Let the numbers talk!

Use Excel to learn how probability & statistics speak for your data

Rutgers Libraries - NB Data Science Workshop Series

Pratiksha Sharma
Oct 06, 2022

Fall 2022 Hours

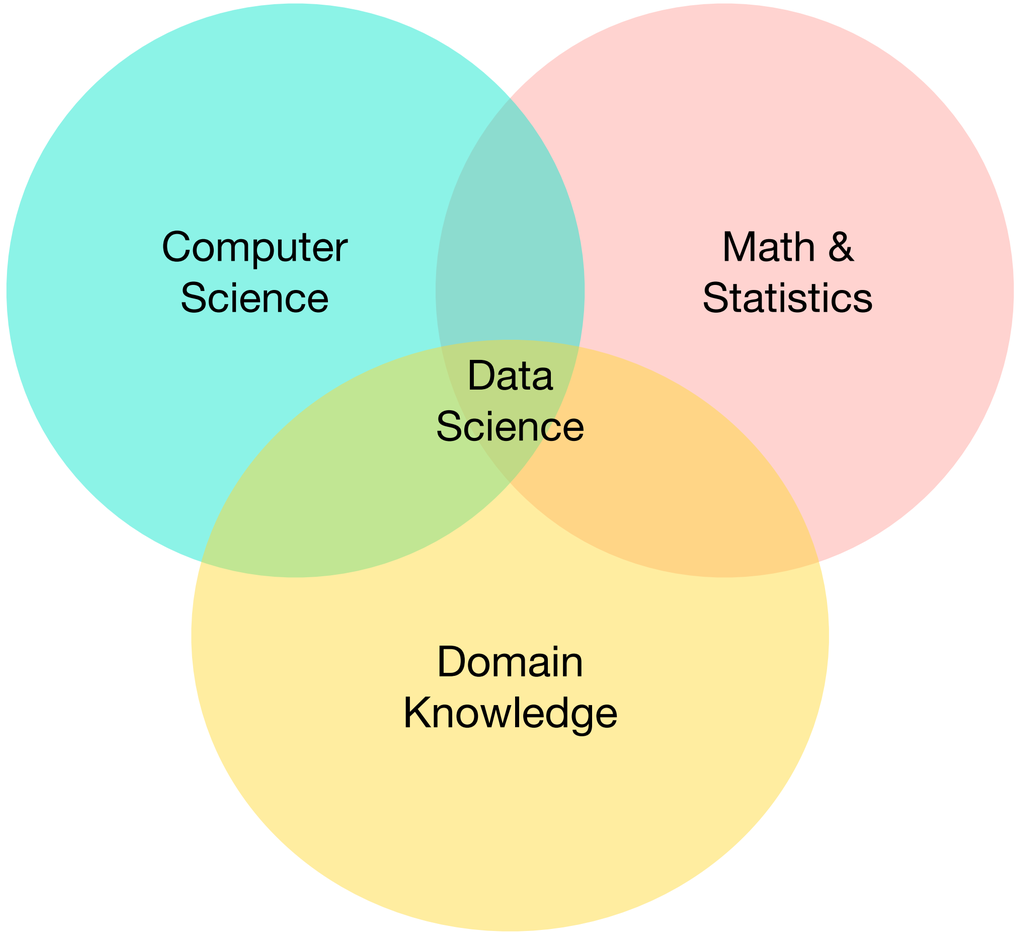
Pratiksha Sharma - Data Science Graduate Specialist

Email: pratiksha.sharma@rutgers.edu
Topics: Data Science, Tableau, Python, SQL & NoSQL Databases
Office Hours (by appointment):
Thursday 12:30 - 01:00 pm (on days when workshop ends at 12:30 pm)
Thursday 01:00 - 01:30 pm (on days when workshop ends at 01:00 pm)
General Consultation: Request an appointment via email
Location:
Zoom Meeting Link
Meeting ID: 926 5210 0393
Passcode: 772895

Data Science

What's all the fuss about?

- A combination of Maths & Statistics, Computer Science and Domain Knowledge.
- This workshop is about **Maths & Statistics!**
- You don't need to be an expert - but maths & statistics are the building blocks of nature :)



Motivation

Source: Bill Waterson | Universal Press Syndicate

Before we begin..

- We will be using MS Excel in this workshop
- Particularly some examples from the Analysis ToolPak: Load Analysis ToolPak for Excel
- You are encouraged to participate & follow along!
- Data would be provided in the Zoom chat & will be available later on with the workshop materials on: Rutgers Libguides Data Science Workshops

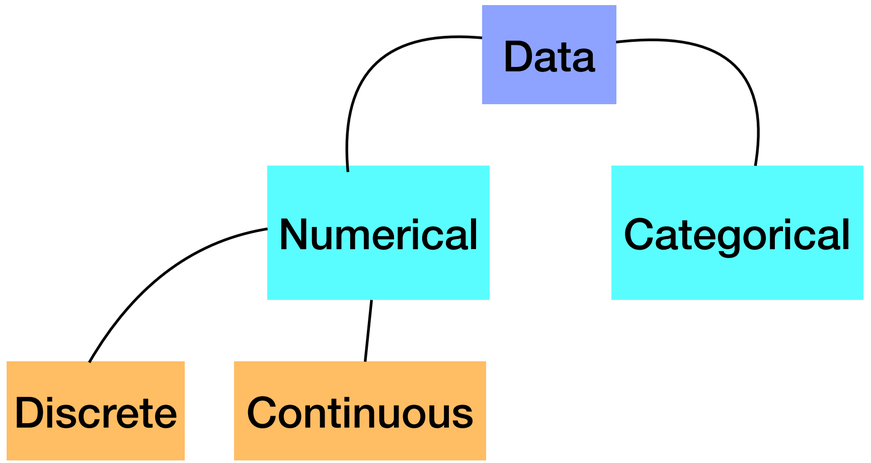
Probability & Statistics

We will cover..

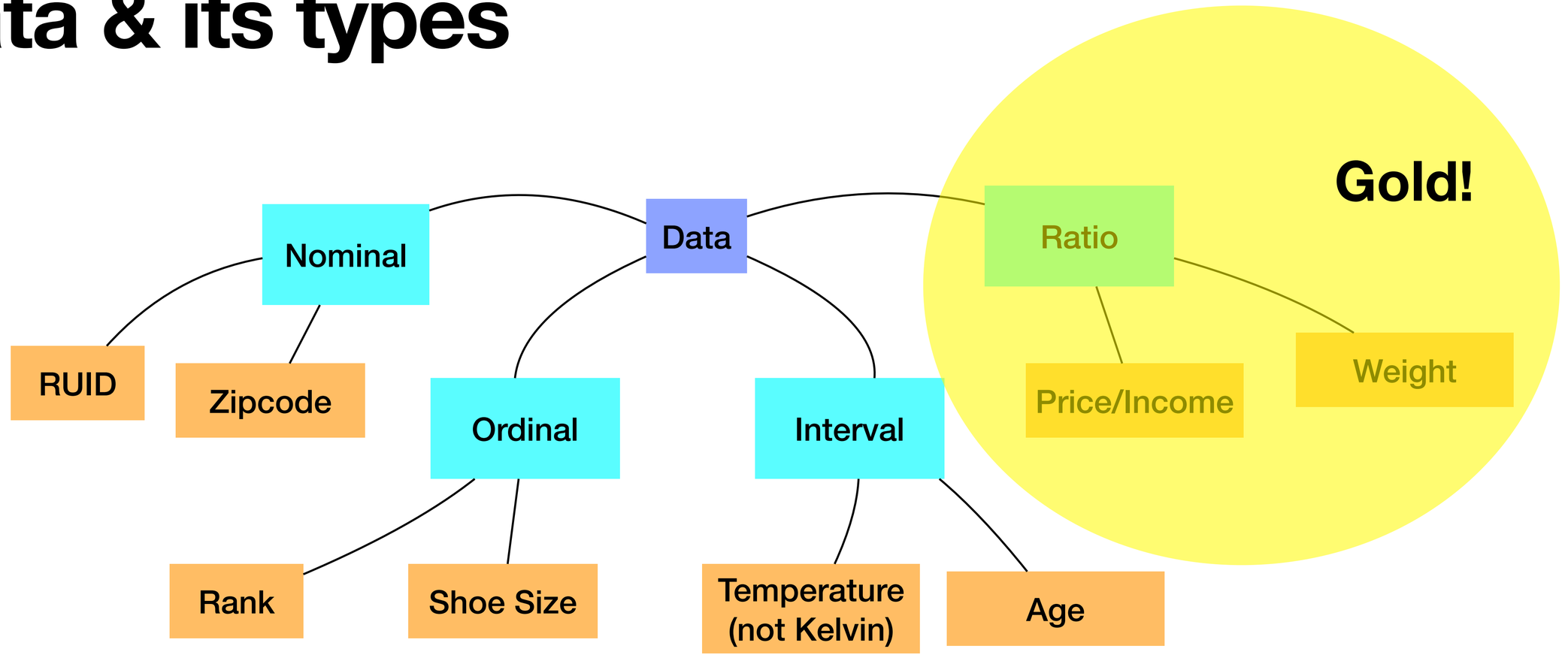
- Data & its types
- Back to the Basics: Probability
- Back to the Basics: Statistics
- Probability Distribution: Normal Distribution

Data & its types

- Data can be categorized into two types:



Data & its types



Back to the Basics: Probability

- Probability is the likelihood of the occurrence of an event

$$p(x) = \frac{number\ of\ outcome\ x}{total\ number\ of\ outcomes}$$

$$x = picking\ a\ red\ card$$

$$p(x) = \frac{total\ number\ of\ \boldsymbol{red}\ cards\ in\ the\ deck}{total\ number\ of\ cards\ in\ the\ deck}$$

$$p(x) = \frac{26}{52} = \frac{1}{2}$$

Back to the Basics: Probability

- What is the probability of getting an ace or a red card?

$A = picking\ an\ ace$
$B = picking\ a\ red\ card$

$$p(A) = \frac{total\ number\ of\ \boldsymbol{aces}\ in\ the\ deck}{total\ number\ of\ cards\ in\ the\ deck} = \frac{4}{52}$$

$$p(B) = \frac{total\ number\ of\ \boldsymbol{red}\ cards\ in\ the\ deck}{total\ number\ of\ cards\ in\ the\ deck} = \frac{26}{52}$$

$$p(A\ and\ B) = \frac{total\ number\ of\ \boldsymbol{red\ aces}\ in\ the\ deck}{total\ number\ of\ cards\ in\ the\ deck} = \frac{2}{52}$$

$$p(A\ or\ B) = p(A) + p(B) - p(A\ and\ B)$$

$$p(A\ or\ B) = \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$

Back to the Basics: Probability

- Repetition: What is the probability of picking a red card (with replacement), three times in a row?

$$x = picking\ a\ red\ card\ three\ times\ in\ a\ row$$

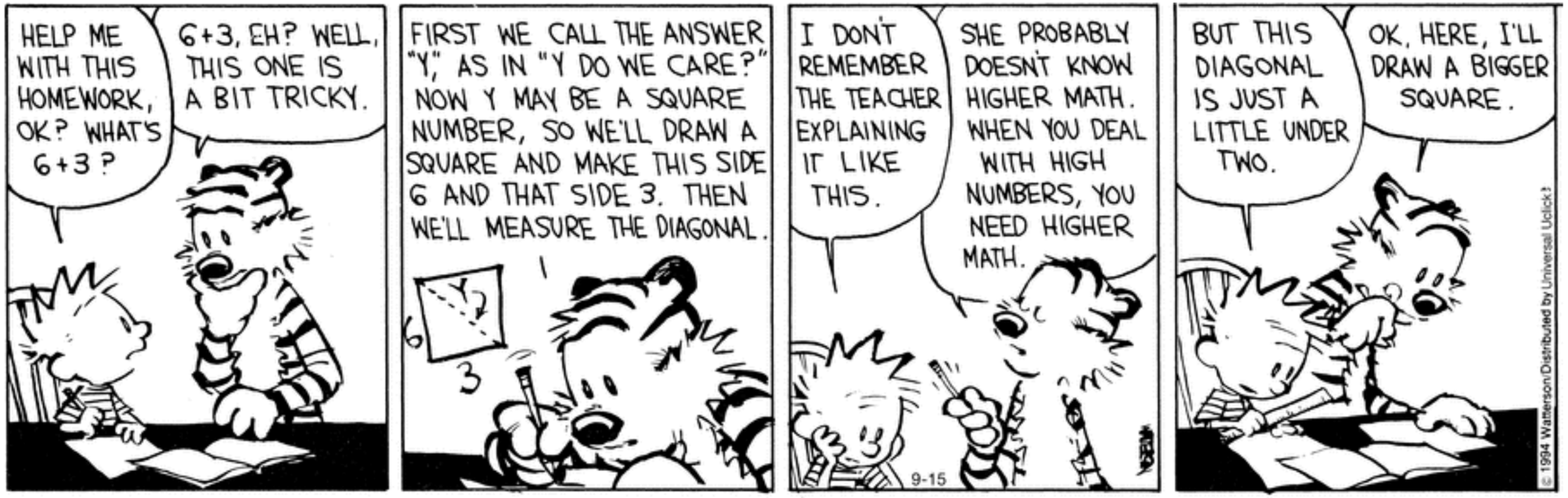
$$p(x) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

- Enter probability distributions!

Back to the Basics: Statistics

- Statistics is the practice of collecting & analyzing numerical data
- Let’s jump to MS Excel with our first example dataset!
- In this case, our population itself was the height of 30 students in a class, but what if our population is massive?

Motivation



Source: Bill Waterson | Universal Press Syndicate

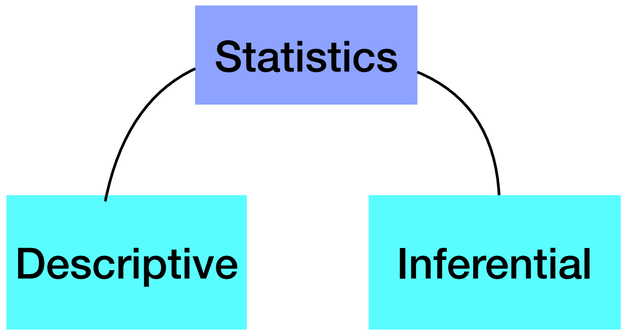
Back to the Basics: Statistics

- Statistics is the practice of collecting & analyzing numerical data
- Let’s jump to MS Excel with our first example dataset!
- In this case, our population itself was the height of 30 students in a class, but what if our population is massive?
- More appropriate: Statistics is the practice of collecting & analyzing ***massive*** numerical data
- The relationship between statistics & probability theory exists because you now want to approximate characteristics about your population data given you have some characteristics from sample data

Back to the Basics: Statistics

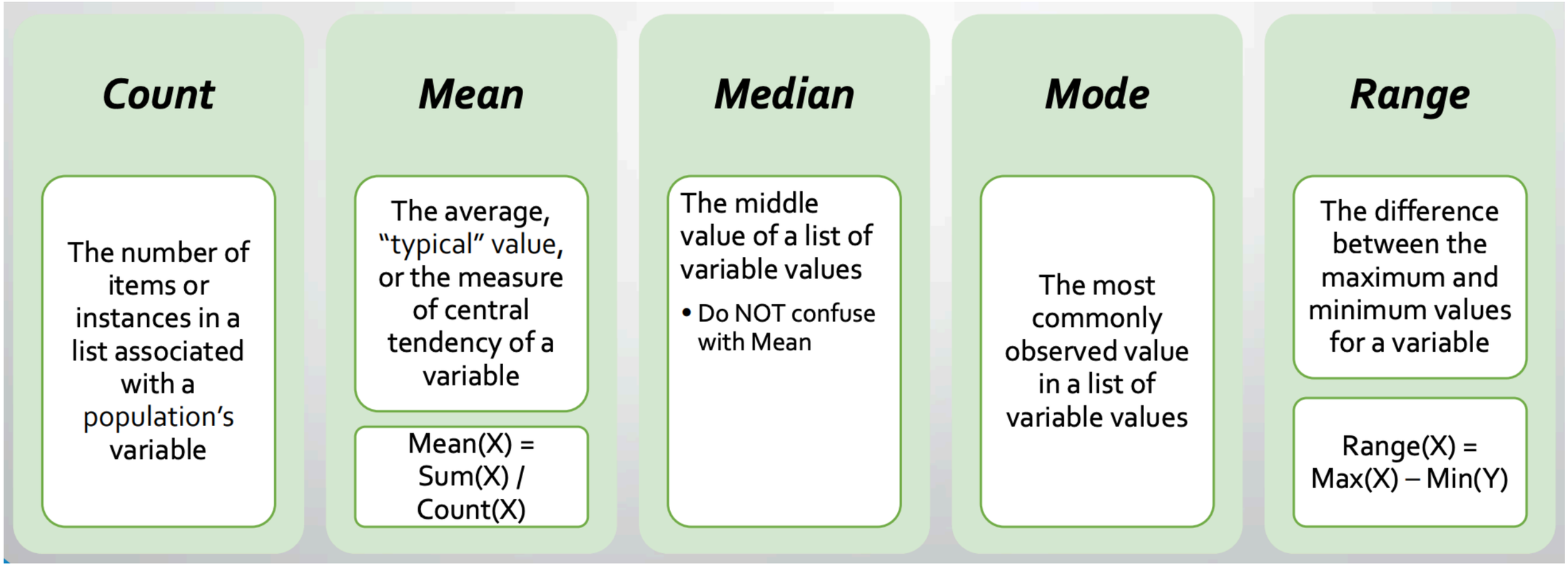
Before we jump into Probability Distributions

- Descriptive Statistics focus on describing the visible/visual characteristics of your data such as Measures of Central Tendency (Mean, Median, Mode, Range) or Measures of Variability/Dispersion (Standard Deviation, Variance, Interquartile Range) or Measures of Asymmetry.



Back to the Basics: Statistics

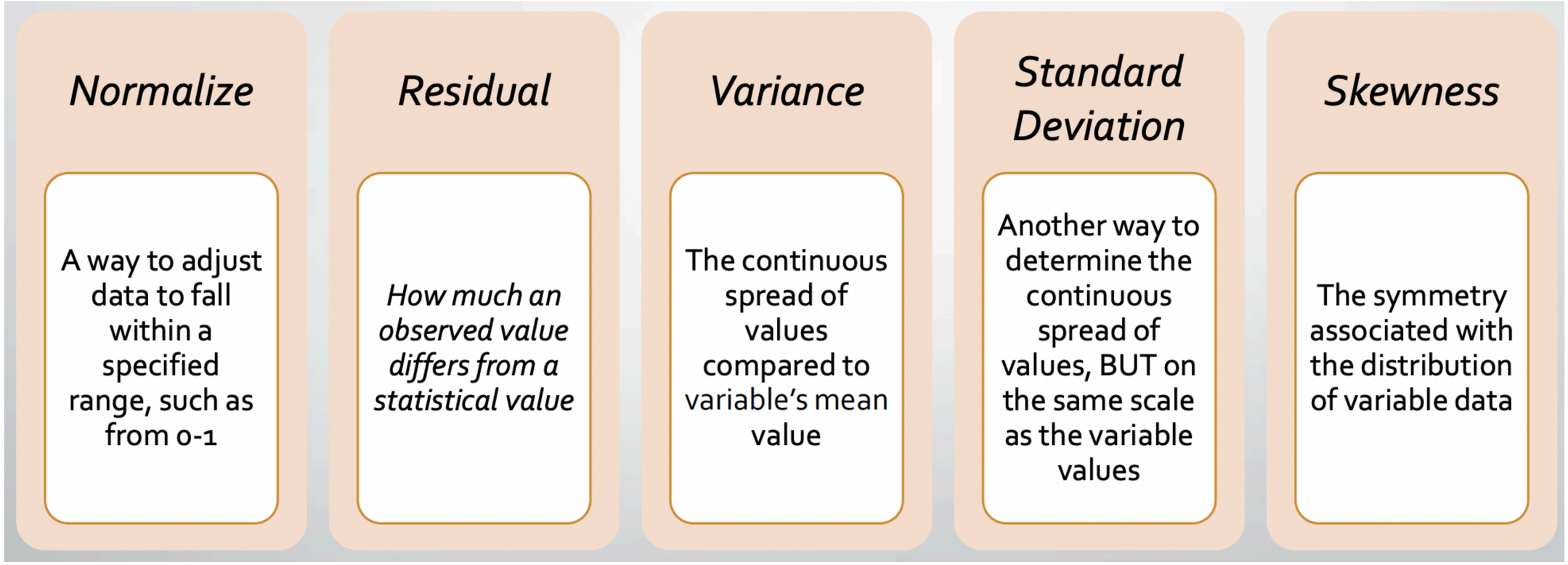
Measures of Central Tendency



Source: Robert Young | Rutgers NB Library

Back to the Basics: Statistics

Measures of Variability (Dispersion)



Source: Robert Young | Rutgers NB Library

Back to the Basics: Statistics

Measures of Asymmetry (Skewness)

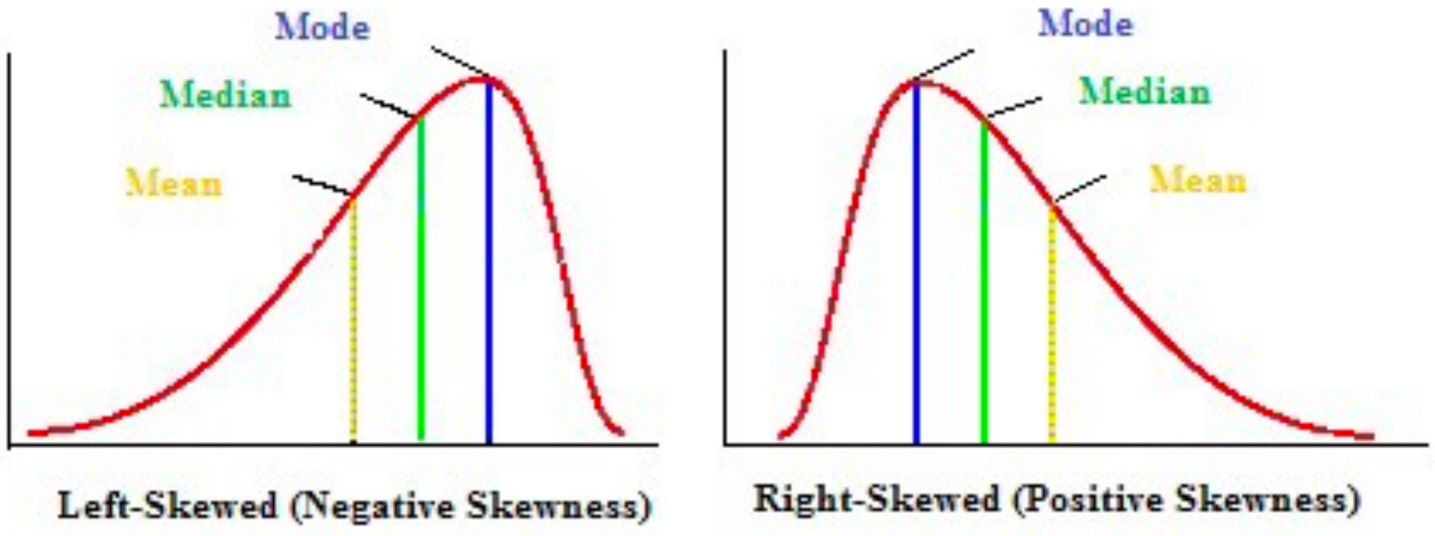
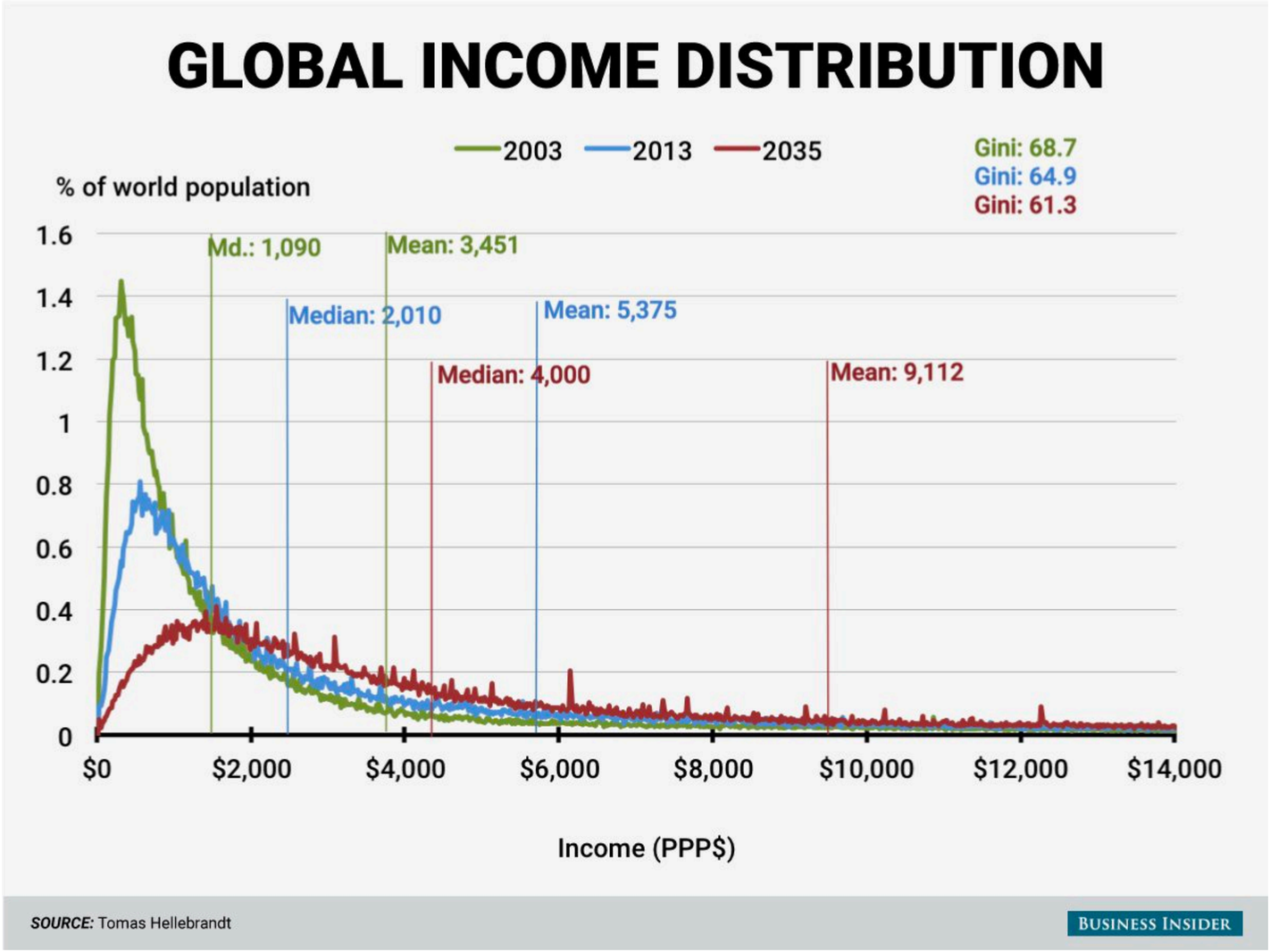


Image Source: https://www.statisticshowto.com/pearson-mode-skewness/

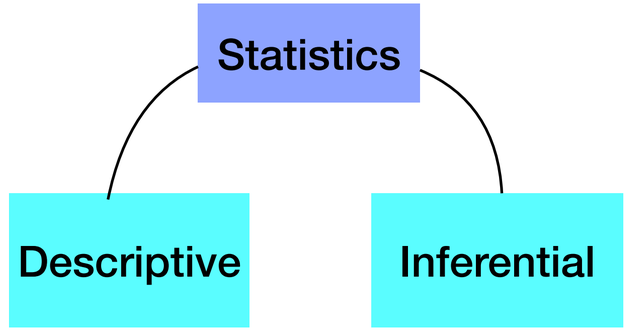
Back to the Basics: Statistics

Measures of Asymmetry (Skewness)



Back to the Basics: Statistics

Before we jump into Probability Distributions



- Inferential Statistics focus on making a generalization or prediction on the population based on a sample that is ***appropriately representative of the population***.
- Your sample needs to be unbiased i.e. random
- Inferential Statistics includes Hypothesis Testing, Confidence Intervals, Correlation & Regression Analysis

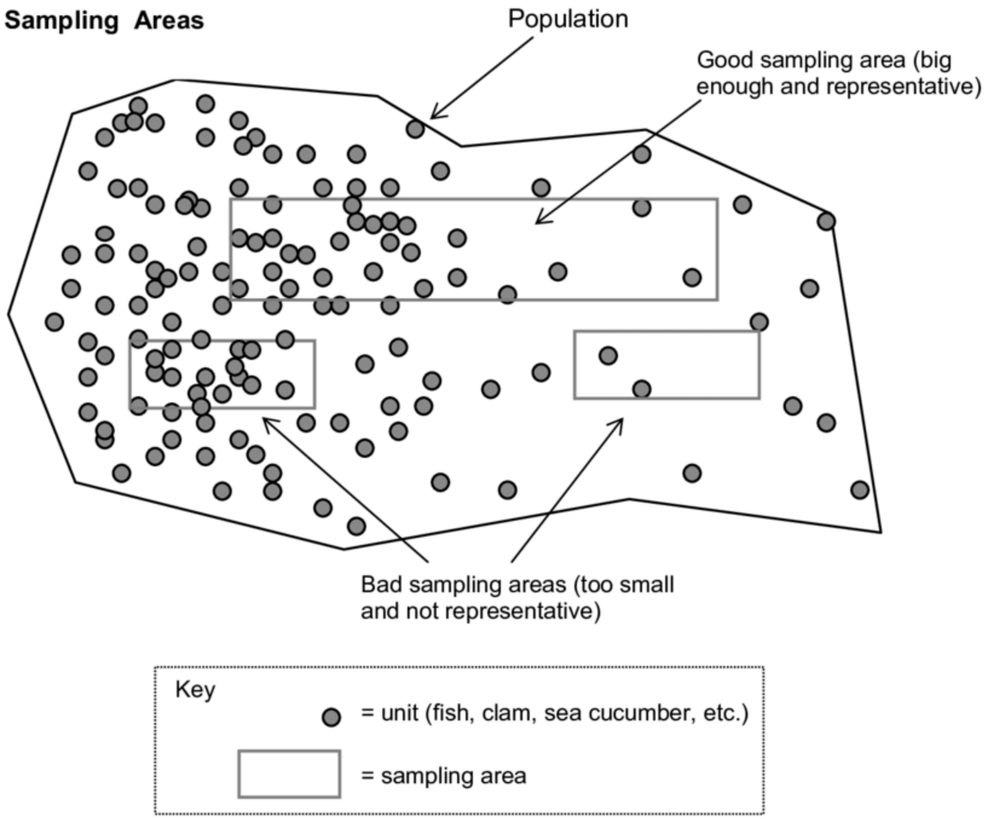
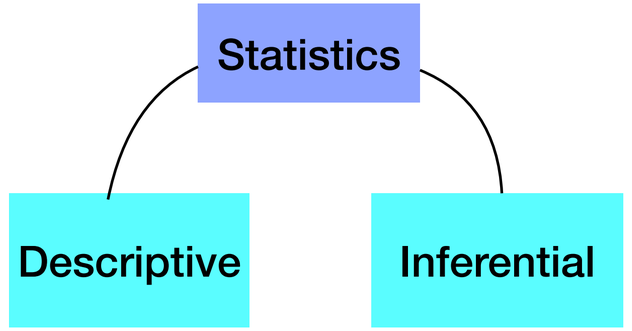


Image Source: Hugh Govan | researchgate.net

Back to the Basics: Statistics

Before we jump into Probability Distributions



- Hypothesis Testing: Eliminate the chance of getting a result by chance!
- Confidence Intervals: Remember that you are ***approximating*** the characteristics of a population. There will be error associated with your approximation & it is always better to say, “we are 95% confident that the mean height of students is between 5’3” & 5’8””

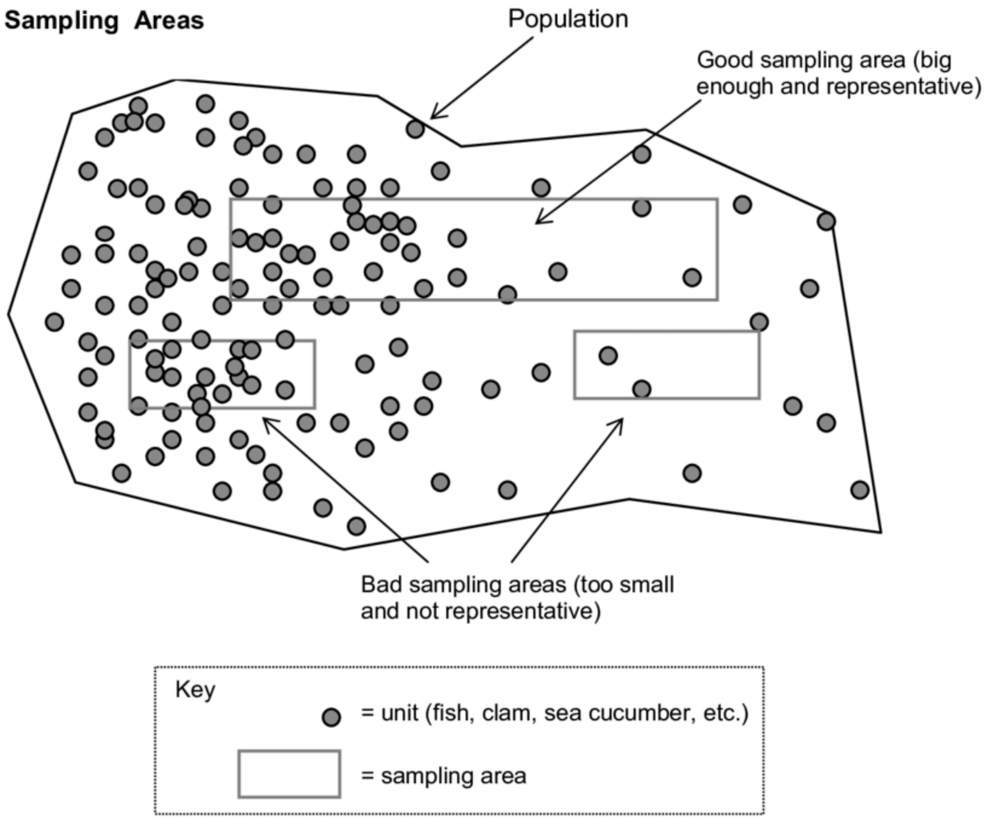
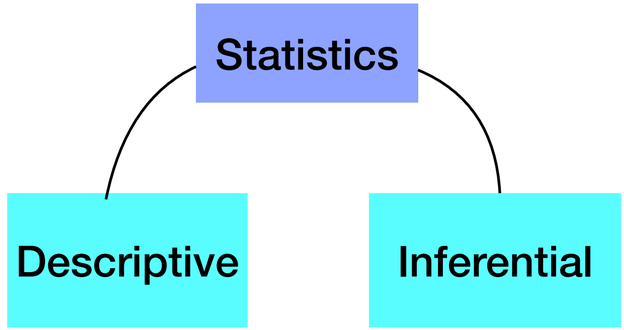


Image Source: Hugh Govan | researchgate.net

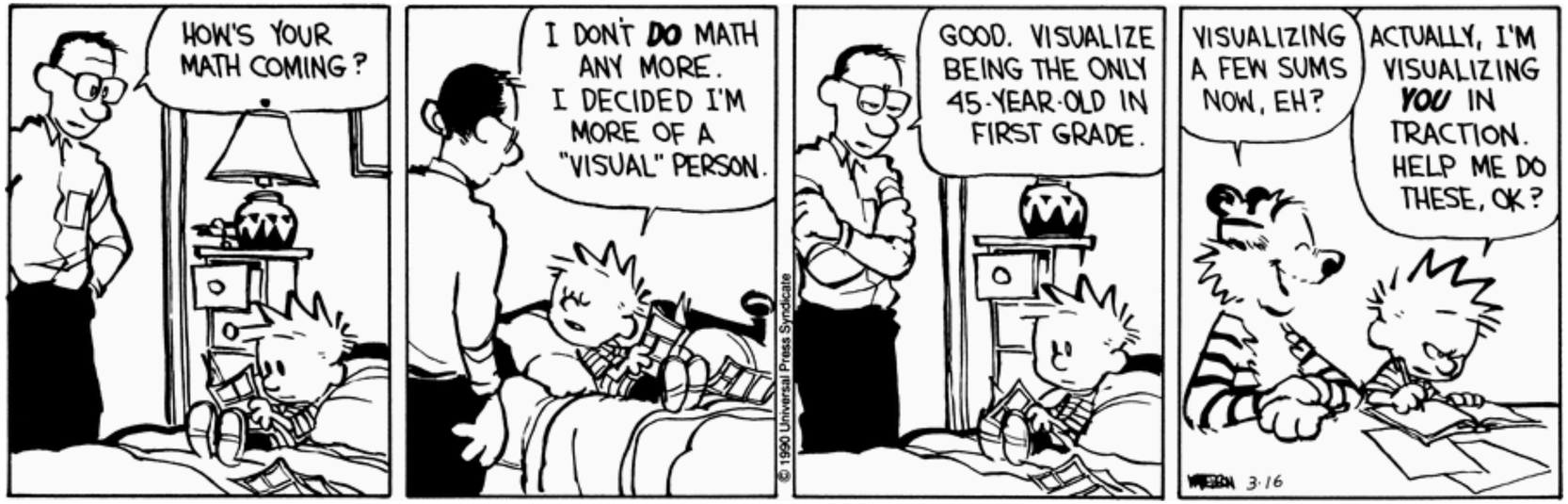
Back to the Basics: Statistics

Before we jump into Probability Distributions

- Correlation & Regression Analysis:
- Both study the relationship between two or more variables.
- Regression particularly studies cause & effect. For eg. Does eating a chocolate cause a rise in blood sugar level?
- Correlation measures the degree of association. For eg. Restaurant sales & sunburns are both high on sunny days near the beach. There is definitely no cause/effect relationship here.



Motivation



Source: Bill Waterson | Universal Press Syndicate

Probability Distribution

Normal Distribution

- The most commonly occurring distribution in nature
- Its a probability ‘bell’ curve - symmetric about the mean
- The values near to the mean are the most frequently occurring values
- The mean & the standard deviation are capable of explaining the underlying distribution of data

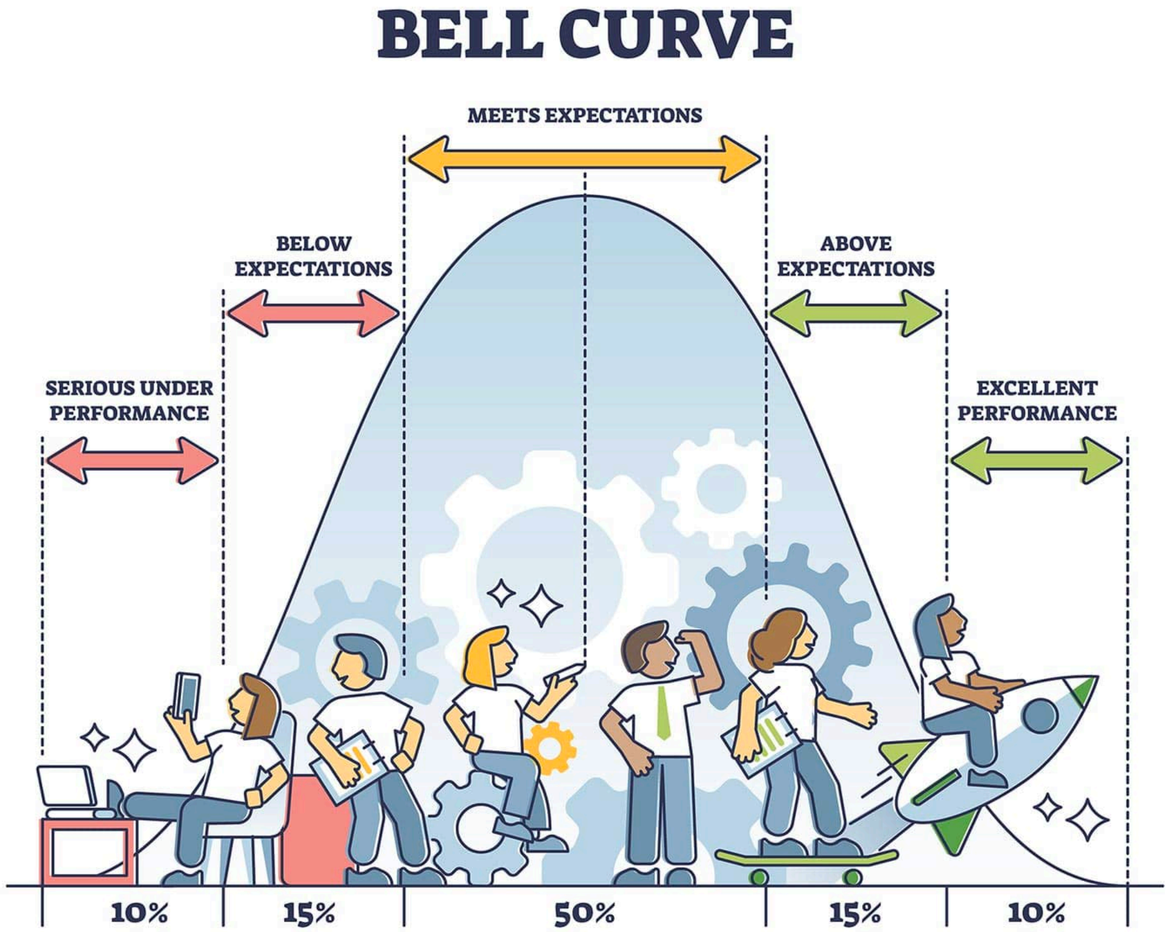


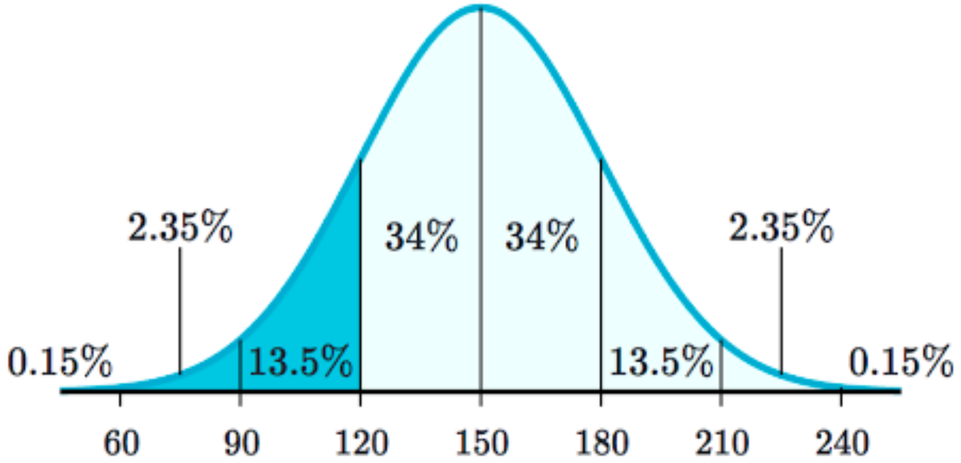
Image Source: Vikram Poddar | The Economic Times

Probability Distribution

Normal Distribution

- Look at data in terms of ‘distance’ from the mean
- ‘Distance’ can be measured in terms of standard deviations
- The values within 1 standard deviation are the most frequently occurring values
- Reinforce the idea that the mean & standard deviation are a capable of explaining an underlying normal distribution

Image Source: KhanAcademy.org



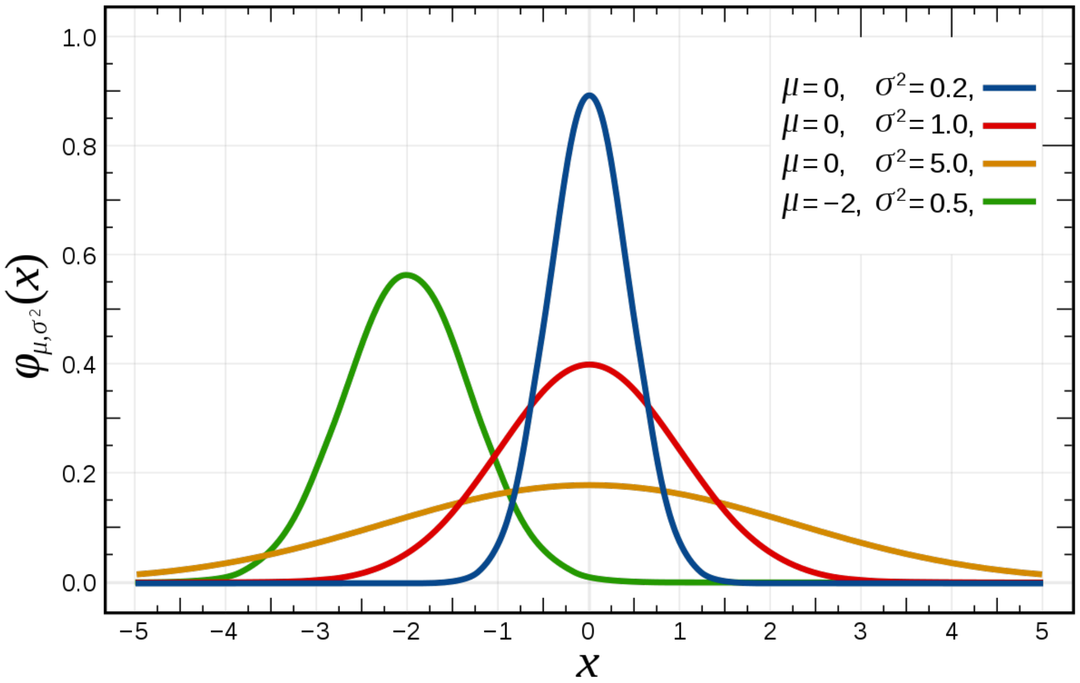


Image Source: Wikipedia

MS Excel Analysis ToolPak

We will cover..

- Descriptive Statistics (Summary Statistics)
- Sampling
- Correlation Analysis
- ANOVA (Single Factor)
- Next Time: Regression Analysis

Takeway

Image Source: https://www.statisticshowto.com/

Upcoming Workshops

https://libcal.rutgers.edu/nblworkshops

- “You will spend nearly 70% of your time doing this!" Organize & pre-process your Data : Oct 13
- The Power of Visual Storytelling: Learning Tableau Public : Oct 20
- “Make your computer work for you!" Learn how to use Python to program your tasks- Part 1 : Oct 27
- “Make your computer work for you!" Explore popular Data Science libraries in Python - Part 2 : Nov 03

Feedback Form

https://rutgers.libwizard.com/f/graduate_specialist_feedback