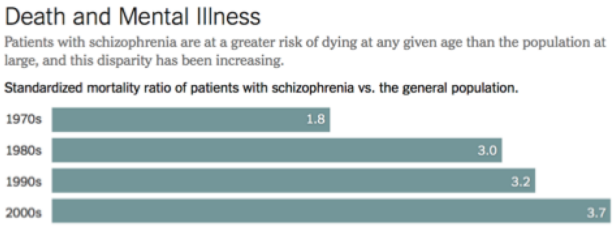
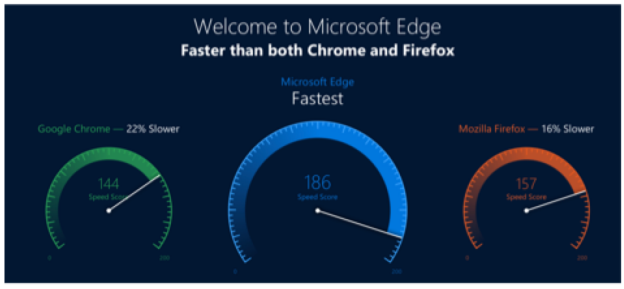
Stat 324 – Introduction to Statistics for Engineers

LECTURE 1: INTRODUCTION TO STATISTICS, COURSE, AND A LITTLE R/RSTUDIO

Why Study Statistics?

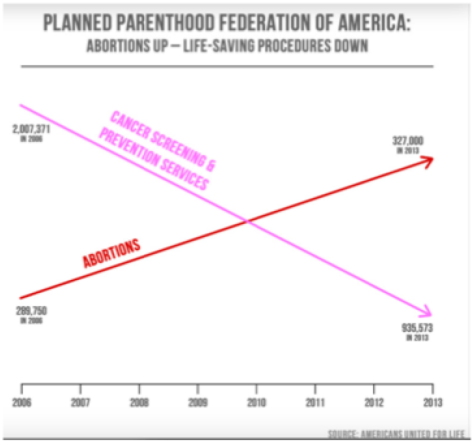


NYTimes, Dhruv Khullar, May30, 2018

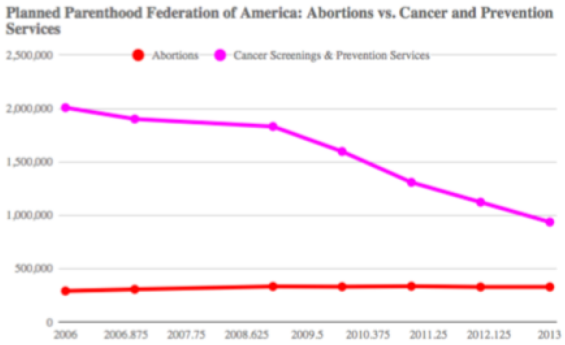


Microsoft Welcome Page June 12, 2018

Why Study Statistics?



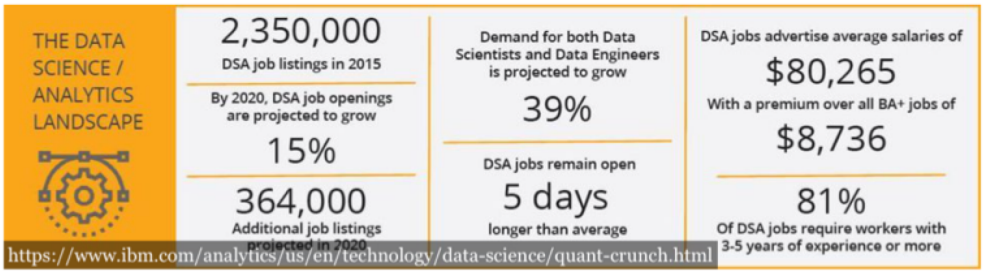
Presented by Jason Chaffetz on Sept 29th, 2015 in Congressional Hearing



Alternative presentation of data

Why Study Statistics?

When scientific measurements are repeated, they come out somewhat differently each time (even if experiment or observation is performed exactly the same) because of random chance!



"I wish I or my PI knew more Statistics so I could better understand publications and how to better conduct my studies. Can you come work in my lab?"

– PhD student in Mechanical Engineering June 7th, 2018 in my office

A little about me…

Department of Mathematics
Van Vleck Hall, 480 Lincoln Drive, Madison, WI

Epic Systems
Software company

Department of Statistics

SAN JOSE COLLEGIATE
KIPP
EST. 2008
AD VITAM PARAMUS

Milwaukee Pulaski
High School Rams

A little about me…

A little about you..

1. Prefer Memorial Terrace or the beach?

2. How many courses this semester?

3. Do/have done work in labs/internships? Asked to read/evaluate/perform statistics?

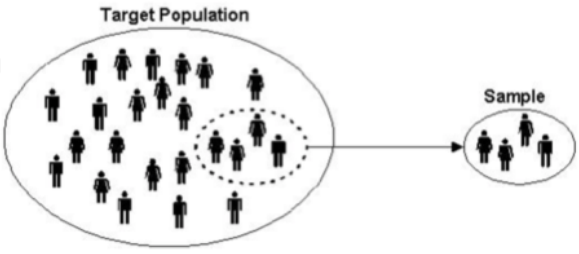
What is Statistics?

Necessary Vocabulary

- A ***population*** is the entire collection of objects or outcomes about which information is sought.
  - Census:
- A ***parameter*** is a numeric summary of a Population.

- A ***sample*** is a subset of a population, containing the objects or outcomes that are actually observed
- A statistic is a numeric summary of the sample.

- Ex: I'd like to know the percent of UW Madison students who would cheer for Ohio State if they made it to the Final 4.



What is Statistics?

What can we get information about? And is this the only group we're interested in?

Population or Sample Information?

Sports : A single player's batting average sample

Company/Organization. : Median employee salary population

Politics : *Polls* sample

Historical Weather Patterns : Average temperature for a day population

Science: Number of ash trees in 10 sections of forest that have emerald ash borer sample

What is Statistics?

A. Descriptive Statistics

- _____brief_____ and _____descriptive_____ summaries of observed data.

- Summarizing data in "useful" ways that could reveal interesting patterns lurking within.

- The type of data we've collected and the questions we have will drive what summaries are most useful/appropriate.

What is Statistics?

B. Inferential Statistics

- Making inferences (educated guesses/_____) about a population by studying a relatively small sample chosen from it.
  - A *point estimate* is a statistic used to estimate a parameter. E.g.
  - A ___Confidence interval_____ is a range of probabilistically plausible values for a parameter, in light of a sample.
  - In a *hypothesis test* , we write a hypothesis about a population parameter and then reject or retain it in light of the evidence _____from the sample data.
  - In a *linear model,* we estimate and make claims about the ____correlation_____ and _____Best fit_____of a line relating one variable to another.

Getting Data to Consider

Controlled Experiments **vs** **Observational Studies**

- Values of factors are under the control of ____The experimenter_____.
- Can produce reliable information about ____Cause and effect_____ relationships between factors and response.

E.g. Chemical Engineer wants to know how the concentrations of reagent and catalyst affect the yield of a process. Can run the process several times, changing concentrations, and compare the yields that result.

- Values of factors are **not** under the control of experimenter and are just observed.
- Much more ___difficult_____ to demonstrate cause-and-effect relationship.

E.g. Studies conducted to determine the effect of cigarette smoking on the risk of lung cancer cannot force people to smoke

All Data is not Useful Data

A good sample requires some degree of randomness.

- In a **Simple Random Sample (SRS)** each group of size n (sample size) is _____ Equally likely _____ to be drawn as the sample
  - E.g. Choose 5 lottery winners out of 10,000 lottery tickets by putting the 10,000 tickets in a drum, mix them thoroughly, and then reach in and one by one draw 5 tickets out.
- Simple random samples ___not_____guaranteed to reflect the population perfectly, but there should be no _____ mechanism tending to make the sample _____biased_____.
- Ideally, the differences between the sample and its population are due entirely to random variation between the samples– called _____Stratified random_____.

All Data is not Useful Data

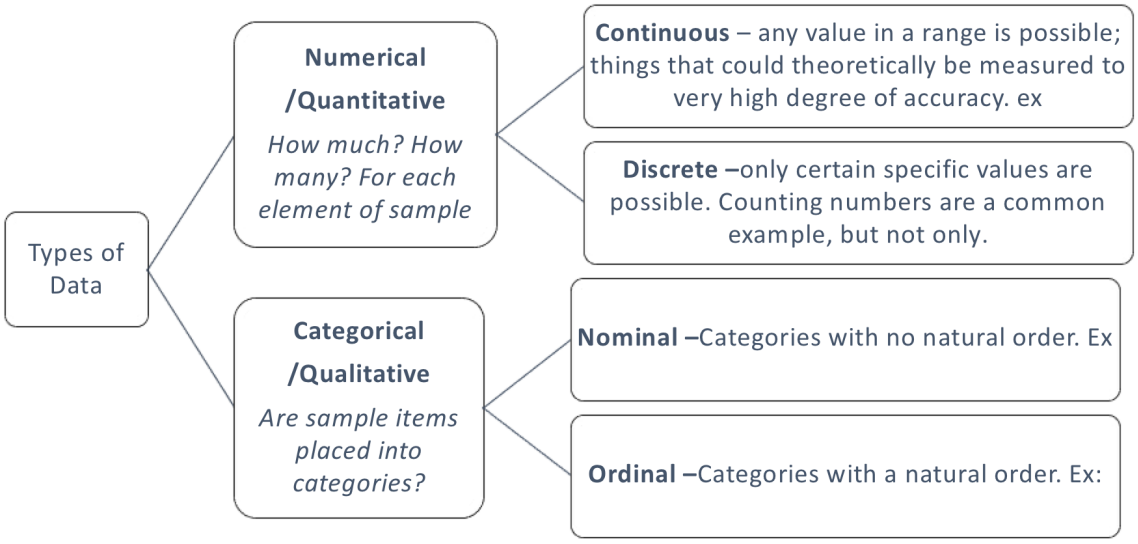
_____convenience_____ **samples are not drawn by a well-defined random method**

1. A quality inspector takes 100 random bolts from the top of one carton of bolts that was shipped to the company to get evidence for whether the entire carton of bolts meets compliance.
2. An industrial engineer measures the paint thickness on the first batch of toys produced each morning to get evidence for whether the automated painting process is working correctly.
3. A news program conducts a call in poll of its audience members to gauge support/opposition of proposed legislation that was the subject of a lead story the previous night.

Problem: A sample of convenience may differ systematically in some way from the population – any estimates we make from these samples may suffer from **bias**

Types of Data

The types of data you are interested in drives the type of descriptive and inferential statistics you use – not always cut and dry designation - more on this later



Types of Data

Often, quantitative and categorical data are collected in the same experiment or observational study.

Ex: Iris data

	Sepal.Length <dbl>	Sepal.Width <dbl>	Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

6 rows

Variables and their Type:

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Course Information

GOALS

1. Articulate the basics of probability and statistics
2. Produce appropriate numeric and graphical summaries of simple data.
3. Recognize and apply statistical methods appropriate for analysis of a data set in some situations, and know when to ask a statistician for help.
4. Relate data and analyses back to the original context in which data was collected.
5. Use R, a free statistical software package, for statistical computations and graphs.

Logistics

1. Bring your computer to lecture and discussion so you can get some R practice as we go.

2. Lecture and Discussion attendance are essential for getting the practice necessary to become proficient.

3. Read the Syllabus for logistics questions.

4. Post Homework, Logistics, class-wide questions to Piazza.

Computational Tools we will use

- R (free statistical programming language) and R Studio (a free integrated development environment)

- Calculator for exams so should also practice using calculator on HWK.

- You will also need to be able to read some R for exams.

For Next Time:

1. Read through the syllabus and put the exam dates on your calendar (if you are a reminders person, also add homework due dates.)

2. Install R and Rstudio on your computer and then bring your computer with you to lecture and discussion.

3. Save off course files for tomorrow so you can work through R with me in class.

A Little R Demo if Time

- Using “catalyst” data set
  - A certain reaction was run several times using each of two catalysts, A and B. The catalysts were supposed to control the yield of an undesirable side product. Results, in units of percentage yield, for 4 runs of catalyst A and 6 runs of catalyst B are as follows:
  - CATALYST A: 4.4 3.4 2.6 3.8
  - CATALYST B: 3.4 1.1 2.9 5.5 6.4 5.0

- Open .Rmd file in R Studio and run along with me if you’ve installed R and R studio. If you haven’t yet, then run through it after you’ve installed them.