Stat 104: Quantitative Methods

Class 1: The Nature of Statistics

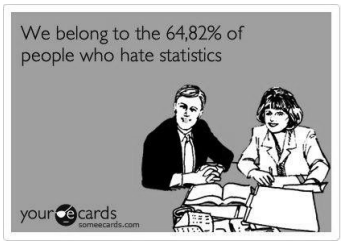
Historical Quote (1911)

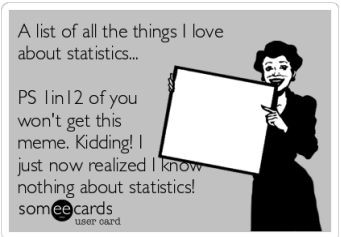
"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."-**H.G. Wells**

We are bombarded with statistics

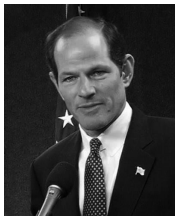
- " Women who breastfeed are half as likely to develop type 2 diabetes as women who do not " -Time Magazine
- " The number of social networking users ages 50 and older nearly doubled in the past year" –PCWorld Magazine
- " 60 percent of American men report that they would be willing to take a male birth control pill, if it was available"- Cosmopolitan
- " Owners of Apple's iPhone on average have more sex partners by age 30 than those who own other handsets" -AppleInsider.com

Yet people don't like statistics classes





And you want to be my latex salesman?



Where Am I?

Michael Parzen: Sci 300b

Phone: 617-495-8711

Email: mparzen@fas.harvard.edu

Who Am I?

HARVARD School of Public Health

CHICAGO BOOTH
The University of Chicago Booth School of Business

EMORY | GOIZUETA BUSINESS SCHOOL

Statistics

The Primary Teaching Staff

- Me (well duh): michaelparzen@gmail.com
  - Feel free to contact me for any problems/concerns/anything about the course/life.
- Department Preceptor and Head Teaching Fellow (TF) Kaitlin Hagan
  - Will do the weekly online section and is in charge of entire teaching staff: stat104kaitlin@gmail.com

Course Website

- Course website:
- https://canvas.harvard.edu/courses/27892
- There you will find:
  - Syllabus
  - Administrative Announcements
  - Lecture Notes (out at least 24 hours in advance)
  - R Tutorial
  - Assigned Homeworks

Weekly Sections

- There will be weekly optional sections. Section starts next week (details in class).
- Sections are on Wed and Thur. You may go to any section you want.
- There will also be a study network which consists of walk in help sessions (details next week).
- There is an online section (also videotaped) Thursdays at 5pm.

Course Intent

- Overcome Math Anxiety.
- Learn skills for other courses and the workplace (no Las Vegas Syndrome).
- Use statistical tools to identify problems and opportunities. Learn techniques which help you make decisions.
- Learn to think about numbers, magnitudes, tradeoffs, constraints, modelling and risks/likelihoods.

What is this course all about ?

- Recognize the importance of correct data collection.
- Learn to use statistical software (R and Rstudio)
- Apply estimation and testing methods to understand natural phenomena and make data-based decisions.
- Investigate relationships between variables.
- Interpret results correctly, effectively, and in context without relying on statistical jargon.

Computing

- For **all exams**, you will need a calculator with log, exponential, square-root functions.
- Statistical Computing Package: R and Rstudio
  - See course website for instructions on accessing R and Rstudio.
  - Tutorial available on the class website
    Next week's section will include an introduction to the software
- You DO NOT need R for the exams.

A Note on R and Rstudio

What is R?

- A computer language, with orientation toward statistical applications
- Relatively new
- Growing rapidly in use
- Its what the professionals are now using

Why R?

- Language of statisticians
- Becoming a common language for data analysis in all fields
- Code is a form of communication
- It's easy to extend
- It's free (and runs on all platforms)

You'll be in good company

We'll be using R pretty stupidly

- **As a fancy calculator**
- To look up areas under curves
- To make plots
- To run statistical tests
- To code and analyze simulations;Nope-see my Stat 109 class in the spring for that.

What's RStudio?

- Invented by JJ Allaire who lives nearby (and is not a statistician)
- R is the language
- RStudio is the place to write R code, run R code and examine the results.
- RStudio is a separate program that runs R.
- RStudio looks identical regardless of whether you are on Mac, Windows, Linux or using a web version.

Assessment

Percentage	Component	Due Date
8%	Homework	Weekly'ish Out Monday, Due next Monday
8%	Online Quizzes	Weekly
4%	Regression Project	Reading Period
20%	Exam 1	October 2
20%	Exam 2	November 6
40%	Final	To Be Determined

For full details read the class syllabus

Why study decision analysis/statistics ?

- Most of us are not very good at analyzing probabilities or understanding data.
- We have inherent biases when assigning probabilities to events, and like to use heuristics (short cuts that lead to decisions that are not necessarily correct).

"It ain't so much the things we don't know that get us into trouble. It's the things we know that just ain't so."

Artemus Ward

Why Study? There are jobs!

Here's a Retail **Job** That's Still in High Demand: **Data Scientist**
Bloomberg - Aug 21, 2017
Despite an onslaught of bankruptcies, store closures and layoffs, there's one retail **job** that's still in hot demand: **data scientists**.

IBM Expecting Demand For **Data Scientists** To Grow By 28% By 2020
Dispatch Tribunal - Aug 15, 2017
The findings does not come as a surprise given that 19% of all **job** openings in the finance and insurance sector are for **Data Science** and ...

EWU and Microsoft join forces on new degree program
The Spokesman-Review - Aug 25, 2017
Students who complete the program will graduate with a B.S. in data ... Industry analysts anticipate 1.5 million unfilled **data science jobs** in the ...

What is a **data scientist**? A key data analytics role and a lucrative ...
CIO - Aug 18, 2017
It's a fast growing and lucrative field, with the BLS predicting **jobs** in this field will grow 11 percent by 2024. **Data scientist** is also shaping up to ...

Jobs(?) Hell Yeah

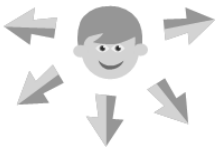
Most in-demand skill in 2014, according to **LinkedIn**.

Among the **fastest-growing jobs**. The McKinsey Global Institute predicts a shortage of up to **190,000 people** with the data analysis skills to work with Big Data.

Great Field to Study

More than **40%** of statistics undergraduate degrees go to **women**.

Ranked by *Fortune* as the **top graduate degree** based on salary, growth and job satisfaction.



Offers a wide variety of **careers to choose from**.

Jobs (again!)

Jobs Rated Report 2016: Ranking 200 Jobs

By CareerCast.com

The job landscape never stops evolving. One need look no further than the 2016 CareerCast.com Jobs Rated report, now in its 28th year.

Over the course of nearly three decades, changes to the economic and career environment have rendered once desirable jobs now almost nonexistent. In their place come exciting new job prospects.

New serves as a central theme for the 2016 Jobs Rated report. Careers newly added to the report made a splash.

Jobs Rated Links

- The Worst Jobs of 2016
- The 10 Best Jobs of 2016
- Our Methodology
- Infographic

Numbers 1 and 2

The Best Jobs of 2016: 1. Data Scientist

2016 Jobs Rated Score: 91
Annual Median Salary: $128,240
Growth Outlook : 16%

Opportunities across a variety of fields make data scientist not just a high-growth job, but also one of the most lucrative tracked by the Jobs Rated report.

The Best Jobs of 2016: 2. Statistician

2016 Jobs Rated Score: 92
Annual Median Salary: $79,990
Growth Outlook : 34%

The growing importance of statistical analysis in a variety of fields makes for abundant opportunities for trained statisticians. At 34%, the growth outlook for statistician is among the highest of any job in the 2016 Jobs Rated report.

Numbers 200 and 199

The Worst Jobs of 2016: 200. Newspaper Reporter

2016 Jobs Rated Score: 734
Annual Median Salary: $37,200
Growth Outlook : -9%

A gradual decline in print publications at the turn of the century became a steep downturn for the past decade. Publications folding mean far fewer job prospects, and declining ad revenue means unfavorable pay for those in the Fourth Estate.

The Worst Jobs of 2016: 199. Logger

2016 Jobs Rated Score: 724
Annual Median Salary: $35,160
Growth Outlook : -4%

Declining use of paper, as more content moves to digital, has negatively impacted employment prospects in the logging industry. The job's high stress and dangerous work environment also contribute to its place in the worst jobs of 2016.

Statistical Knowledge is Vital

- Hal Varian, Google's chief economist, said the following in a McKinsey Quarterly interview:

  "I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take **data**—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely **important skill** in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."

Analytics/Statistics is the New Know

From Wikipedia: Statistics is….

- Statistics is the formal science of making effective use of numerical data relating to groups of individuals or experiments.
- As we will see, statistical thinking is involved in:
  1. Defining scientific hypotheses
  2. Designing experiments and collecting data
  3. Analyzing the data
  4. Interpreting the results

Dr John Snow Saves London

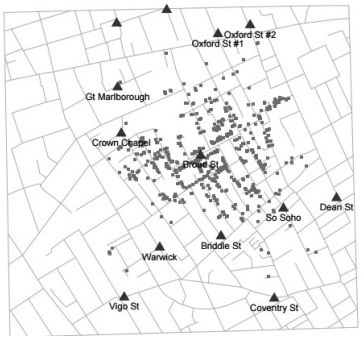
- Cholera is an acute, diarrhoeal illness caused by infection of the intestine with the bacterium *Vibrio cholerae*. The infection is often mild or without symptoms, but sometimes it can be severe.
- In these persons, rapid loss of body fluids leads to dehydration and shock. Without treatment, death can occur within hours.
- 1854 Outbreak in London; 500 deaths in 10 days.

1854 Data Science

- In 1854, London water was provided by competing private firms
- Residents would walk to the nearest street pump for water
- Snow recorded the location of each death in real time
- He placed these spatial data on a map, along with the water pumps
- Was one pump, from a particular company, contaminated with cholera?

The Data

- How do we assess the relationship between deaths (red dots) and pumps (blue triangles)?
- Are we convinced that a relationship exists?

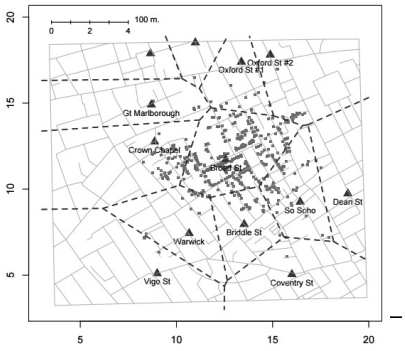


Could Study Distance from Pump

Fact: For any spot on the map, there is a closest pump or pumps

Modeling Assumptions:
- Some (not all) pumps are contaminated
- People use the closest pump

Model prediction: Pattern of deaths should correspond to nearest-pump boundaries (in blue)



Data Directed Decision Making

- Snow used his data and map to convince officials to remove the handle from the Broad Street pump.
- Credited with stopping the outbreak and providing first experimental evidence for germs

Some Questions to Consider (later)

1) Did the Broad Street Pump really cause the cholera outbreak?
2) Did removing the handle stop it?
3) Can we measure our uncertainty about our answers to 1 and 2?

More mundane experiment

- What percentage of Earth is covered by water?
- Lets design an experiment.

Ideally by the end of this course…..

- You'll be able to present and model data in a clear and concise fashion like Dr John Snow, test claims and (very important!) understand the common fallacies that occur everyday with decision analysis and statistics.
- The following slides look at some of the fallacies/problems that come up in statistics and decision making (critical thinking).

Question what you hear

- On NPR a few weeks ago, I heard a most interesting statistic. A researcher at some university had arrived at the astounding conclusion that the more male children you have in a family, the more likely it is that one of them will grow up to become president.
- Why stop there? I would argue that this also increases the likelihood that one of them will grow up to attend Harvard, or a used car salesman, or be hit by a meteorite, or win the lottery, or contract syphilis, or rob a bank, or die while ice-fishing, or...

Question what you hear

- A recent ad on the radio says that "Children's Hospital of Boston was recently ranked in the top 3 of all national pediatric hospitals."
- What conclusions can you draw from this ?

Question what you read

- Consider a headline that invites us to infer a causal connection: BOTTLED WATER LINKED TO HEALTHIER BABIES.
- Without further evidence, this invitation should be refused, since affluent parents are more likely both to drink bottled water and to have healthy children; they have the stability and wherewithal to offer good food, clothing, shelter, and amenities. Families that own cappuccino makers are more likely to have healthy babies for the same reason.

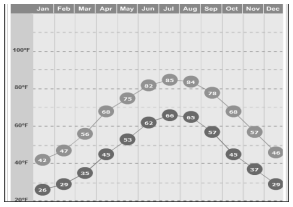
Correlation measures association. But association is not the same as causation

Question Comparisons

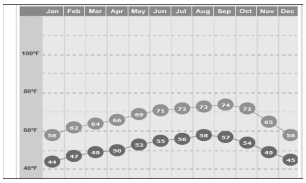
- The average temperature of Alameda,CA is 67, very similar to the average temperature of Charleston, WV (which is 66).
- Indeed.
- Question what is being compared, and why they decided to compare averages, instead of min,max, shape, etc……

Graphical Temp Comparison

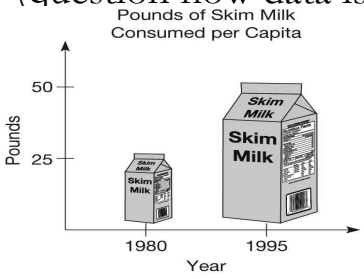
Charleston, WV



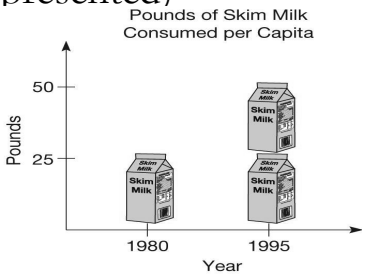
Alameda, CA



What's Wrong with This Picture? (question how data is presented)

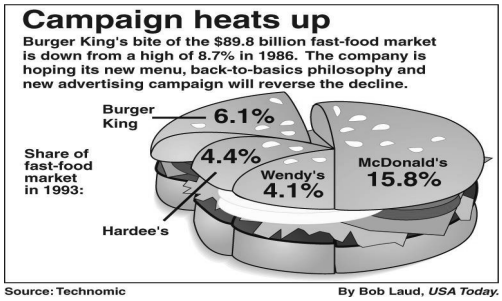


Source: US Department of Agriculture



Source: US Department of Agriculture

What's Wrong?



Scale Matters

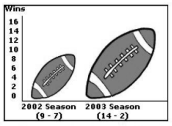


Figure 3: Regular season wins for the New England Patriots, 2002-2003.

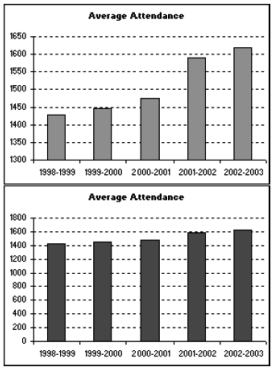


Figure 2: Two different charts showing average attendance at NCAA Women's Soccer (season) matches.

What's Wrong With This Sample?

- A uniformed police officer surveys teens at a local high school track meet about drug activity in the school.
- Third grade students are asked, "How nutritious are your evening meals?"
- Compare the following survey questions:
  - "Should the university use trees harvested during construction to build furniture in the new chemistry building?"
  - "Should the university cut down trees in the construction zone to use as furniture in the new chemistry building?"

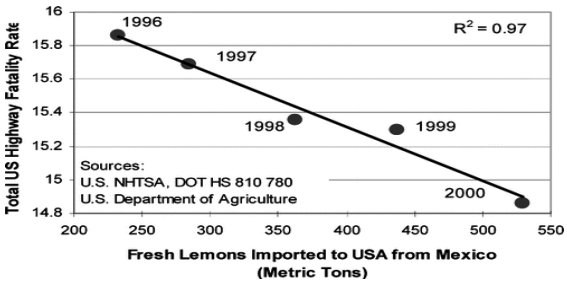
Choice of Sample

- The Hawaii State Senate held hearings when it was considering a law requiring that motorcyclists wear helmets. Some motorcyclists testified that they had been in crashes in which helmets would not have been helpful.
- Which important group was unable to testify?

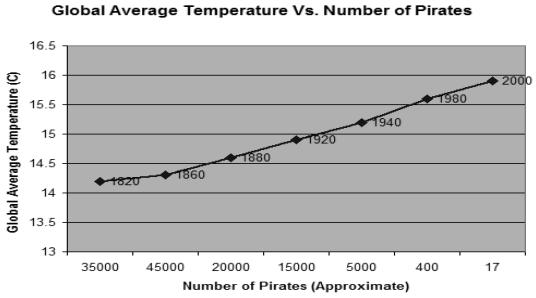
Question Framing

- 97% yes: "Should the President have the line item veto to eliminate waste?"
- 57% yes: "Should the President have the line item veto, or not?"

Correlation is not Causation



Pirates and Global Warming



Use data to prove/disprove a theory

The Hot Hand Theory

"If I'm on, I find that confidence just builds... you feel nobody can stop you. it's important to hit that first one, especially if it's a swish. Then you hit another, and...you feel like you can do anything."
~Lloyd Free (a.k.a. World B. Free)

What is the Hot Hand Theory ?

- Belief that success breeds success, failure breeds failure
- 100 basketball fans surveyed:
- 91% thought "player has better chance of making a shot after having just made his last two or three shots than he does after having just missed his last two or three shots."
- 84% thought "it's important to pass the ball to someone who has just made several shots in a row."

What does the data say ?

- Calculate probability of making a shot after missing previous 1, 2, or 3 shots and after making the previous 1, 2, or 3 shots.

Table 2.1 Probability of Making a Shot Conditioned on the Outcome of Previous Shots for Nine Members of the 76ers

Player	*P*(x|*ooo*)	*P*(x|*oo*)	*P*(x|*o*)	*P*(x)	*P*(x|x)	*P*(x|xx)	*P*(x|xxx)	*r*
C. Richardson	.50	.47	.56	.50	.49	.50	.48	−.02
J. Erving	.52	.51	.51	.52	.53	.52	.48	.02
L. Hollins	.50	.49	.46	.46	.46	.46	.32	.00
M. Cheeks	.77	.60	.60	.56	.55	.54	.59	−.04
C. Jones	.50	.48	.47	.47	.45	.43	.27	−.02
A. Toney	.52	.53	.51	.46	.43	.40	.34	−.08
B. Jones	.61	.58	.58	.54	.53	.47	.53	−.05
S. Mix	.70	.56	.52	.52	.51	.48	.36	−.02
D. Dawkins	.88	.73	.71	.62	.57	.58	.51	−.14
Mean =	.56	.53	.54	.52	.51	.50	.46	−.04

NOTE: x = a hit; o = a miss. r = the correlation between the outcomes of consecutive shots

Implication

- Probabilistic laws and statistical tools will aid in understanding what's going on around us(and hence help us make better decisions).