- Experiment:
  - Followed 18 arthritis patients for 15 months
  - 2 x per month assessed: (1) pain and joint tenderness, and (2) weather
- Results:
  - No correlation between pain/tenderness and weather
  - Patients saw correlation that did not exist... why?

Arthritis study (cont)

- Patients noticed when bad weather and pain co-occurred, but failed to notice when they didn't.
    - Better memory for times that bad weather and pain co-occurred.
    - Worse memory for times when bad weather and pain did not co-occur
- **Confirmation bias**: People often seek information that **confirms** rather than disconfirms their original hypothesis

Extra Examples

Estimating probabilities *(Tversky & Kahneman '73/'74)*

- *Question*: Is the letter **R** more likely to be the 1st or 3rd letter in English words?

- *Results*: Most said **R** more probable as 1st letter

- *Reality*: **R** appears much more often as the 3rd letter, but it's easier to think of words where **R** is the 1st letter

Estimating probabilities (cont)

- *Question*: Which causes more deaths in developed countries?
  (a) traffic accidents or (b) stomach cancer
- *Typical guess*: traffic accident = 4X stomach cancer
- *Actual*: 45,000 traffic, 95,000 stomach cancer deaths in US
- Ratio of newspaper reports on each subject:
  137 (traffic fatality) to 1 (stomach cancer death)

- **Availability heuristic**: Tendency for people to make judgments of frequency on basis of how easily examples come to mind

Gambler's fallacy

- *Gambler's fallacy*: belief that if deviations from expected behavior are observed in repeated independent trials, then future deviations in the opposite direction are then more likely
- T&K: this is an example of the **representativeness heuristic**—where the probability of an event is judged by its similarity to the population from which sample is drawn
- The sequence "H T H T T H" is seen as more representative of a prototypical coin sequence. Why?
    - When people are asked to make up random sequences, they tend to make the proportion of H and T closer to 50% than would be expected by random chance
    - T&K interpretation: people believe that short sequences should be representative of longer ones

Interpretation of these findings

- People do not use proper statistical/probabilistic reasoning... instead people use heuristics which can **bias** decisions
- Heuristics can often be very effective (and efficient) for social inferences and decision-making
    - E.g., the book “Simple Heuristics That Make Us Smart” summarizes research by Gigerenzer and Todd
- ... but be aware that **heuristics can bias** results from **exploratory data analysis and other modeling efforts**