Microsoft

MLADS

MACHINE LEARNING, AI,
AND DATA SCIENCE CONFERENCE

November 14 - 18

1

Microsoft

Introduction to Statistics

Shep Sheppard
shep.sheppard@microsoft.com
https://www.linkedin.com/in/sqlshep/
@sqlshep

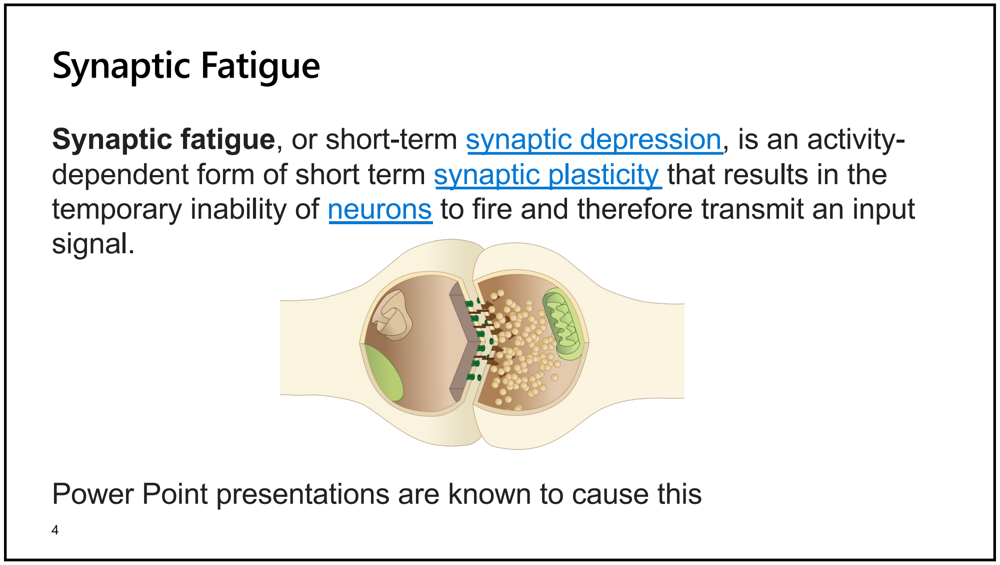
2

Microsoft

Introduction to Statistics

Shep Sheppard
Customer Engineer
Fast Track ISV

3



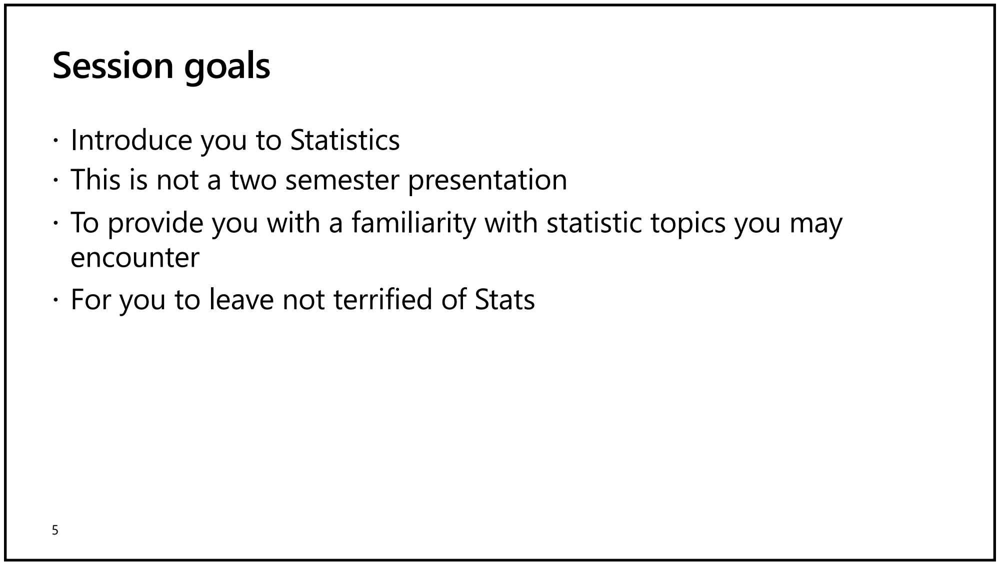
Synaptic Fatigue

Synaptic fatigue, or short-term synaptic depression, is an activity-dependent form of short term synaptic plasticity that results in the temporary inability of neurons to fire and therefore transmit an input signal.

Power Point presentations are known to cause this

4

4

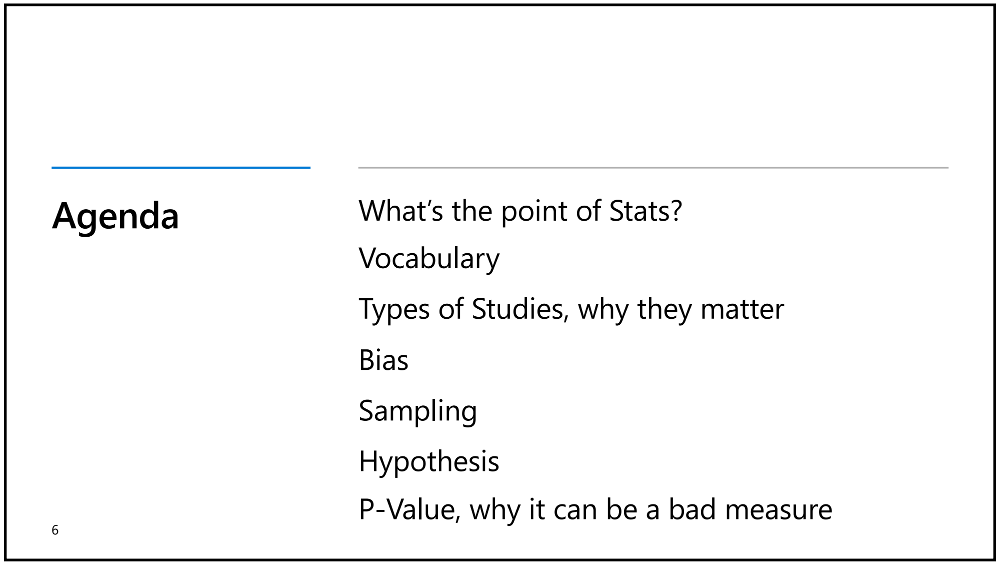


Session goals

- Introduce you to Statistics
- This is not a two semester presentation
- To provide you with a familiarity with statistic topics you may encounter
- For you to leave not terrified of Stats

5

5



Agenda

What's the point of Stats?

Vocabulary

Types of Studies, why they matter

Bias

Sampling

Hypothesis

P-Value, why it can be a bad measure

6

6

What’s the point of Statistics?

- Identify a question or a problem
- Collect relevant data on the topic
- Analyze the data
- Form a conclusion

YOU KNOW HOW STUDIES ALWAYS SAY ONE OUT OF TEN PEOPLE HAVE A PARTICULAR PROBLEM?
Dilbert.com DilbertCartoonist@gmail.com
I'M **ALWAYS** THAT GUY. STATISTICALLY SPEAKING, I KEEP NINE PEOPLE SAFE JUST BY EXISTING.
1-24-15 © 2015 Scott Adams, Inc. /Dist. by Universal Uclick
THAT'S NOT HOW STATIS— TICS WORK.
AND... EVERYONE ELSE IN THE DEPARTMENT KNOWS THAT?

7

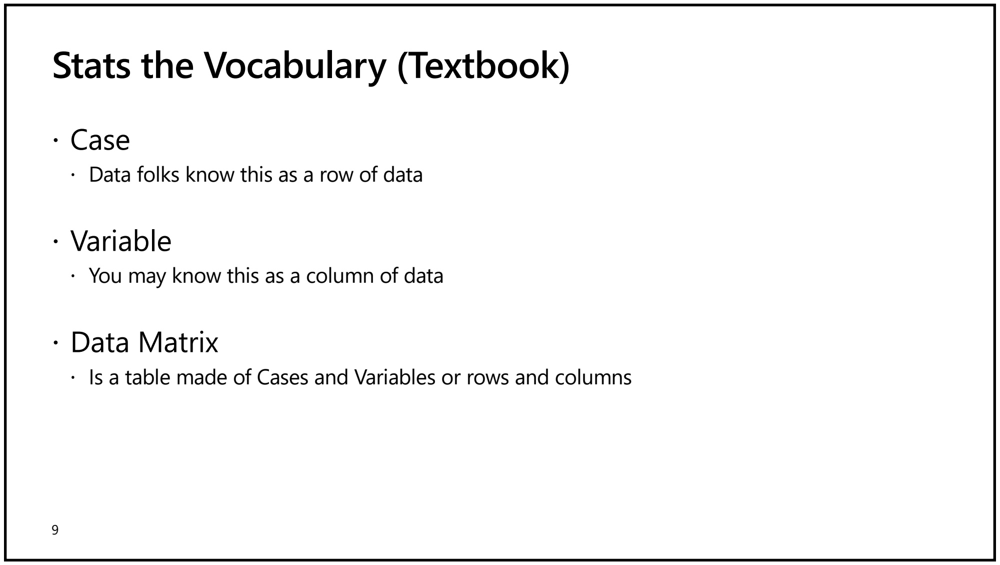
7

Stats the Vocabulary

- Treatment Group
  - The group you are experimenting on and monitoring
- Control Group
  - The group you are not experimenting on and monitoring
- Be aware of ethics of both groups, even if it is A/B testing
  - Tuskegee experiments
  - OKCupid
  - Facebook

8

8

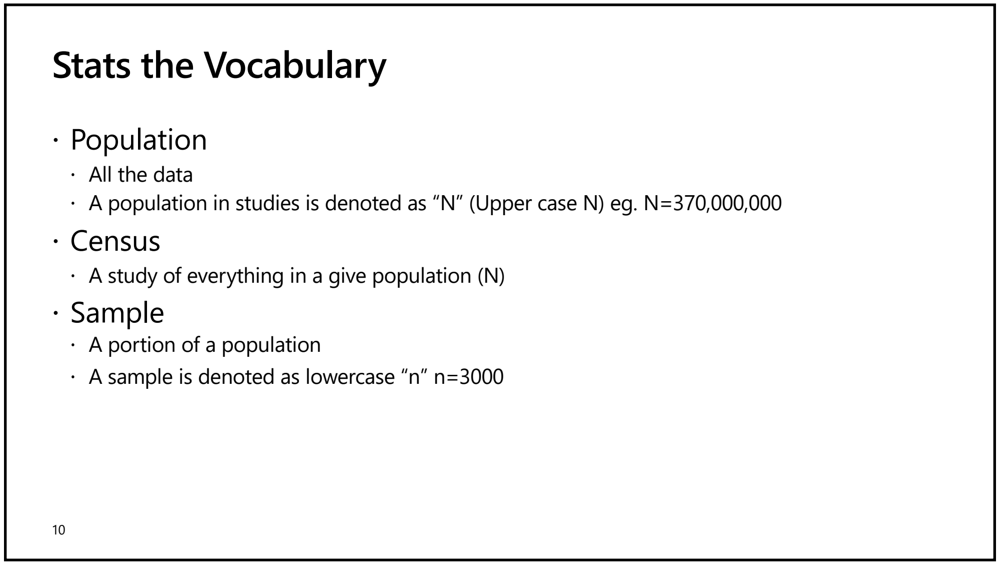


Stats the Vocabulary (Textbook)

- Case
  - Data folks know this as a row of data
- Variable
  - You may know this as a column of data
- Data Matrix
  - Is a table made of Cases and Variables or rows and columns

9

9



Stats the Vocabulary

- Population
  - All the data
  - A population in studies is denoted as “N” (Upper case N) eg. N=370,000,000
- Census
  - A study of everything in a give population (N)
- Sample
  - A portion of a population
  - A sample is denoted as lowercase ”n” n=3000

10

10

Stats the Vocabulary

- Case
  - In statistics is a single row of data
  - This can also be a single case for a patient
  - Imagine capturing 20 variables about a patient and putting them together in a row.
- Parameter
  - A numerical quantity that tells us something about a ***population***
  - E.g.. Quantity of specific ethnicity, number of high school graduates, proportion of singles.

11

11

Stats the Vocabulary - Qualitative Data

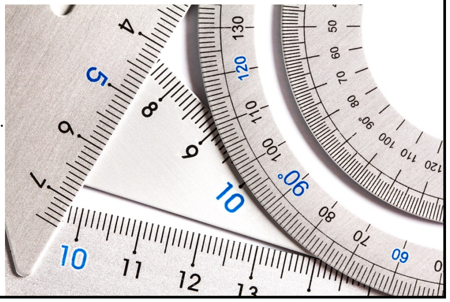
- Qualitative Data, Categorical or Category
  - Defined by Which or What
  - What color,
  - What dog breed,
  - What grade,
  - Which model of car,
  - What county Precinct number
  - Can be a number even an ordered number (ordinal) but would not make sense to do arithmetic against it.

12

12

Stats the Vocabulary - Quantitative Data

- Quantitative – Is always a number, can do math against it
  - Continuous – Measuring Data, asks how much
    - What is your height? 5’11”
    - What is your Weight?
    - What is the weight of your vehicle?
    - What is the MPG of your vehicle?
  - Discrete – Counting Data, asks How Many
    - How many people are on the bus? Never half a person.
    - How many cars in the driveway? Never half a car.
    - How may books do you own? One book per ISBN.



13

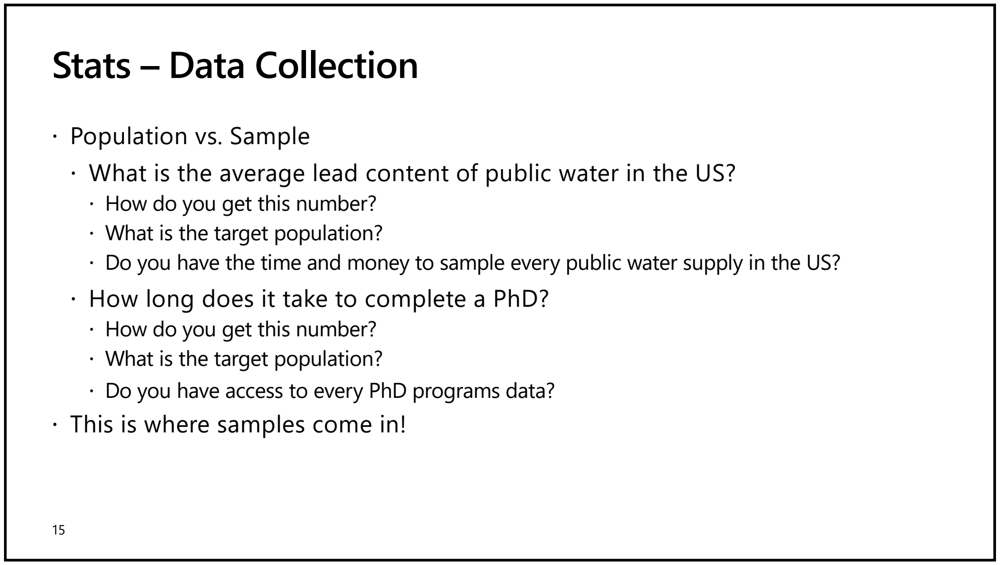
13

Stats the Vocabulary – Qual, Quan, Review

- Qualitative - Categorical, or a Category
- Quantitative – I can do math against it
  - Continuous – Measuring Data, asks how much
  - Discrete – Counting Data, asks How Many

14

14

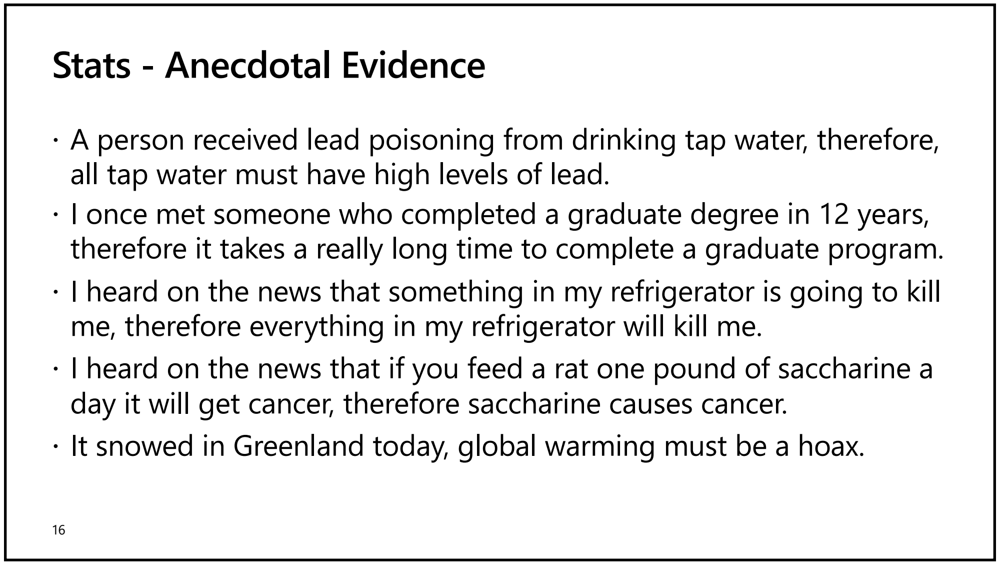


Stats – Data Collection

- Population vs. Sample
  - What is the average lead content of public water in the US?
    - How do you get this number?
    - What is the target population?
    - Do you have the time and money to sample every public water supply in the US?
  - How long does it take to complete a PhD?
    - How do you get this number?
    - What is the target population?
    - Do you have access to every PhD programs data?
- This is where samples come in!

15

15



Stats - Anecdotal Evidence

- A person received lead poisoning from drinking tap water, therefore, all tap water must have high levels of lead.
- I once met someone who completed a graduate degree in 12 years, therefore it takes a really long time to complete a graduate program.
- I heard on the news that something in my refrigerator is going to kill me, therefore everything in my refrigerator will kill me.
- I heard on the news that if you feed a rat one pound of saccharine a day it will get cancer, therefore saccharine causes cancer.
- It snowed in Greenland today, global warming must be a hoax.

16

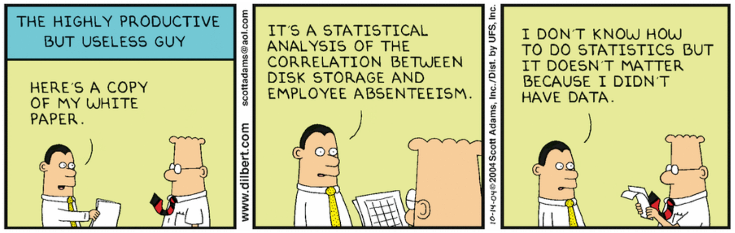
16

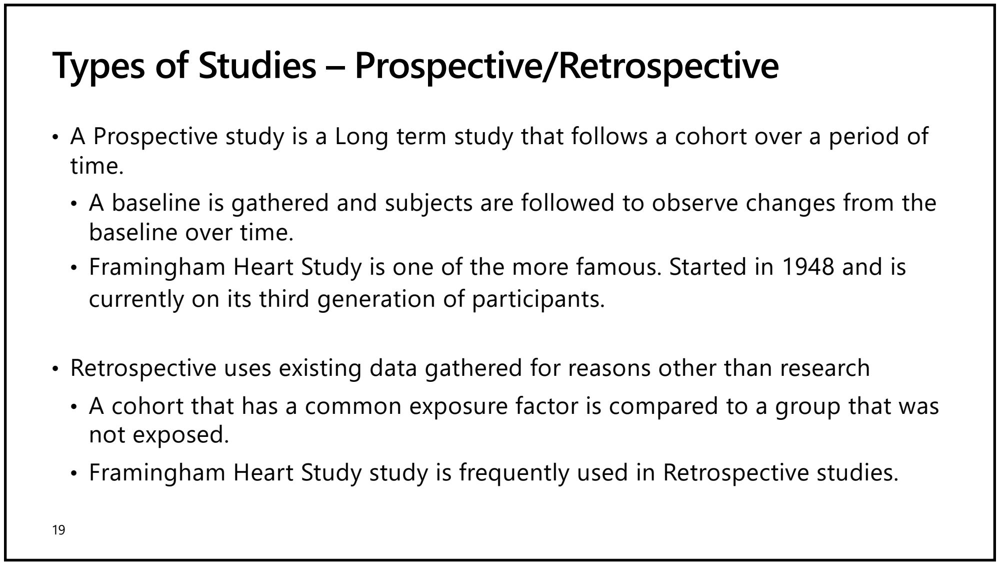
Types of Studies - Observational

- Observational - You are watching, or using data already collected
  - Data collected in such a way it does not interfere with the subjects response.
  - Survey data – Framingham Heart study, Harvard Happiness.
  - Cohort studies - follow a group of people for years, decades, generations.
  - Retrospective Study - Performing a study on data collected in the past.
  - **Causation cannot be implied from observational studies – EVER!**

Types of Studies - Experimental

- Experimental – Derive causal connection via treatment groups
  - Control and control groups– Literally have full control of factors in each group, treatment and placebo as well as other factors related to health, diet, exercise, etc…
  - Randomization – True randomization of treatment and control, if half the participants are of a specific ethnicity, it should be equal proportion among treatment groups.
  - Replication – The more cases the better


THE HIGHLY PRODUCTIVE BUT USELESS GUY
HERE'S A COPY OF MY WHITE PAPER.
scottadams@aol.com
www.dilbert.com
IT'S A STATISTICAL ANALYSIS OF THE CORRELATION BETWEEN DISK STORAGE AND EMPLOYEE ABSENTEEISM.
10-14-04 ©2004 Scott Adams, Inc./Dist. by UFS, Inc.
I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DIDN'T HAVE DATA.

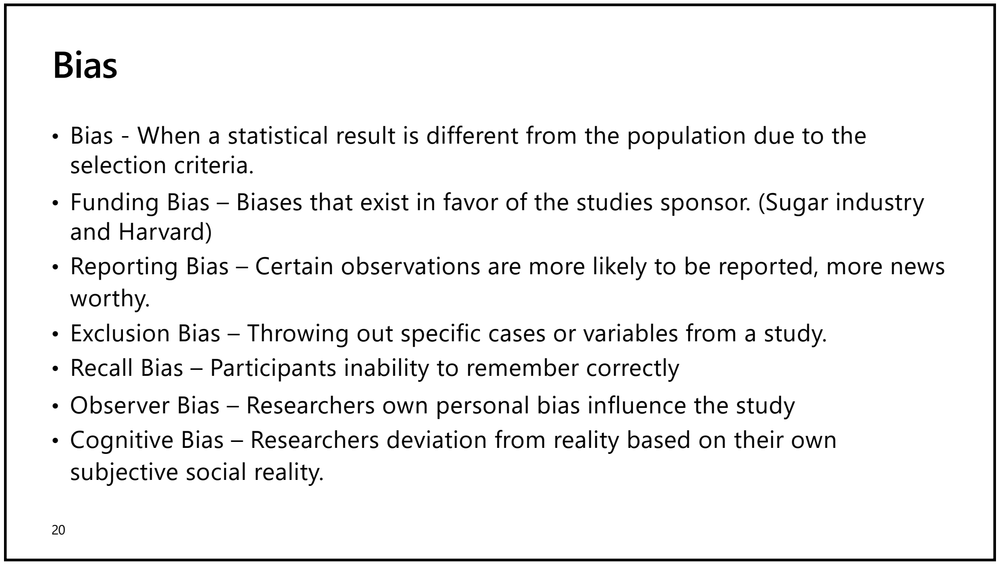


Types of Studies – Prospective/Retrospective

- A Prospective study is a Long term study that follows a cohort over a period of time.
  - A baseline is gathered and subjects are followed to observe changes from the baseline over time.
  - Framingham Heart Study is one of the more famous. Started in 1948 and is currently on its third generation of participants.
- Retrospective uses existing data gathered for reasons other than research
  - A cohort that has a common exposure factor is compared to a group that was not exposed.
  - Framingham Heart Study study is frequently used in Retrospective studies.

19

19



Bias

- Bias - When a statistical result is different from the population due to the selection criteria.
- Funding Bias – Biases that exist in favor of the studies sponsor. (Sugar industry and Harvard)
- Reporting Bias – Certain observations are more likely to be reported, more news worthy.
- Exclusion Bias – Throwing out specific cases or variables from a study.
- Recall Bias – Participants inability to remember correctly
- Observer Bias – Researchers own personal bias influence the study
- Cognitive Bias – Researchers deviation from reality based on their own subjective social reality.

20

20

Common Sampling Methods

- Simple Random Sampling
  - Randomly select X values from population
- Stratified Sampling
  - Group population by some factor then randomly select X random values from each group
- Cluster Sampling
  - Divided the population into X clusters, then randomly select X entire clusters.
- In the real world, a combination of all three is best.

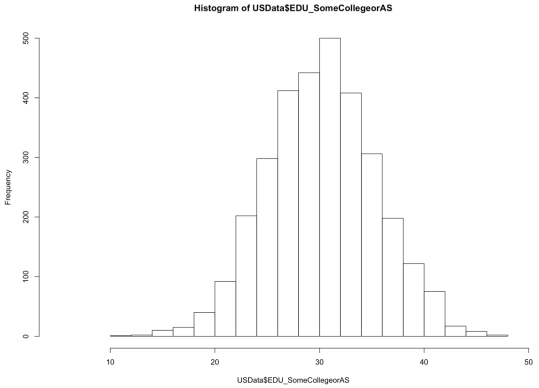
Mean Median, Mode, Quartiles

- Mean or average
  - The Sum of the values divided by the number of values summed
  - Typically known as the central value
- Median
  - The sorted middle value or average of the two middle values
- Mode
  - The number that shows up most frequently
- Inter Quartile
  - ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data, 25th , 50th, 75th.

Central Tendency

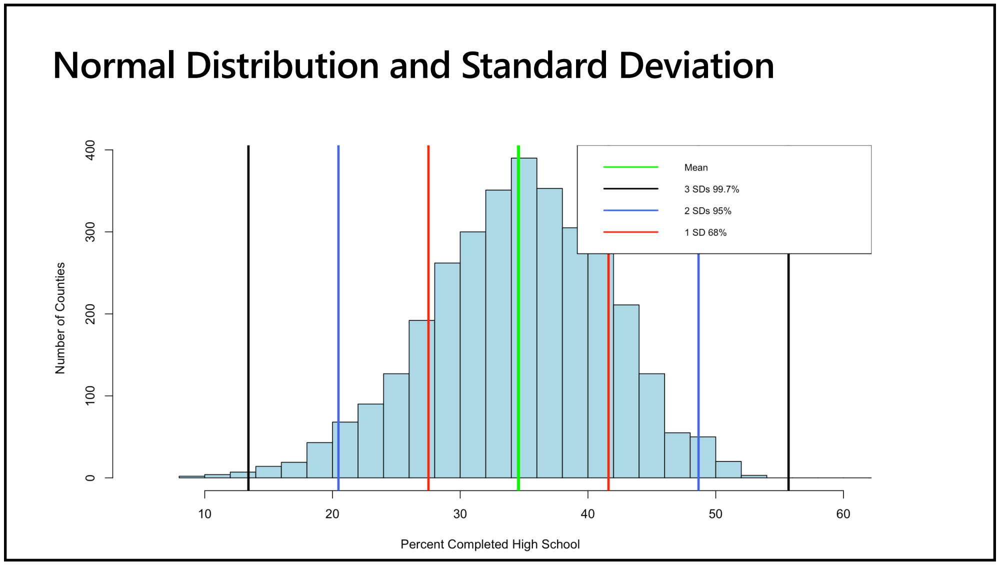
Generally speaking, data likes to be centered in a specific location.

The more data you have the more likely it is to be centered, or piled up.



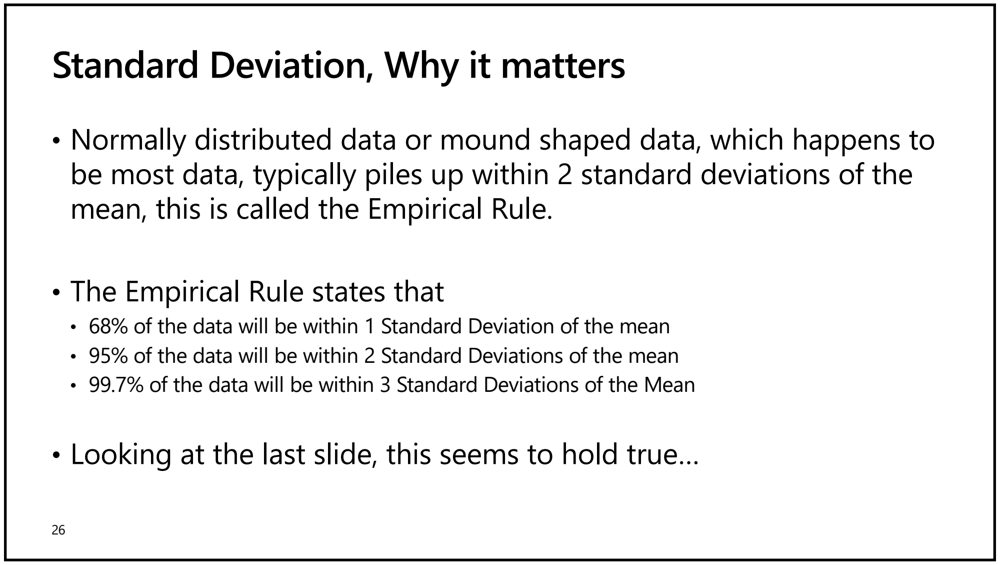
Standard Deviation and Variance

- Variance
  - Variance is the expected value of the squared deviation of a random variable for its mean. Looking at the formula is easier.
  - **Variance = sum((x – mean(x)) ^2) / (length(x)-1).**
- Standard Deviation
  - Measure of the variation of dispersion of the data. This has a nice easy formula as well, and it is based on variance.
  - **Standard Deviation** = **sqrt(sum((x – mean(x)) ^2) / (length(x)-1)).**
  - It’s the square root of the variance, how cool is that, you only need to know one formula! sqrt(VARIANCE)



Normal Distribution and Standard Deviation

25

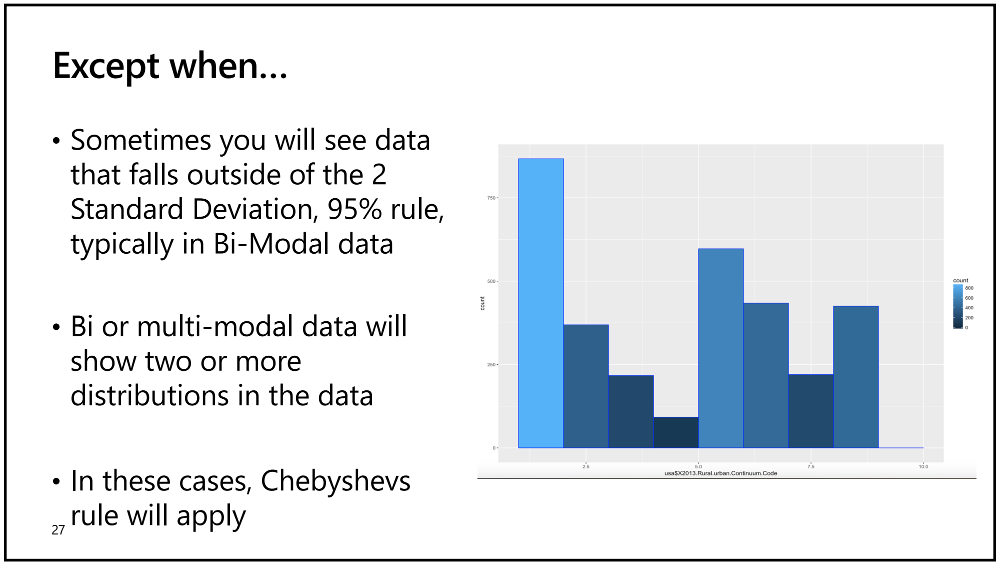


Standard Deviation, Why it matters

- Normally distributed data or mound shaped data, which happens to be most data, typically piles up within 2 standard deviations of the mean, this is called the Empirical Rule.
- The Empirical Rule states that
  - 68% of the data will be within 1 Standard Deviation of the mean
  - 95% of the data will be within 2 Standard Deviations of the mean
  - 99.7% of the data will be within 3 Standard Deviations of the Mean
- Looking at the last slide, this seems to hold true…

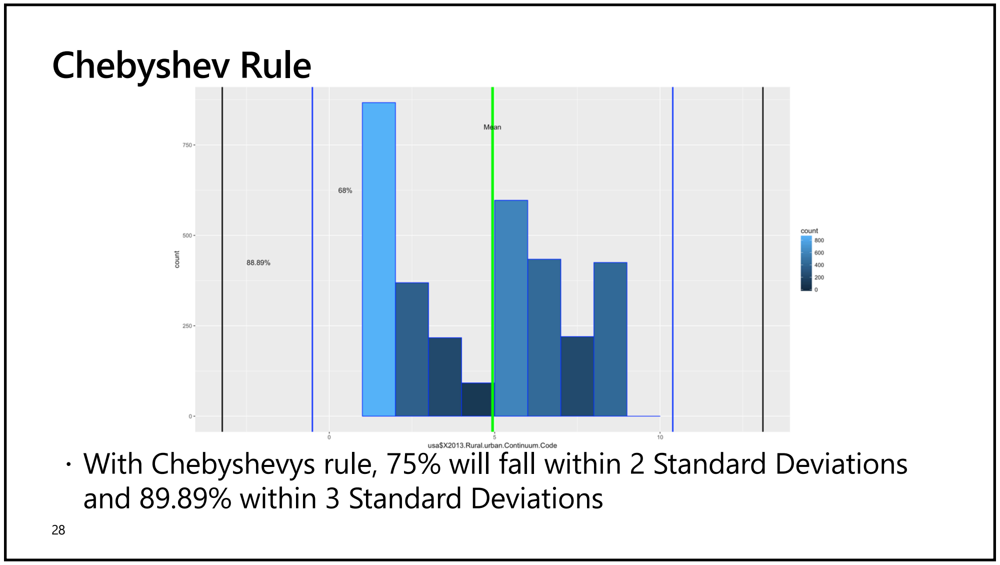
26

26



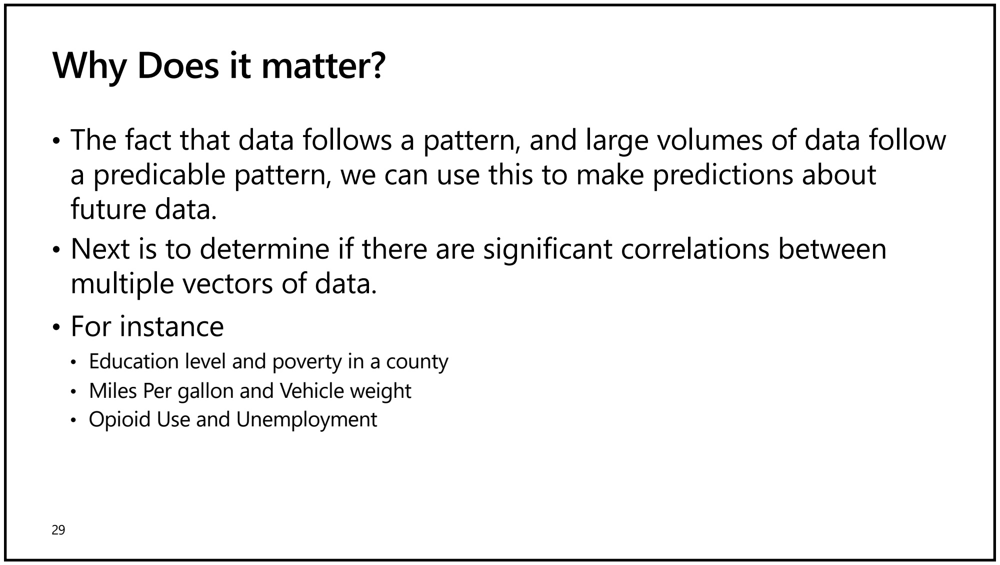
Except when…

- Sometimes you will see data that falls outside of the 2 Standard Deviation, 95% rule, typically in Bi-Modal data
- Bi or multi-modal data will show two or more distributions in the data
- In these cases, Chebyshevs rule will apply



Chebyshev Rule

- With Chebyshevys rule, 75% will fall within 2 Standard Deviations and 89.89% within 3 Standard Deviations



Why Does it matter?

- The fact that data follows a pattern, and large volumes of data follow a predicable pattern, we can use this to make predictions about future data.
- Next is to determine if there are significant correlations between multiple vectors of data.
- For instance
  - Education level and poverty in a county
  - Miles Per gallon and Vehicle weight
  - Opioid Use and Unemployment

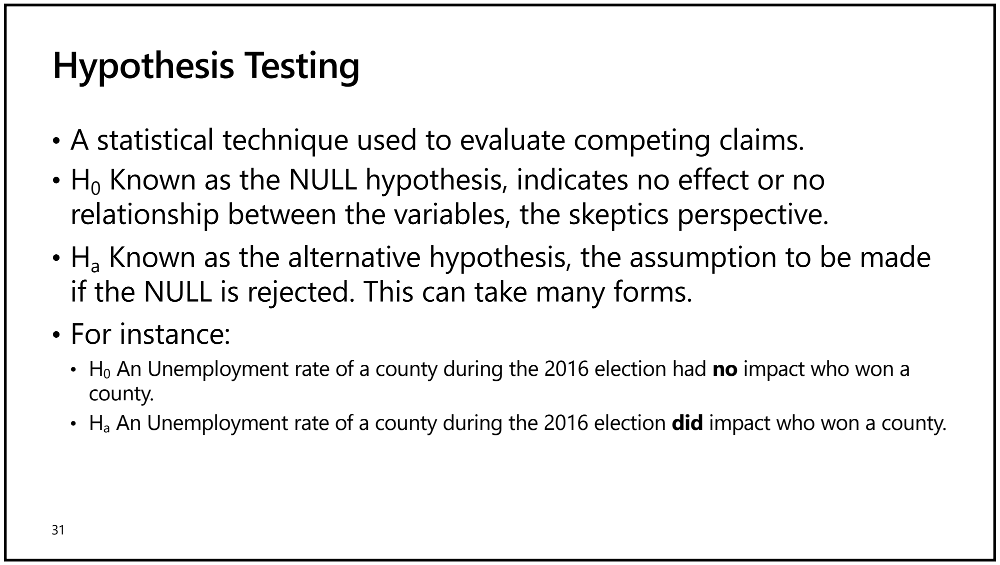
29

29

Microsoft

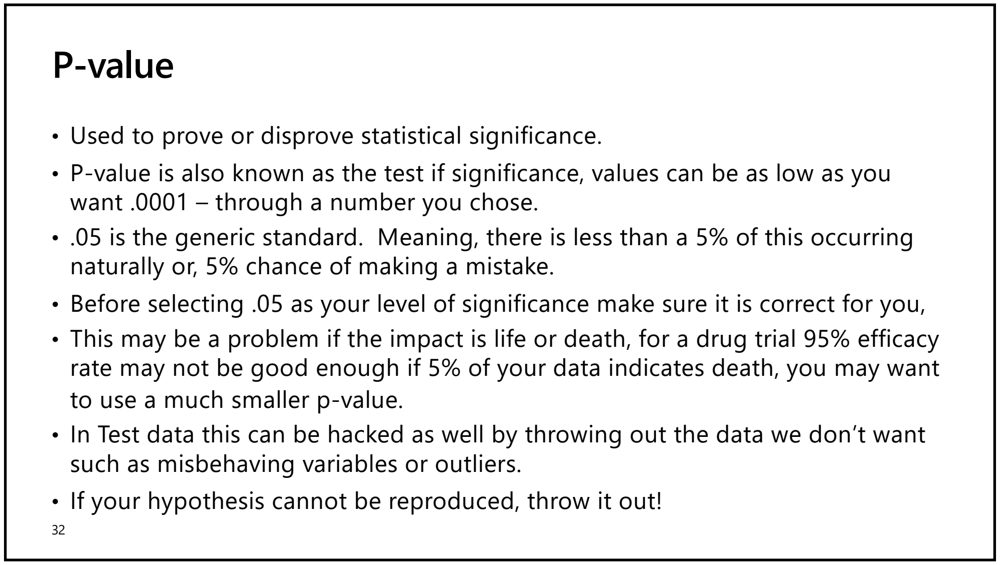
Insert 6 weeks of Statistics here, of which we will be skipping today!

30



Hypothesis Testing

- A statistical technique used to evaluate competing claims.
- H_0 Known as the NULL hypothesis, indicates no effect or no relationship between the variables, the skeptics perspective.
- H_a Known as the alternative hypothesis, the assumption to be made if the NULL is rejected. This can take many forms.
- For instance:
  - H_0 An Unemployment rate of a county during the 2016 election had **no** impact who won a county.
  - H_a An Unemployment rate of a county during the 2016 election **did** impact who won a county.



P-value

- Used to prove or disprove statistical significance.
- P-value is also known as the test if significance, values can be as low as you want .0001 – through a number you chose.
- .05 is the generic standard. Meaning, there is less than a 5% of this occurring naturally or, 5% chance of making a mistake.
- Before selecting .05 as your level of significance make sure it is correct for you,
- This may be a problem if the impact is life or death, for a drug trial 95% efficacy rate may not be good enough if 5% of your data indicates death, you may want to use a much smaller p-value.
- In Test data this can be hacked as well by throwing out the data we don’t want such as misbehaving variables or outliers.
- If your hypothesis cannot be reproduced, throw it out!

Q&A

This slide is required—do not delete—please read the notes for this slide, then delete this text box

33

33

Thank you for attending the MLADS Conference and helping to build a strong community

To find recordings, presentations, and other resources from the event, go to: http://aka.ms/november2022mlads

More information about the AI & ML Connected Community: http://aka.ms/aiml-cc

We want your feedback!

We read each evaluation and incorporate your comments into future MLADS events.
Both sessions and event evaluations will be available within the MyMLADS event portal during the conference dates—to access them, navigate to the Evaluations tab.

We request that session evaluations are submitted following each day of programming. Evaluations for both weeks of the conference will be open until **11:59 PM PT on Friday, June 17th**

34

34

35