Data Analysis Boot Camp

29400M_7.0_2017

Introduction

What we will cover

- Data: How we get it, how we manage it
- Basics of Probability and Statistics
- Numerical and Visual Modelling
- Prediction
- Telling the Story

Data and Information

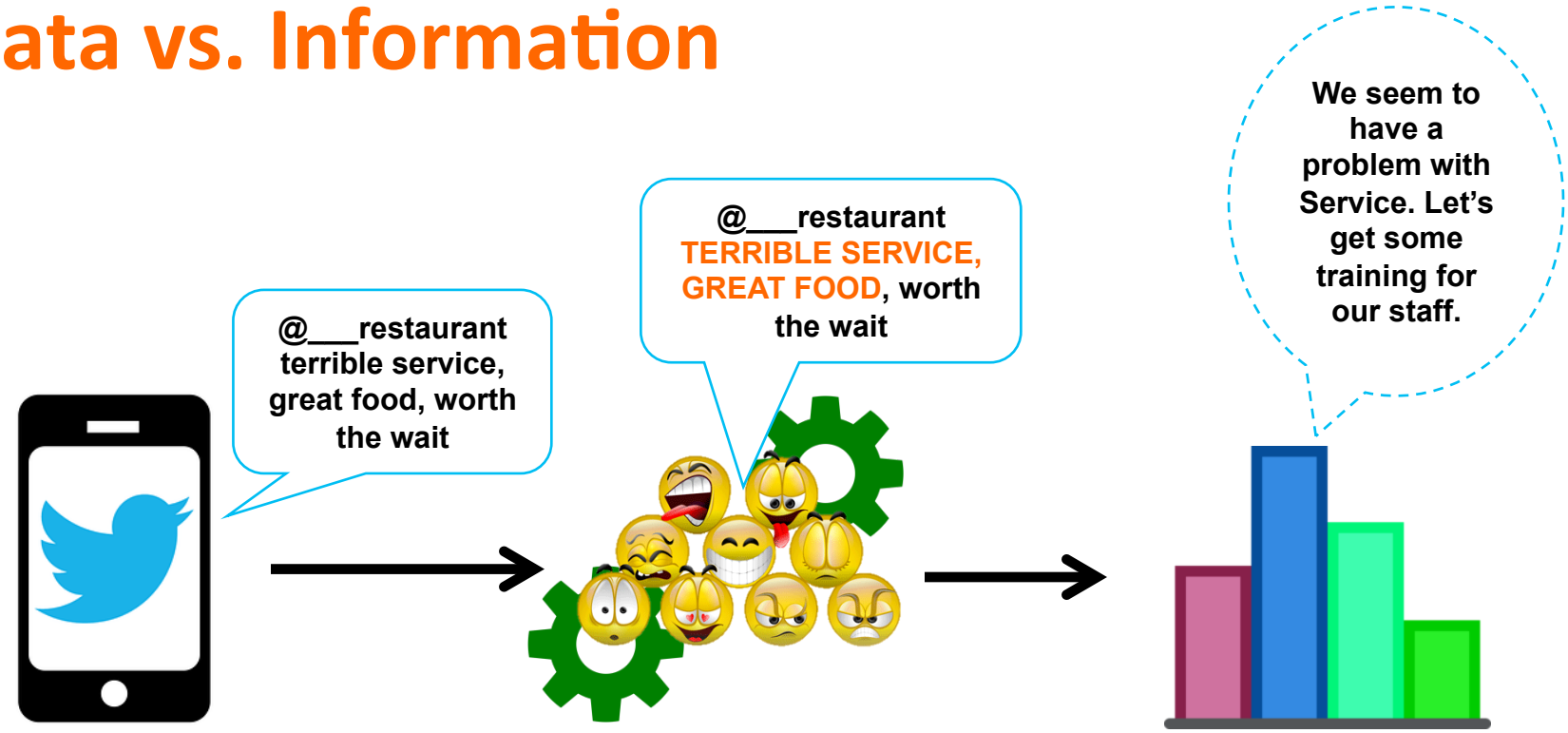
We discuss the difference between data and information in the enterprise and the world at large

Section 1

Data in the Real World

- It can come from just about anywhere
- Comes in two basic flavors: structured and unstructured
- It's messy
- It's non-standard
- Relationships can be tough
- Outliers exist
- …And it's growing rapidly

Data vs. Information



Tweets are a source of data

Sentiment Analysis uses analytics to identify relevant tidbits

The outcome of such analyses is actionable information

Data And Information

CLASSROOM WORK

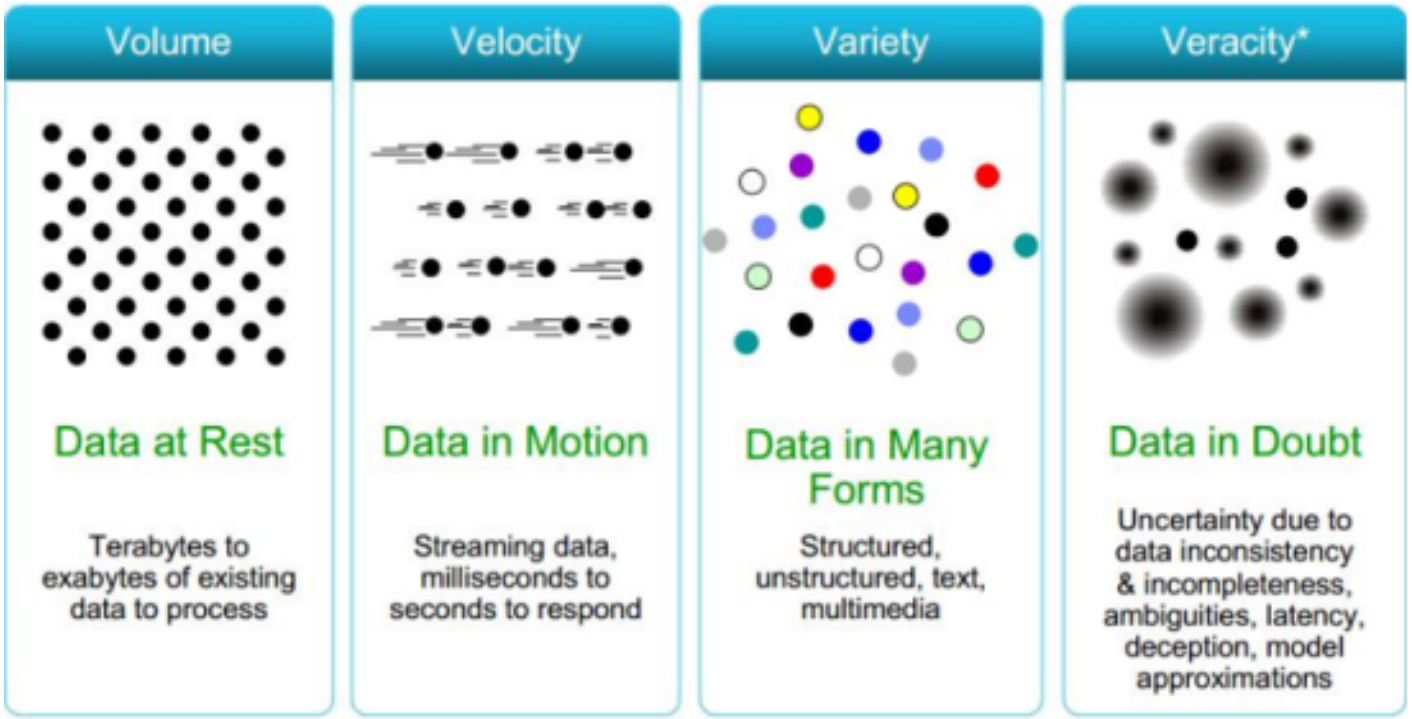
20 minutes

Use “sentiment.xlsx”

Column B-E: Add terrible, good, awful, bad
Add this formula
=IF(ISNUMBER(SEARCH(B$1,$A2)), 1,0)

Create pivot table, charts

The Many "Vs" of Data



Structured Data

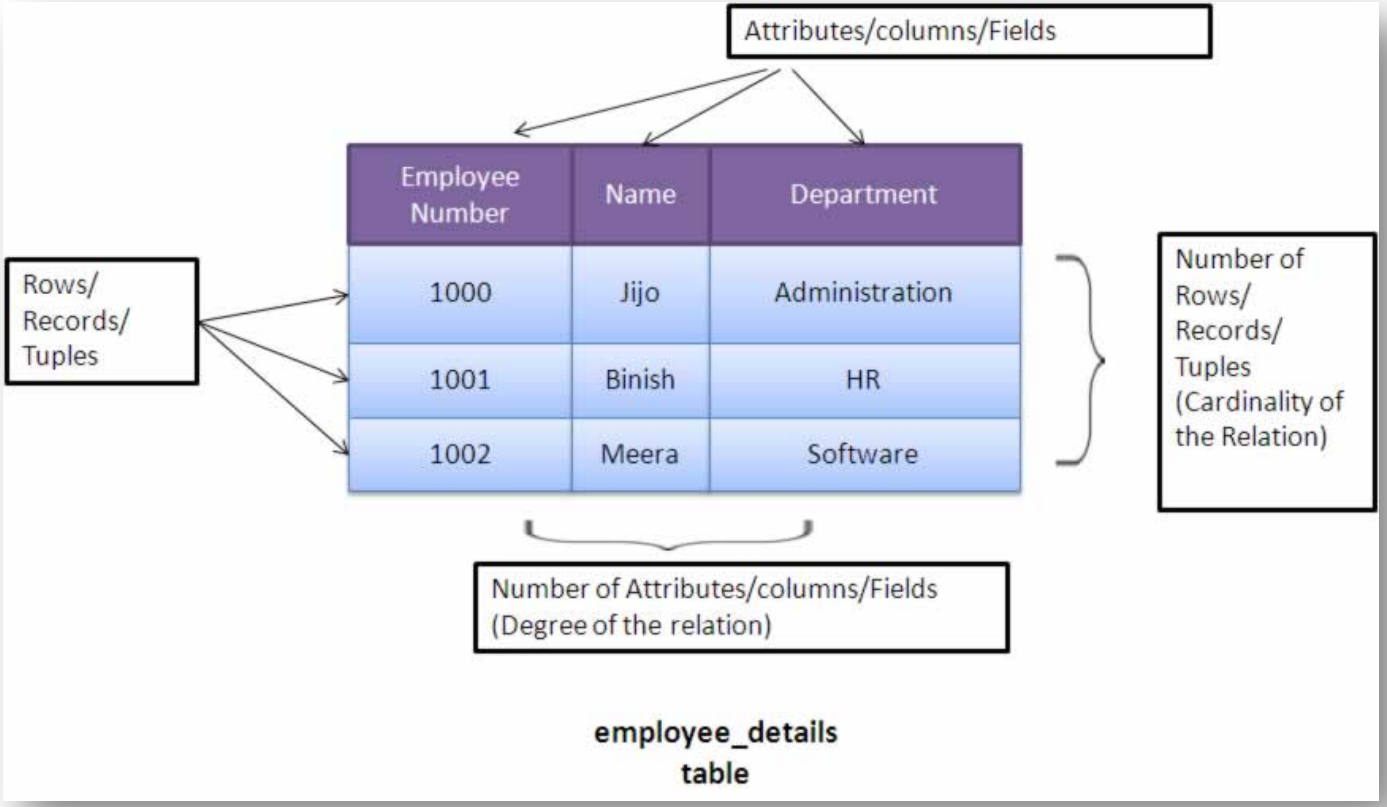
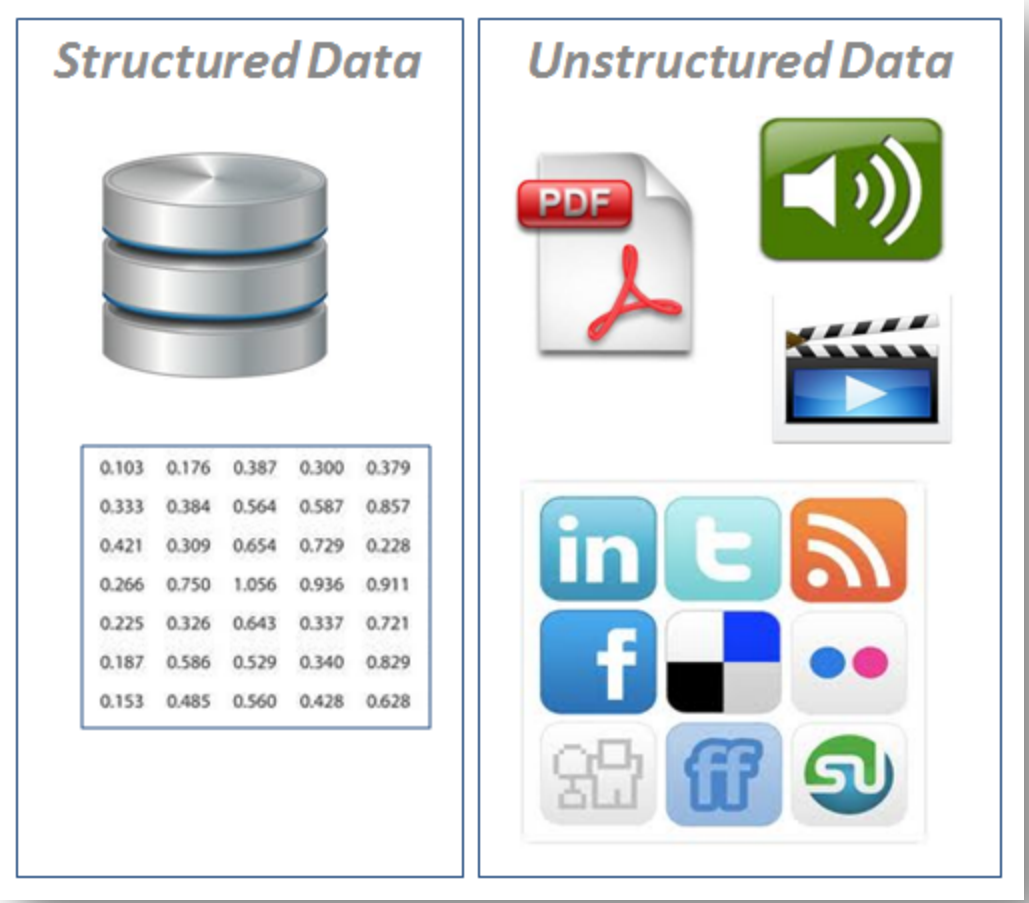


Image: Careerbless

Unstructured Data



…everything else.

Image: Infosys

Types of Data

Internal/External

- Internal data most relevant to tactical
- Great external sources exist, but often at a cost

Data Quality should be a priority

Data Quality

CLASSROOM WORK

30 minutes

Use “survey.xlsx”

Apply filter and check if you find same id with different names
For column G

(a) Average of rows from 15 to 900
(b) (Sum of rows from 15 to 900)/(900-15)
note the difference between a and b. Why ?

Starting on the right foot

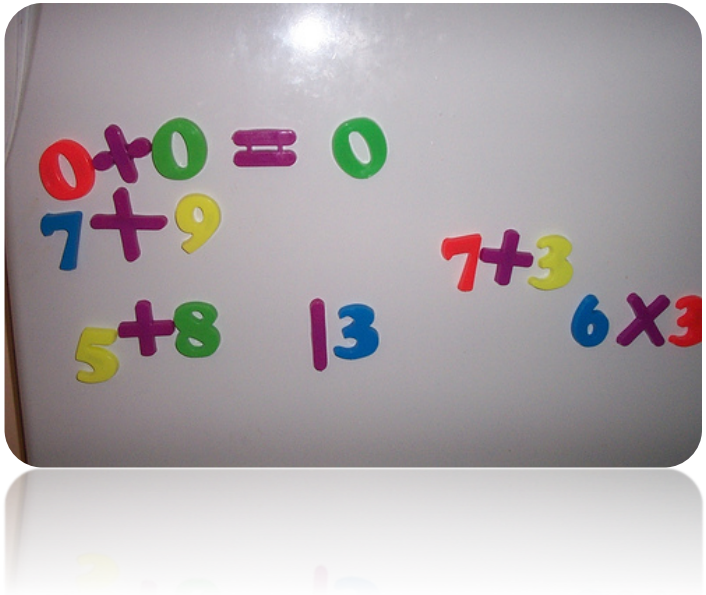
Cleansing means:

- Removing duplicates
- Having a single source of truth (SSOT) for core records
- Standardizing core fields Identifying sparsely populated fields
- Unique IDs

Starting on the right foot

The risks of skipping or neglecting data cleansing:

- **Inaccurate analysis**
- **Wasted time/money**
- **Inconsistent results**



Common Issues/Causes

Key Issues

- **Duplicates**
- **Misseplings**
- **Inconsistent formatting**
- **Inconsistent data population**
- **Missing/duplicated unique keys**
- **Special characters**

Why they happen

- **Manual Entry**
- **Old Business Rules Applied**
- **Software version changes**
- **Poor future planning**
- **Botched administrative efforts**
- **Poor interface/process design**

Data Quality

GROUP WORK

20 minutes

- When do you encounter dirty data
- What are the your issues with data quality
- How do you work around them (resolve)

How to fix some of these issues

- SSOT Services
  - Example: Buy cloud time on Azure to compare and sanitize your data records
- Review policies and business rules that affect data capture
- Implement projects to improve your data quality
- Implement unique identifiers
- Build a policy for quality “threshold” and error tolerance

Data Analysis Defined

We define some essential terms

Section 2

Data Analysis Defined

- **Data analysis:** **researching, organizing, and changing data in order to draw out useful information.**

Why do we analyze data?

- **Decision Support**
- **Trying to figure out what is going on**
  - Ferret out the context or story
  - So we can intervene, fix or improve something
  - Looking for causes
- **Prediction**
  - Using available data to project into the future
- **Monitoring and reporting**
- **SWOT (strengths, weaknesses, opportunities and threats)**

Data Analysis Mindset

- Flexibility
- Imagination
- Persistence
- Problem solving
- Diligence
- Patience

Hone Skills

It's the user not (just) the tool

- Hone deductive and inductive reasoning skills
- Will have to know some programming languages
- Interpretation skills are key
  - Do you know what you are seeing?
  - How will you minimize bias
  - Who will you collaborate with to ensure accuracy?
  - How will you share your findings?

Data Analysis Steps

1. Define your objectives first
2. Determine the levers (metrics are key)
3. Collect the data (ensure an adequate sample)
4. Clean the data
5. Model the data (test, test, test)
6. Identify repeatable processes

Data Analysis Defined

Data Analysis

- Considered a subcomponent of analytics
- Often the goal of analysis is more exploratory than required for decision making

Analytics

- The systematic computational analysis of data
- Describes the data analysis that clusters, segments, scores, and predicts what is likely to happen next
- The outcome should be an implementable decision based on the data

Common Terms

- **Population:** the pool from which the statistical sample is pulled
- **Sample:** the set of collected data gathered from a statistical population
- **Average:** the number that measures central tendency of a given set of numbers
- **Correlation:** describes the strength and direction of a relationship between variables
- **Distribution:** description of the relative number of times an outcome will occur in a number of trials
- **Observation:** an element of a population or data set

Two domains of data analysis

- **Descriptive Statistics**
  - **When you have all of something**
- **Inferential Statistics**
  - **When you have a *representative* sample of a population**
  - **You use the characteristics of the sample to make probabilistic estimates of the parameters of a population**
  - **Formal "Hypothesis Testing"**

There is some overlap in analytical techniques in these two domains

Our focus here is Descriptive Statistics.

- **Data with characteristics analyzed with these tools are more common in business applications**
- **The word "hypothesis" is used in this class not as formal "Hypothesis Testing" as found in Inferential Statistics but as a part of exploratory data analysis.**
  - **Hypotheses are often formed and abandoned as one approaches a better understanding of the data and the reality that generated them.**

Descriptive Statistics

For example, if given a set of annual starting salaries:

- **Mean/Median/Mode of salaries**
- **Salary averages**
- **Variation in salaries**

- ***Descriptive statistics emphasizes measures of central tendency.***

Inferential Statistics

Inferential statistics allows for generalizations about the population using samples.

Limitations:

1. **Providing data about population that may not be fully measured**
2. **Some tests require educated guesses based on theory to run tests**

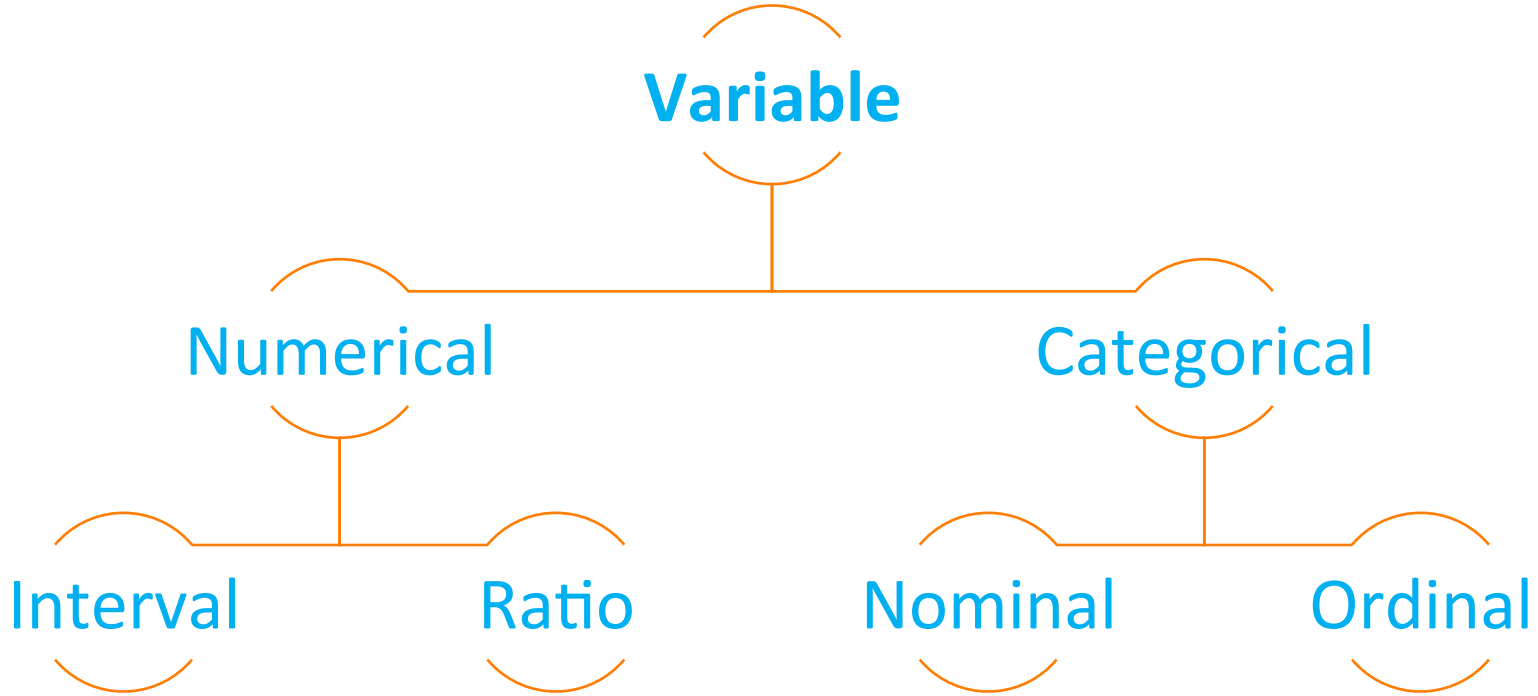
- ***Methods are focused on parameters and testing statistical hypotheses.***

Types of Variables

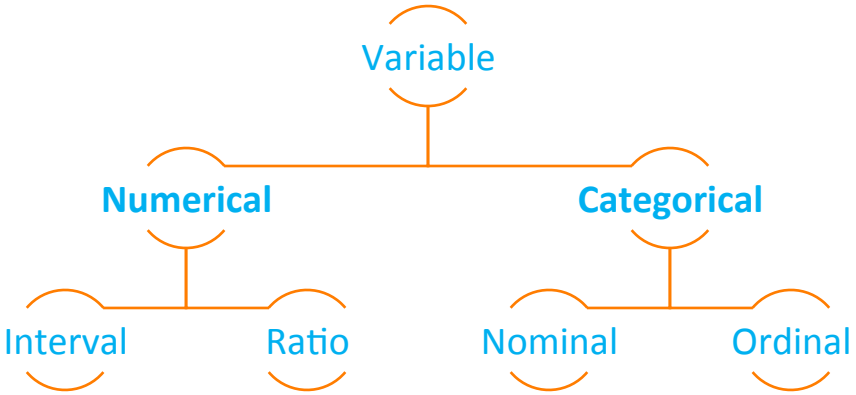
We look at what kinds of variables are involved in data analysis

Section 3

Definitions: Variables



Definitions: Variables

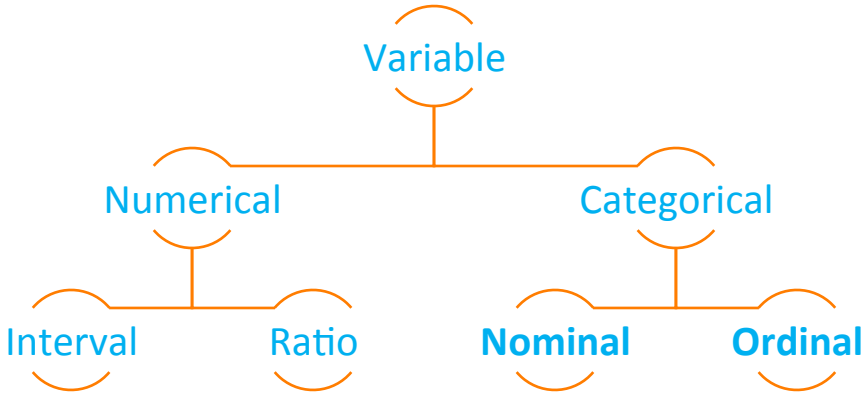


Categorical: categorizes or describes an aspect of a population

(E.g. blue, green, yellow)

Numerical: quantifies an element of a population (E.g. 1, 2, 3)

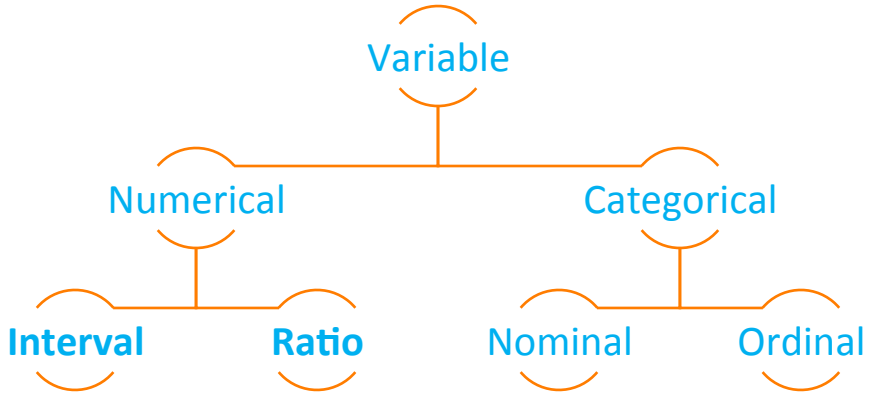
Definitions: Variables



Nominal: Categorical values that doesn't have ordering. E.g. Blue, Green, Yellow

Ordinal: categorical values that has ordering. E.g. low, medium, high

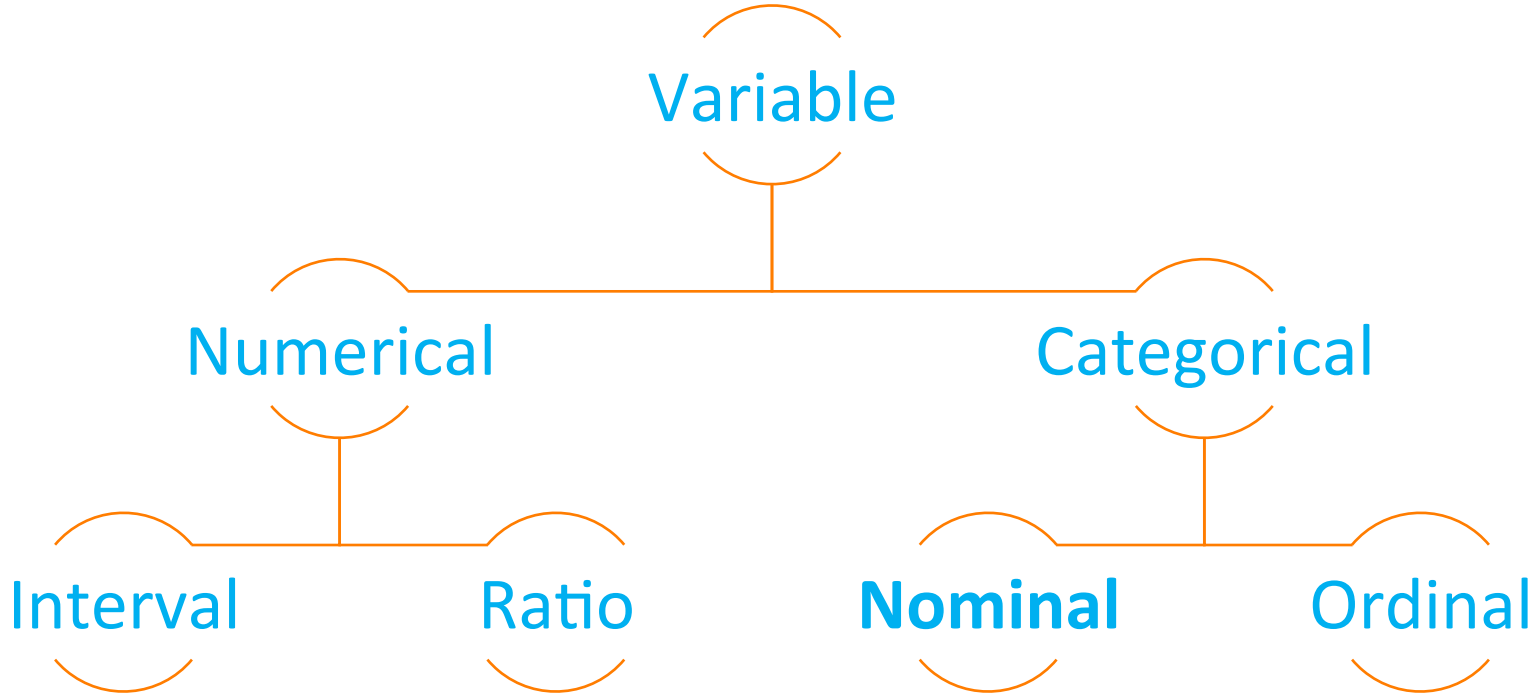
Definitions: Variables



Interval: the distance between two values is meaningful (0 to 100 Celsius on a thermometer)

Ratio: possesses a unique, meaningful and non-arbitrary zero value (Kelvin on the thermometer)

Definitions: Variables

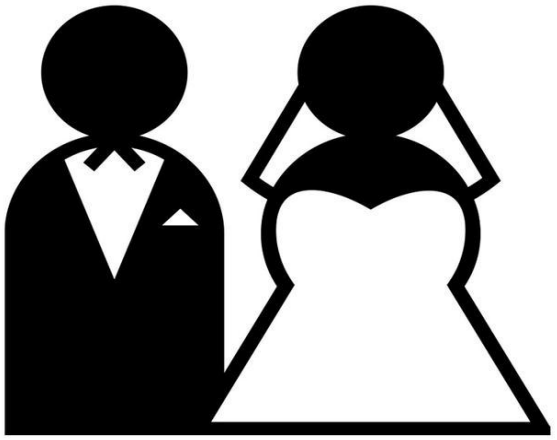


Nominal Variables

- Nominal= “Name ONLY”
- Nominal variables contain mere codes assigned to objects as labels, they are not measurements.
- Not a measure of quantity. Measures identity and difference. People either belong to a group or they do not.

Nominal Variables: More Examples

- **Marital Status: Married, Single, Divorced, Widowed**



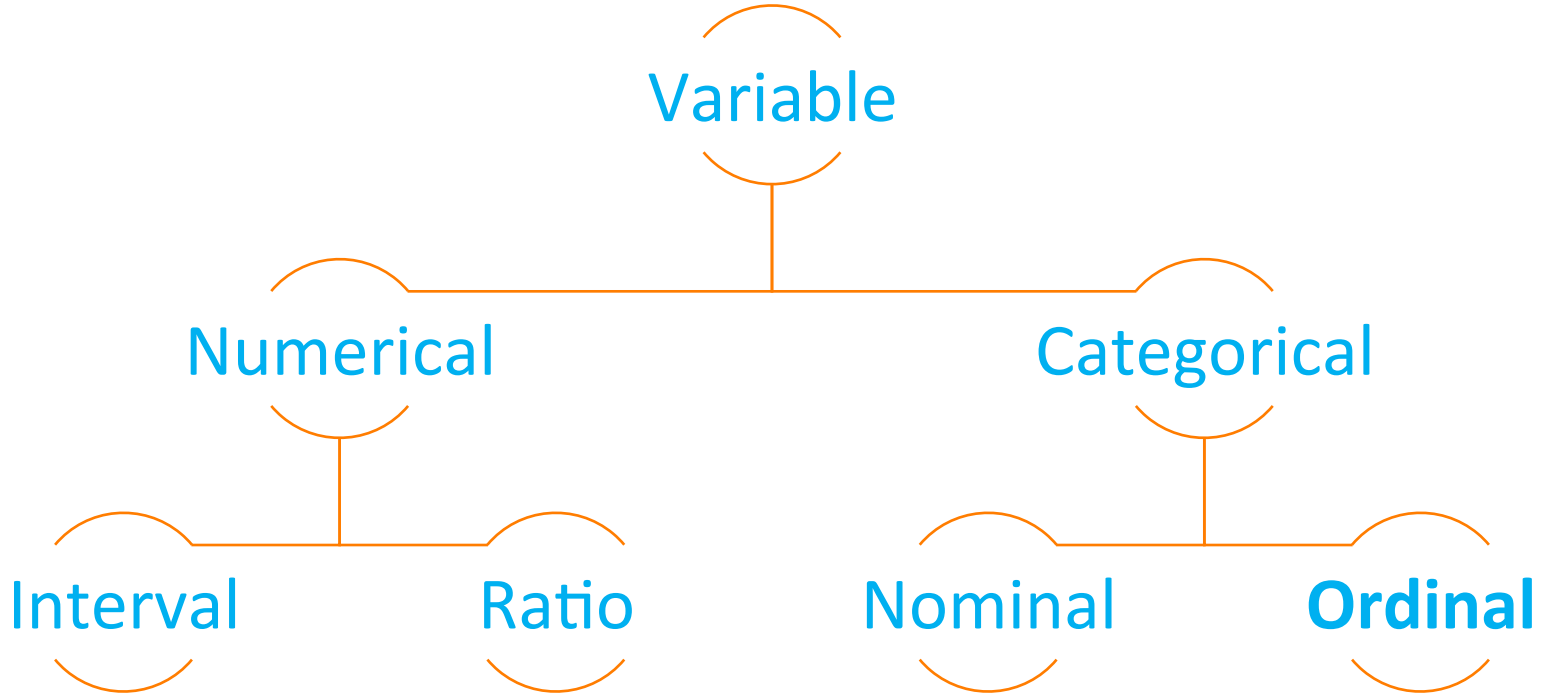
- **Country of Origin:**
  - **1 = United States**
  - **2 = Canada**
  - **3 = Mexico**
  - **4 = Other**

(Notice: Numbers are only labels!!!)

Nominal Variable: What Statistics/ Operations Can I Compute?

Statistics/Operation	OK to Compute?
Frequency Distribution and mode	Yes
Median And Percentiles.	No
Addition Or Subtraction.	No
Mean, Standard Deviation, Standard Error of The Mean.	No
Ratio, Or Coefficient Of Variation.	No

Definitions: Variables



Ordinal Variables: Order Matters

- Ranks Individual attributes in same group
- Unit of measure not available
- Designates an ordering: greater than, less than.
- Does not assume that the intervals between numbers are equal.
- Example: student A is taller than student B

Ordinal Variable: Examples

- **Rank your food preference where**

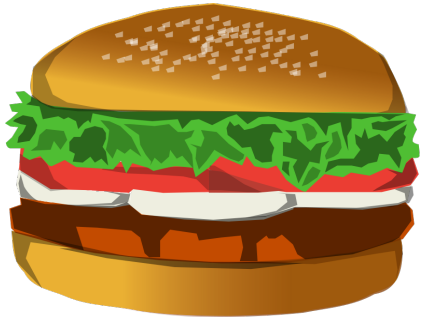
  1 = favorite food and 4 = least favorite:

  _____ sushi

  _____ chocolate

  _____ hamburger

  _____ papaya

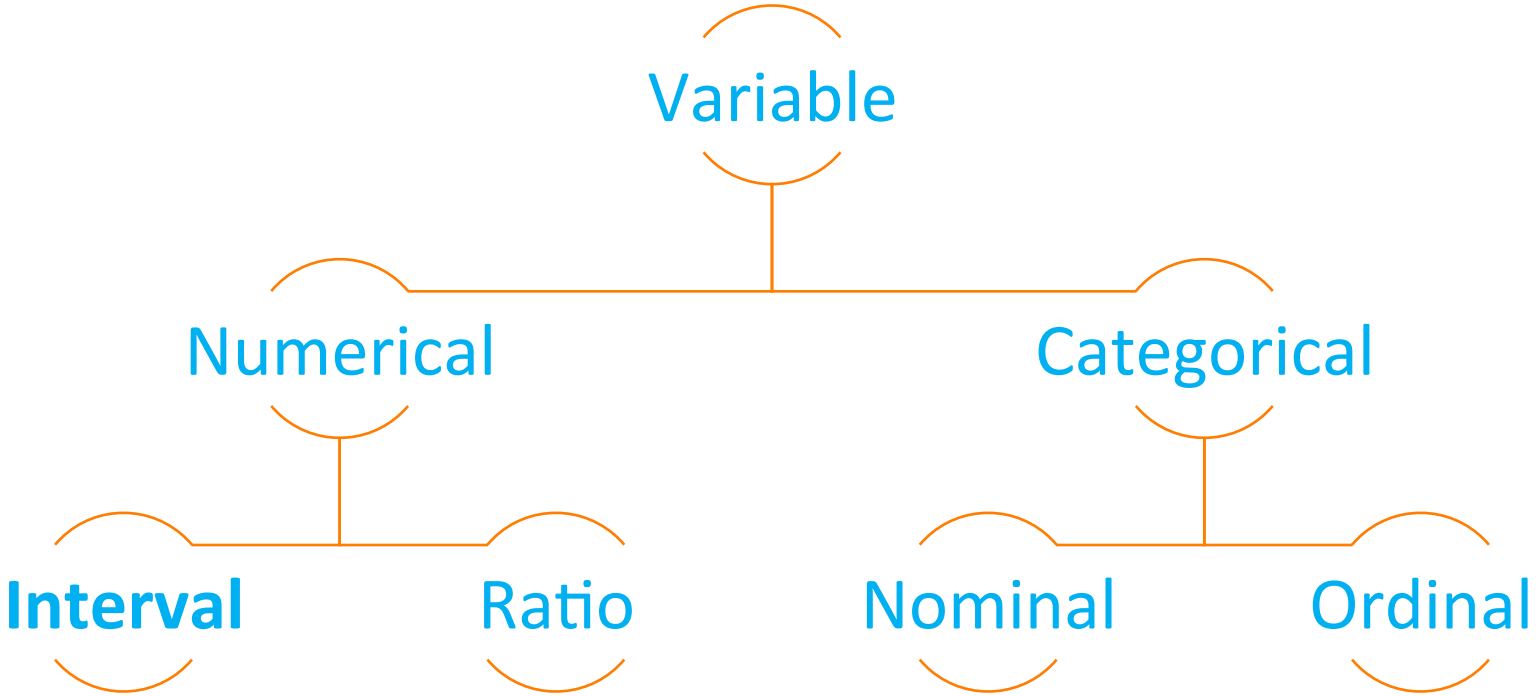


- **Final position of horses in a thoroughbred race is an ordinal variable. The horses finish first, second, third, fourth, and so on.**

Ordinal Variable: What Statistics/ Operations Can I Compute?

Statistic/Operation	Compute?
Frequency Distribution.	Yes
Median And Percentiles.	Yes
Add Or Subtract.	No
Mean, Standard Deviation, Standard Error Of The Mean.	No
Ratio, Or Coefficient Of Variation.	No

Definitions: Variables



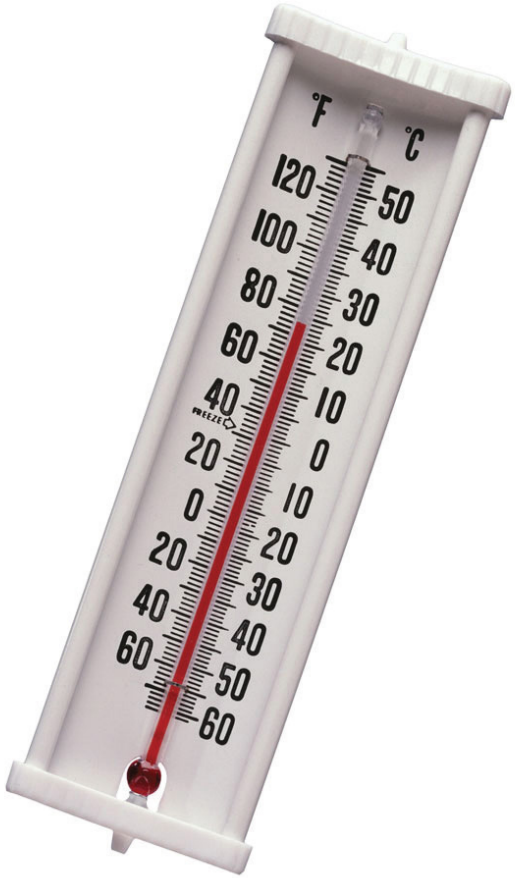
Interval Variables

- Classifies data into groups or categories
- Determines the preferences between items
- Zero point on the internal scale is arbitrary zero, it is not the true zero point
- Designates an equal-interval ordering.

Interval Variables: Examples

Some Temperature Scales

- **Fahrenheit (25 is 5 degrees hotter than 20)**



Interval Variables: Examples

How do you feel about "Intro to Data Analysis"?

1 = I'm totally dreading this class!

2 = I'd rather not take this class.

3 = I feel neutral about this class.

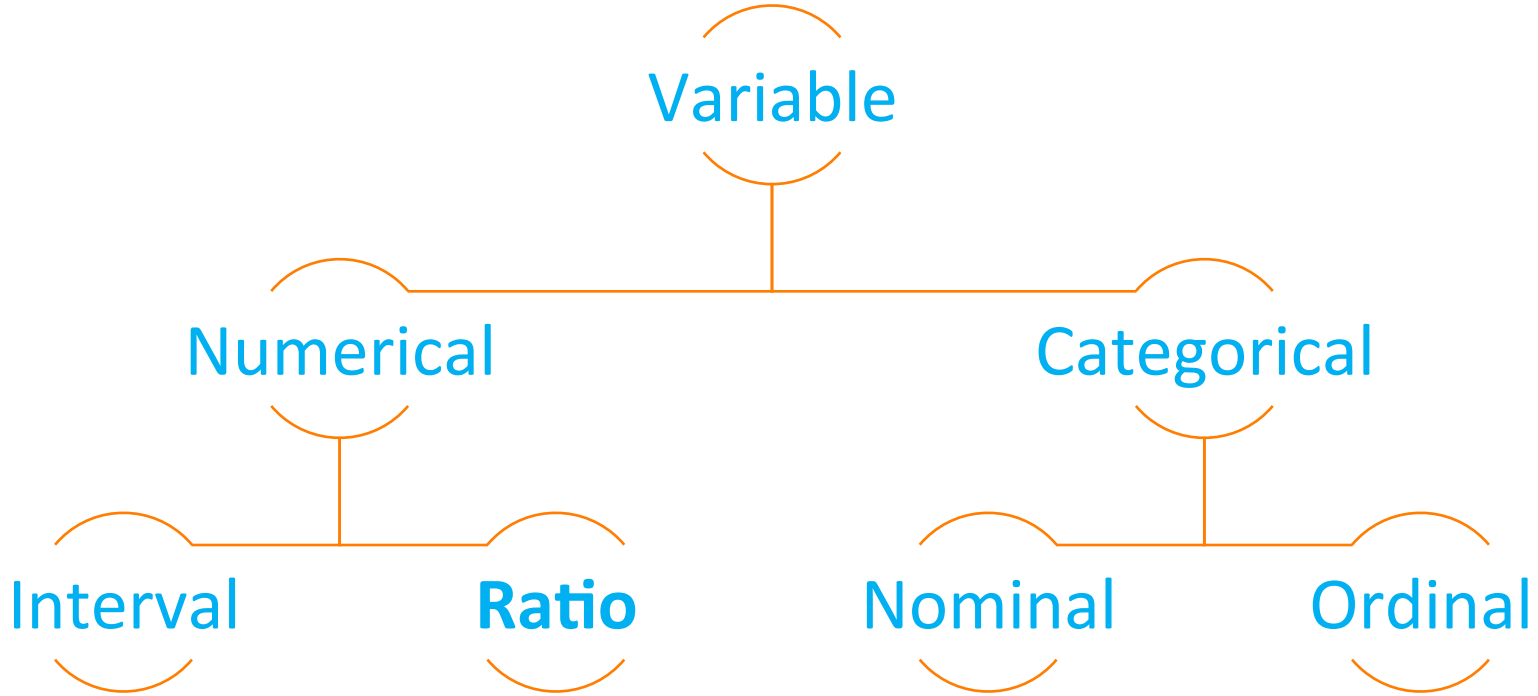
4 = I'm interested in this class.

5 = I'm SO excited to take this class!

Interval Variables: Which Statistic/ Operation Can I Compute?

Statistic/Operation	Compute?
Frequency Distribution.	Yes
Median And Percentiles.	Yes
Add Or Subtract.	Yes
Mean, Standard Deviation, Correlation, Regression, Analysis Of Variance	Yes
Ratio, Or Coefficient Of Variation.	No

Definitions: Variables



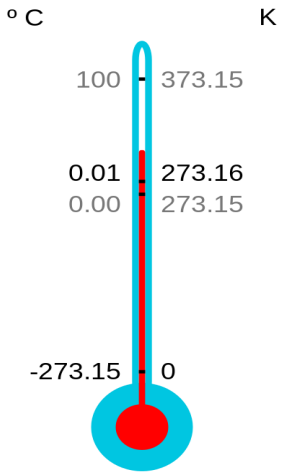
Ratio Variables

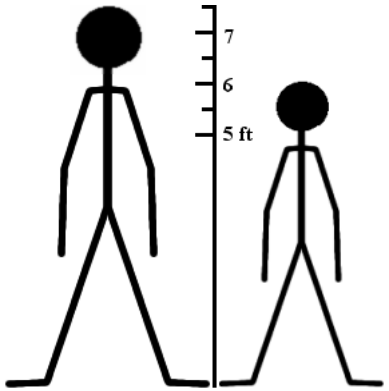
- **This is the highest level of measurement and has the properties of an interval scale; coupled with fixed origin or zero point.**

- **It clearly defines the magnitude or value of difference between two individual items or intervals in same group.**

Ratio Variables

- Temperature in Kelvin (zero is the absence of heat. Can't get colder). 0 kelvin is -273 degree celsius
- Measurements of heights of students in this class (zero means complete lack of height).
- True Multiples: Someone 6 ft tall is *twice* as tall as someone 3 feet tall.
- Heart beats per minute has a very natural zero point. Zero means no heart beats.





Ratio Variable: What Statistics/ Operations Can I Compute?

Statistic/Operation	Compute?
Frequency Distribution.	Yes
Median And Percentiles.	Yes
Add Or Subtract.	Yes
Mean, Standard Deviation, Correlation, Regression, Analysis Of Variance	Yes
Ratio, Or Coefficient Of Variation.	Yes

Summary of Variable Types

Level of measurement: (Variables)	Put data in categories	Arrange data in order	Subtract data values	Determine if one data value is a multiple of another
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Central Tendency of Data

We look at the basic description of the typical values of data sets

Section 4

Central Tendency (Averages)

- Characterizes a set of data by a central or typical number
- Mean, Median, and Mode are most common measures of central tendency
- Each of them has strengths and weaknesses

(Arithmetic) Mean

- Most common meaning of “Average”
- Sum of all individual elements divided by the total number of elements
- Affected by sensitive to outliers
- Be careful about weights and units

(Arithmetic) Mean

Given a data set of

1,2,3,3,4,11

$$\text{Mean} = 1+2+3+3+4+11/6 = 4$$

Median

- The midpoint of a distribution
- Half of the data above, half below
- Insensitive to outliers
- In the case of an even number of data points, take the mean of the middle two points

$$Median(1,2,3,3,4,11)=3+3/2=3$$

Mode

- The value which appears most in a distribution
- A distribution may have more than one mode

$$Mode(1,2,3,3,4,11)=3$$

Central Tendency

CLASSROOM WORK

30 minutes

Exercise 2.1 Use "Gradebook Single.xlsx"
Grades for 44 students out of 20

Task:

- **Create a column for out of 100 grades**
- **Find Average (=Average), Standard deviation (=STDEV.P), Min (=min), Max (=max)**
- **Find frequency count and plot the graph (=Frequency)**
- **Find cumulative frequency and plot the graph**

Central Tendency

CLASSROOM WORK

30 minutes

Exercise 2.2 Use "Gradebook.xlsx"
Three tests - Grades out of 100

Task:

- **Find Average (=Average), Standard deviation (=STDEV.P), Min (=min), Max (=max) for all three tests**
- **Find frequency count and plot the graph (=Frequency) for all three test together**
- **Find frequency count and plot the graph (=Frequency) for individual tests**

Central Tendency

GROUP WORK

20 minutes

- **Draw Conclusion and Give Recommendations**

- # Day 1

Basic Probability

We look at the basics of gambling and how they apply to pretty much everything

Section 5

Probability

- The measure of the likelihood that an event will occur.
- Often Associated with gambling, but has applications in business, finance, manufacturing, science and other fields.
- Expressed as a number between 0 and 1, where 0 means no chance of occurring and 1 means it will certainly occur.

Probability Uses In Business

- Quality Control (Manufacturing Defects, etc.)
  - Six sigma Quality
- Sample Design for Surveys (Random Samples)
- Predictive Analytics (Decision Trees, Monte Carlo Simulations, etc.)
- Risk Management
- Stock Control (Predict when items need to be reordered)

Probability: Several Ways We Can Calculate It

- **Simple probability calculated from frequencies**

RELATIVE FREQUENCY

Color	Frequency	Relative Frequency
Purple	7	7/20= 35%
Blue	3	3/20= 15%
Pink	5	5/20=25%
Orange	5	5/20=25%
Total	20	20/20 = 100%

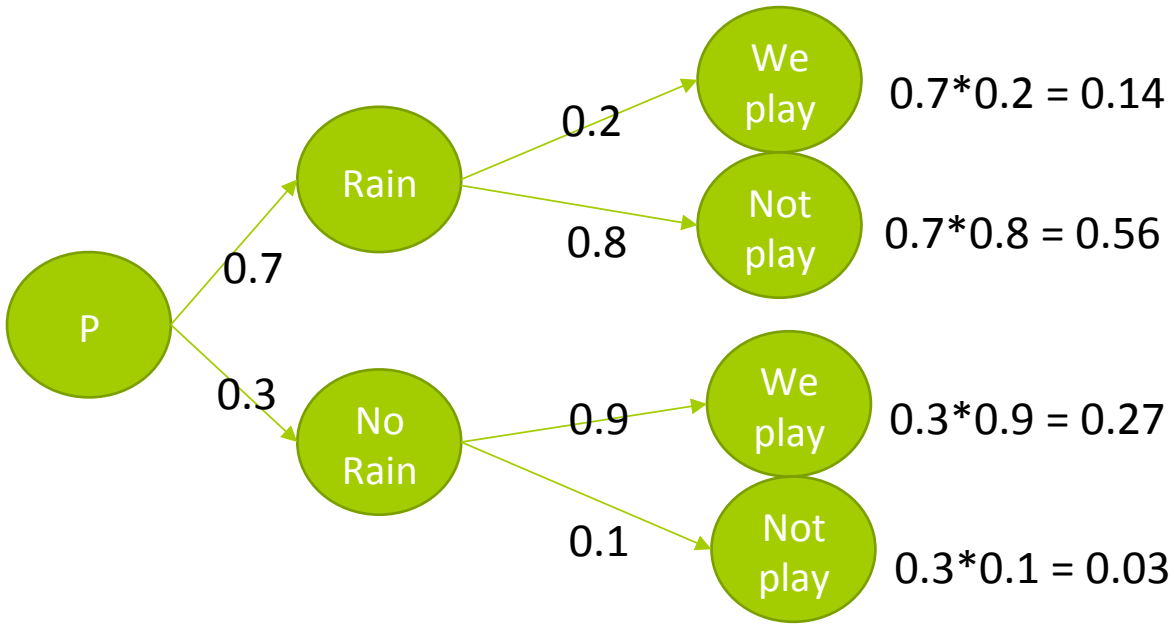
Probability: Several Ways We Can Calculate It

- **Conditional probability from tree diagrams**

Simple probability:
P(rain) = 0.7
P(no rain) = 0.3
Sum = 1.0

Joint probability:
P(play) = 0.14+0.27 = 0.41
P(no play) = 0.56+0.03 = 0.59
Sum = 1.0

Conditional probability
P(play given rain) = 0.14
P(play given no rain) = 0.27
P(no play given rain) = 0.56
P(no play given no rain) = 0.03
Sum = 1.0



P(rain given play) = p(rain|play) = ???

Probability Terms

- **Event:** The outcome to which the probability is assigned.
  - “Heads” on a flipped coin.
  - 4 on a rolled die
  - A Queen drawn from a deck of cards
- **Sample Space:** The set of all possible outcomes
  - Heads and Tails
  - 1, 2, 3, 4, 5, 6
  - The 52 cards
- **Independent:** The outcome of one event does not effect a different event
  - Flipping two coins
- **Dependence:** The outcome of one event effect a different event
  - (playing given rain) OR (playing given no rain)

Calculating Probability

- Assuming a sample space of equally likely events

$$P(A) = \text{Outcomes where A occurs} / \text{Total number of outcomes}$$

- For example, on a flipped coin the sample space contains Heads and Tails, so:

$$P(\text{Heads}) = \text{Outcomes of Heads} / \text{All possible outcomes} = 1/2 = 0.5$$

Calculating probability

CLASSROOM WORK

30 minutes

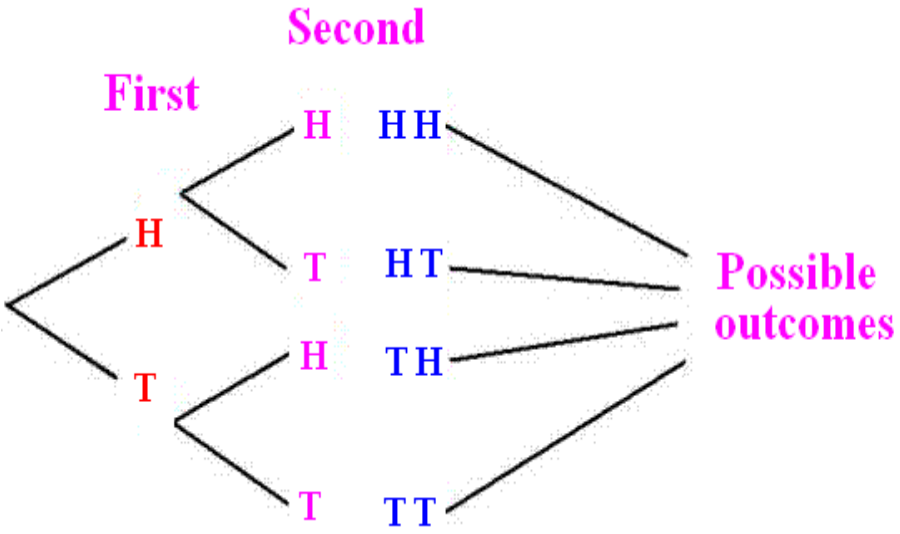
1. If I flip 2 *fair* coins one by one:
   - what is the sample space
   - what is the probability I get at least one head?

2. If I roll two fair dices at once:
   - what is the sample space
   - what is the probability that the sum is 4?

Calculating probability Answer 1

If I flip 2 *fair* coins one by one:

- what is the sample space
- what is the probability I get at least one head?



Sample space:- *{HH, HT, TH, TT} = Total 4 possible outcomes*

P(At least one head)= Outcomes of HH or HT or TH/Total outcomes = **3/4**

Calculating probability Answer 2

If I roll two fair dices at once:

- what is the sample space
- what is the probability that the sum is 4?

Outcomes for Two Dice

	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Sample space:- *{11, 12, …. 65, 66} = total 36 possible outcomes*

P(Sum 7)= *Outcomes of* 31, 22,13/ *Total outcomes* =**3/36**

Calculating Probability from a Contingency Table

A supermarket did a survey to investigate customers' drink spending in relation to their mode of travel.

	Amount spent on drink			
Mode of travel	None	$1 and under $20	At least $20	Total
On foot	40	20	10	70
By bus	30	35	15	80
By car	25	33	42	100
Total	95	88	67	250

$$P(\text{Customer spends} < \$20) = 95 + 88 / 250 = 0.732$$

Conditional Probability

	Amount spent on drink			
Mode of travel	None	$1 and under $20	At least $20	Total
On foot	40	20	10	70
By bus	30	35	15	80
By car	25	33	42	100
Total	95	88	67	250

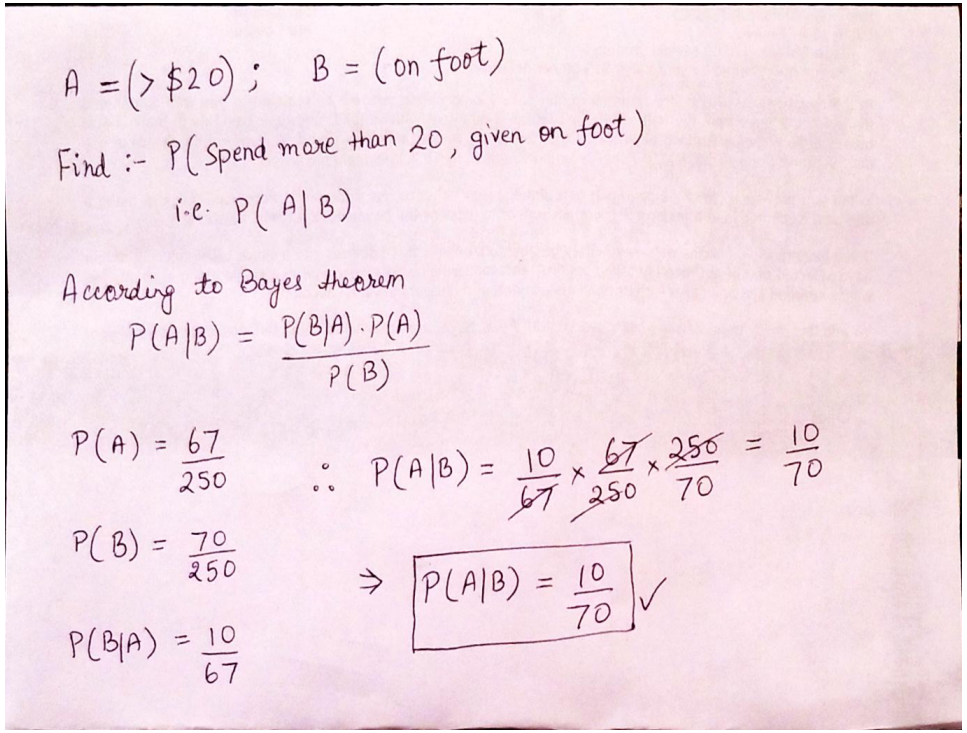
Probability that a customer will spend more than $20, GIVEN that they arrived on foot?

$$P = 10/70 = 0.143$$

Conditional Probability – Bayes Theorem

	Amount spent on drink			
Mode of travel	None	$1 and under $20	At least $20	Total
On foot	40	20	10	70
By bus	30	35	15	80
By car	25	33	42	100
Total	95	88	67	250

Same as the last slide but more formally



Expected Value

Expectation:-

mean or expectation - measure of 'central location' of a random variable

Discrete Random Variable $\mu_X = \mathrm{E}(X) = \sum x\, p_X(x),$

Continuous Random Variable $\mu_X = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$

Expected Value

- Given the probabilities and their values, what value should I expect
- Similar to a weighted average

Possibility (n)	Probability P(n)	Return x(n)
1	0.27	105
2	0.30	100
3	0.23	156
4	0.20	130

- $E(x)=\sum n\uparrow\blacksquare P(n)x(n)$
$=0.27\times105+0.30\times100+0.23\times156+0.20\times130=120.23$

Frequency Distribution

The frequency with which observations are assigned to each category or point on a measurement scale.

- Most basic form of descriptive statistics
- May be expressed as a percentage of the total sample found in each category

Frequency Distribution

How the distribution is read depends on the measurement level.

- Nominal scales are read as discrete measurements at each level
- Ordinal measures show tendencies, but categories should not be compared
- Interval and ratio scales allow for comparison among categories

Frequency Distribution

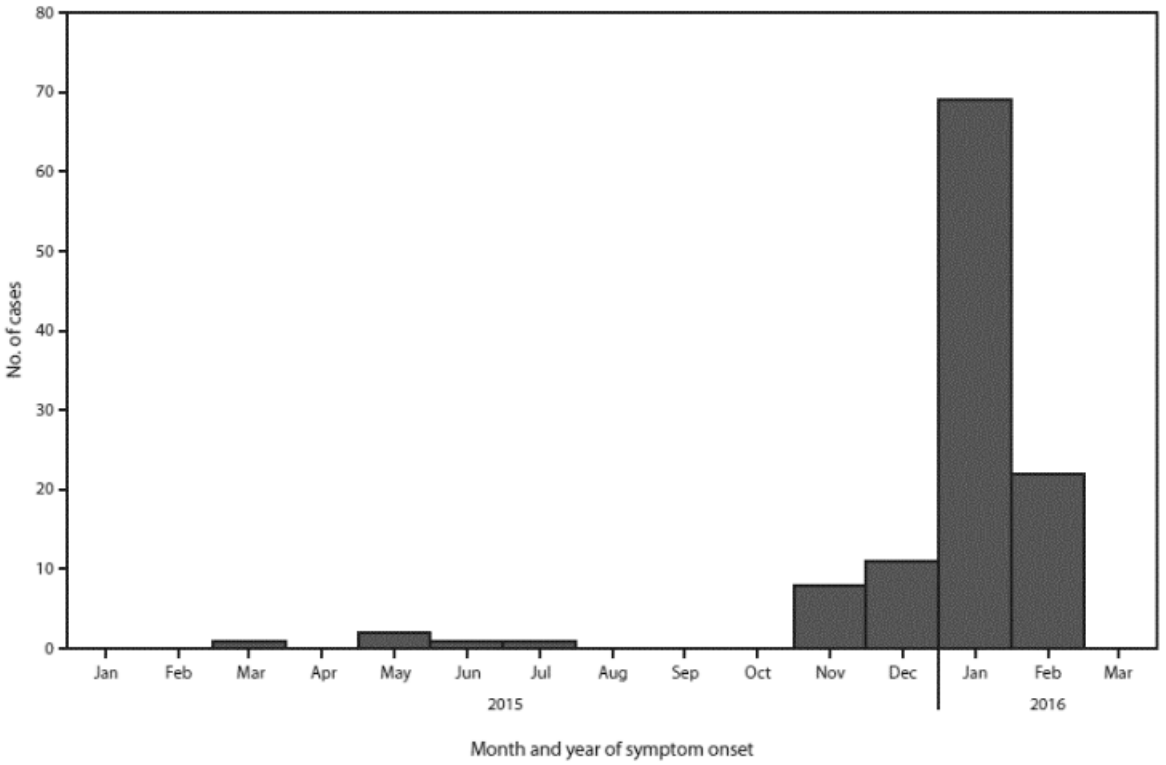


Figure. Month of illness onset for 115 patients with laboratory evidence of Zika virus infection among residents of U.S. states and the District of Columbia — January 1, 2015–February 26, 2016*

Probability – Frequency Distribution

CLASSROOM WORK

30 minutes

Start with Probability.xlsx

1. Plot frequency distribution plot for {2 heads, 1H 1T, 2 tails} outcomes from 2 flips of a coin
2. Plot frequency distribution plot for all possible outcomes of sum of 2 dice rolls

Distributions

We look at at how probabilities of events lead to distributions, how to describe these distributions, and what they tell us

Section 6

Distributions

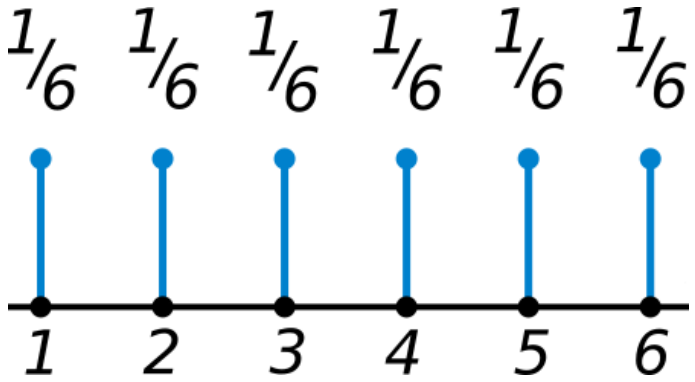
- Mathematical description of the probabilities of events
- May be discrete or continuous
- All probabilities should sum to 1

Discrete Distributions

- **Probability Mass Function (PMF) –**

If X is a discrete random variable, and its range is a countable set, the possible outcomes can be = {x1, x2, x3, ….}. The probabilities of events {X = xi} are formally shown by the probability mass function

Thus it is used to calculate **P(X = x) = some discrete value**



Continuous Distributions

- **Probability Density Function (PDF) –**

If X is a continuous random variable, and it's range is not a countable set, the possible outcomes can be anything.

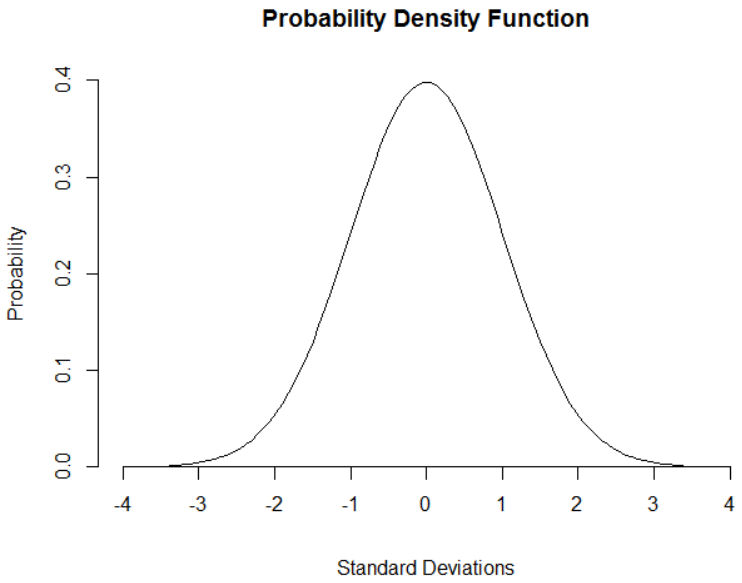
Since there are infinite number of possible values, the absolute likelihood of a random variable to take a value X is zero.

Thus **P(X = x) = 0 for all x belongs to real number**

Continuous Distributions

How much more likely it is that the random variable would equal one sample compared to the other sample.

- relative likelihood of a certain X value occurring
- **Sum over a range to get a probability.**

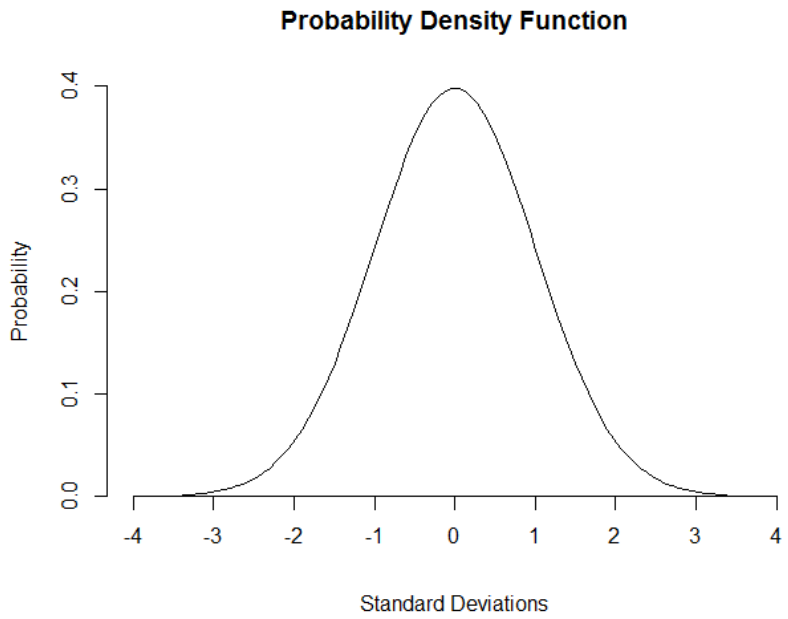


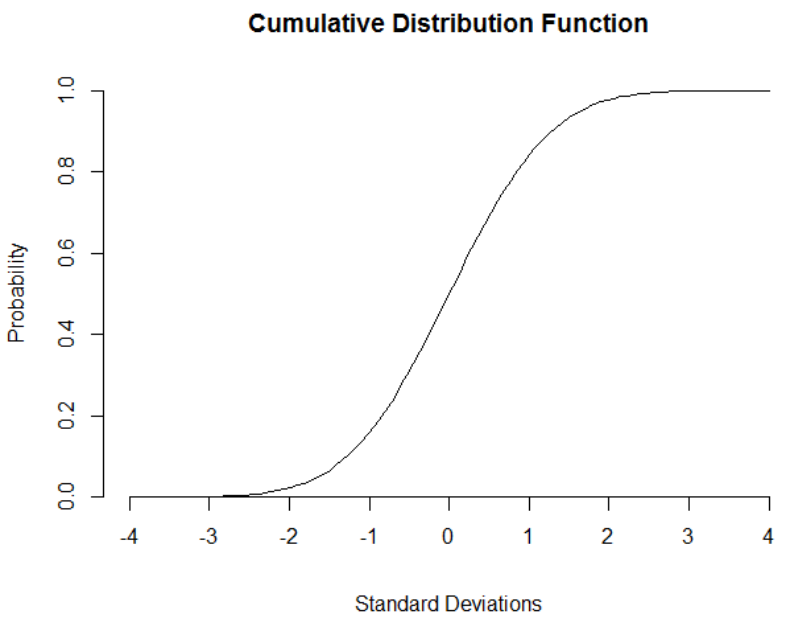
$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)\,dx.$$

This probability is given by the integral of this variable's PDF over that range

Continuous Distributions

- **Cumulative Distribution Function - probability of a particular X value or less**





Measure and Variability

- There will always be variability in the data.
- Statistics helps measure and characterize variability.
- For example, controlling (or reducing) variability in a manufacturing process equates to statistical process control.

Variation by Range

- The difference between the greatest and least data point
- Useful
  - **but susceptible to outliers**

Variation by Quartiles

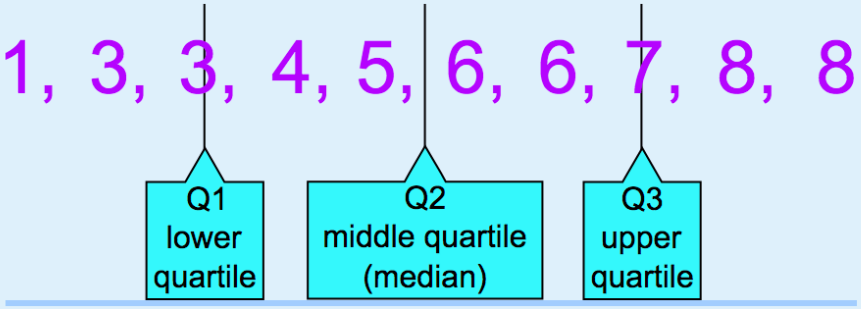
- Divide distribution into 4 sets with equal numbers of points
- 25% of the data is less than the First (or Lower) Quartile
- The Median is the Second Quartile
- 25% of the data is greater than the Third (or Upper) Quartile
- The Interquartile Range is the difference between the Third and First Quartiles

Quartiles

Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:



In this case Quartile 2 is half way between 5 and 6:

$$Q2 = (5+6)/2 = \mathbf{5.5}$$

And the result is:

- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**

Variation by Variance

- **Squared** differences between the mean and each observation, divided by N
- It is always positive but not scaled to the observations
  - Say, miles squared versus miles

$$\operatorname{var}(X) = \sigma_X^2 = \mathrm{E}[(X - \mu)^2], \qquad \text{where } \mu = \mathrm{E}(X).$$

discrete

$$\operatorname{var}(X) = \sigma_X^2 = \mathrm{E}[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx.$$

continuous

Variance

$$Variance(1,1,2,2,3,3,4,4,5,5,100)?$$

- Take the Mean ($130/11 = 11.81$)
- Sum the squared differences with the mean

$$(1-11.8)^2 + (1-11.8)^2 + (2-11.8)^2 + \ldots = 8569$$

- Divide by the count (11)

$$8569/11 = 779$$

Standard Deviation

- The square root of the variance
- It is **scaled to the observations**
- is **always positive**
- Standard Deviation of $(1,1,2,2,3,3,4..,100)$?

$$\sqrt{Variance(arr)}$$
$$=\sqrt{779}=\mathbf{27.9}$$

Compare 27.9 with 99, that was the Range
Which explains data variability better ?

Population vs. Sample

- The definitions above are for the Population variance and standard deviation, i.e. when you have all of the data.
  - VAR.P() and STDEV.P() in excel.

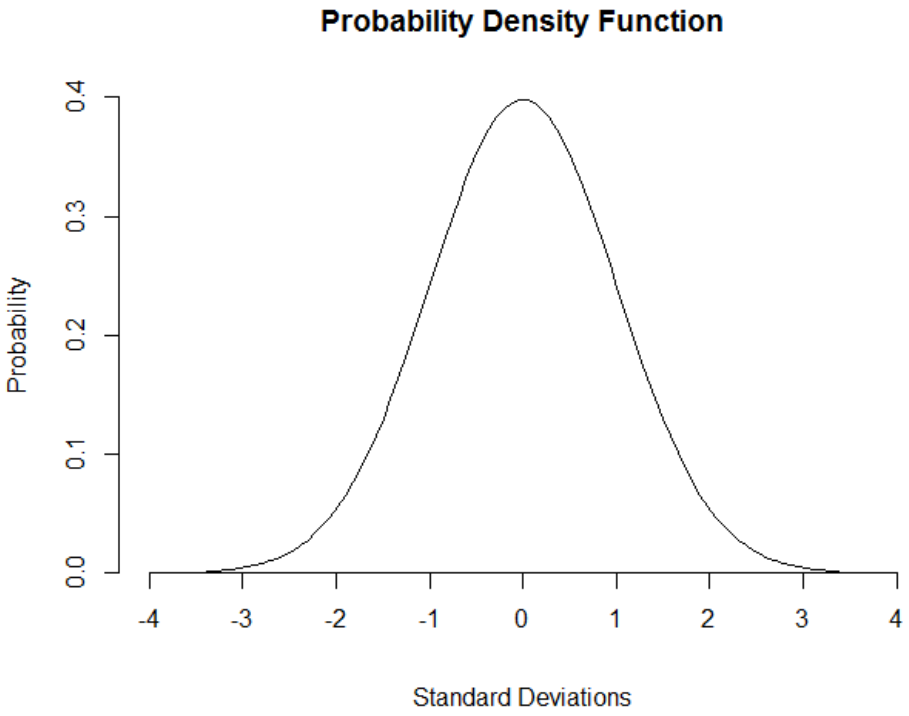
- If you only have a Sample of the data divide by N-1 instead of N for technical reasons.
  - VAR.S() and STDEV.S() in excel.

Application of the Standard Deviation

- Measures deviation of any distribution
- Low value indicates values are close together
- High value indicates values are spread over wider range
- But… when applied to the Normal Distribution, it means specific things

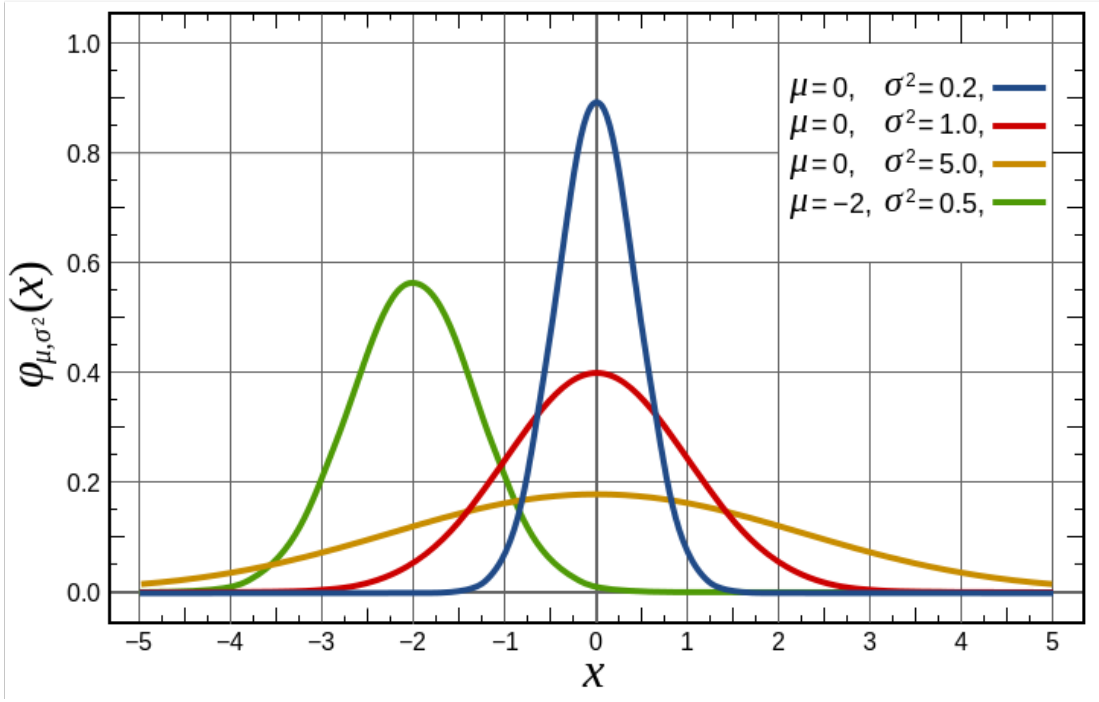
Normal Distribution

- Also called **Gaussian** – the most common continuous probability distribution
- Events are **independent** and **heteroskedastic**
- Useful because of the **Central Limit Theorem** - under some conditions (mentioned above), it states that averages of samples of observations of random variables independently drawn from independent distributions converge in distribution to the normal



$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Normal Distribution



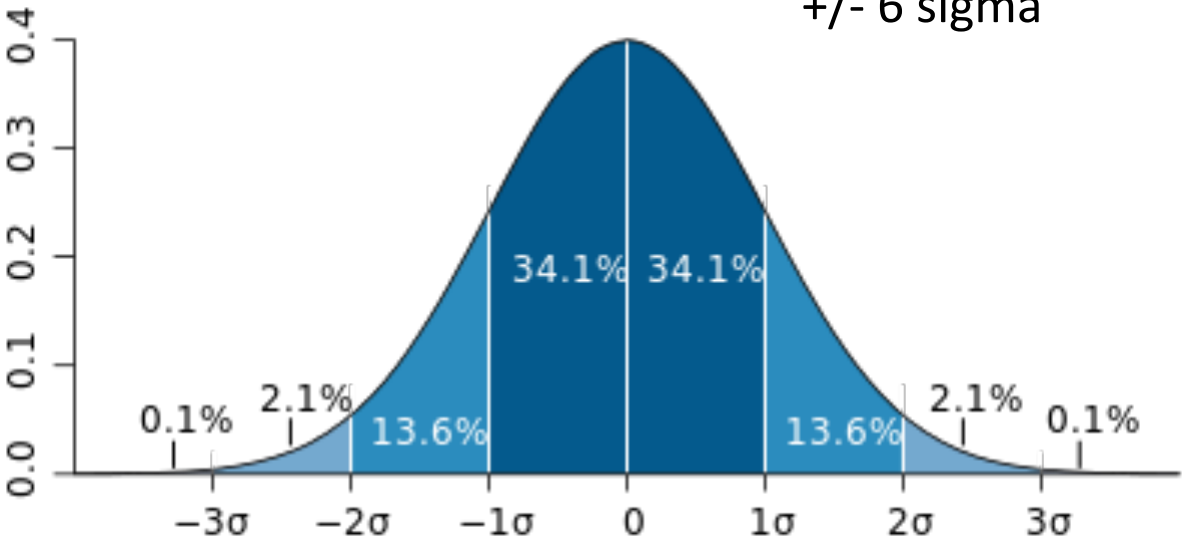
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Ideal normal distribution:-
Mean, median, mode overlaps – but this never happens in real world data

Standard Deviation and the Normal Distribution

For Normally Distributed Data

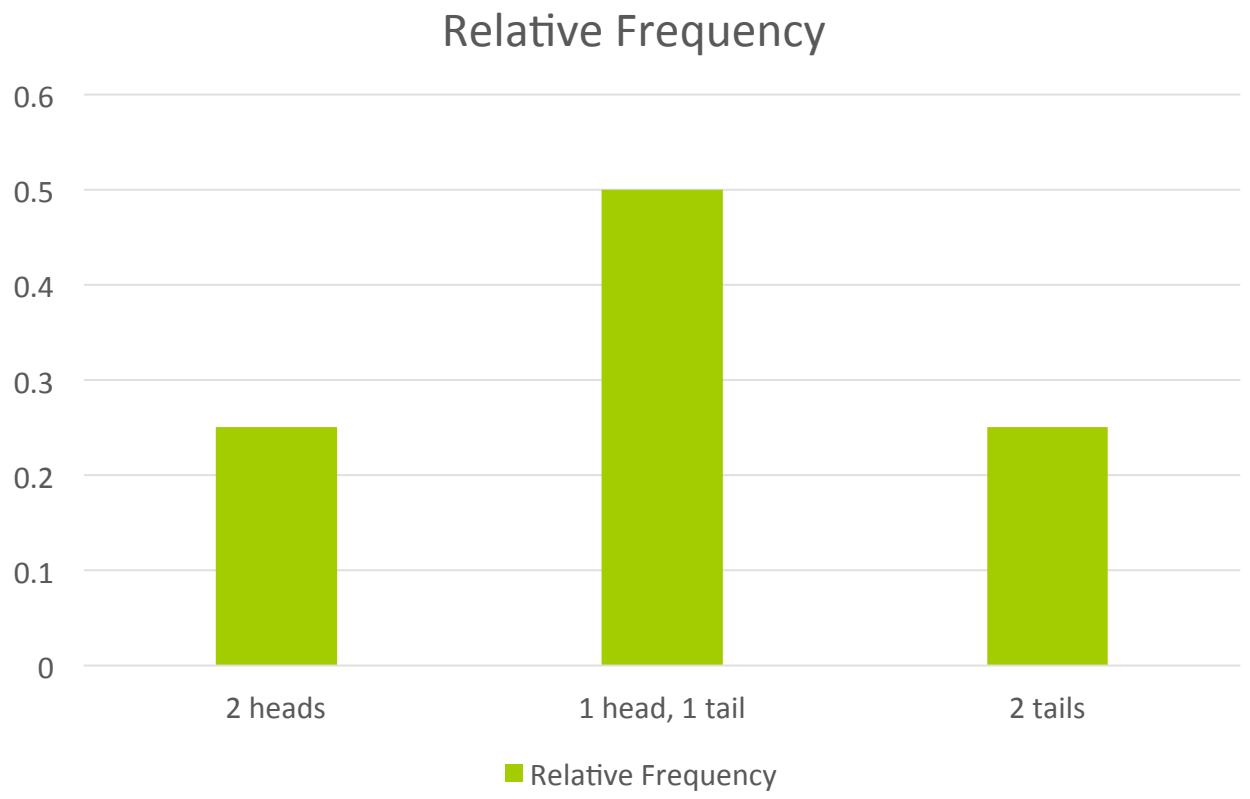
Sigma	Percent
+/- 1 sigma	68.27%
+/- 2 sigma	95.45%
+/- 3 sigma	99.73%
+/- 6 sigma	99.9999998%



Six Sigma Quality

Image: https://en.wikipedia.org/wiki/Standard_deviation

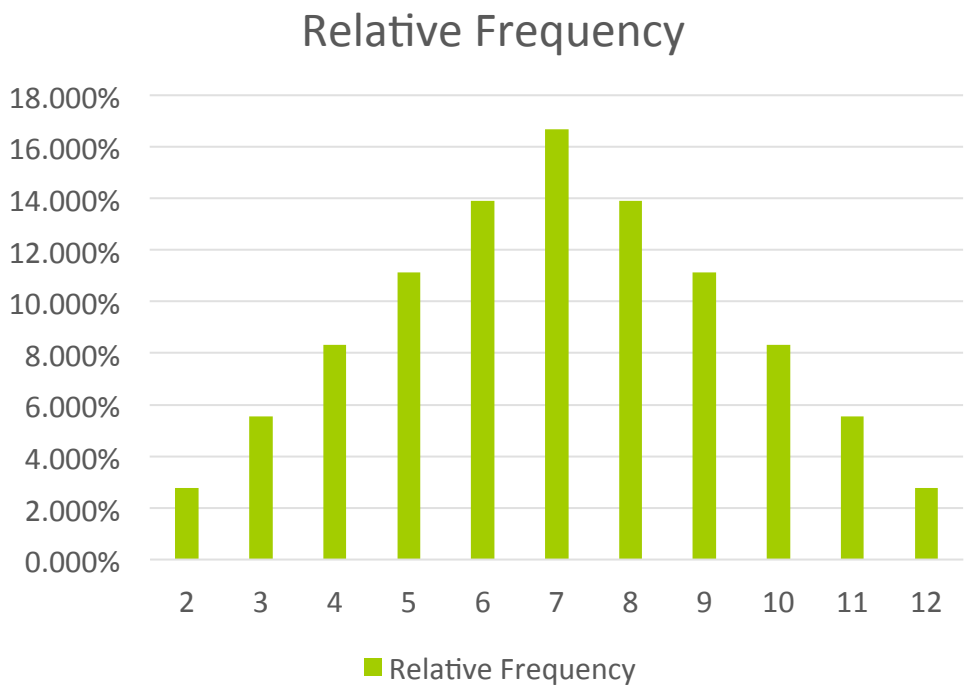
In real-world data – 2 coin flips



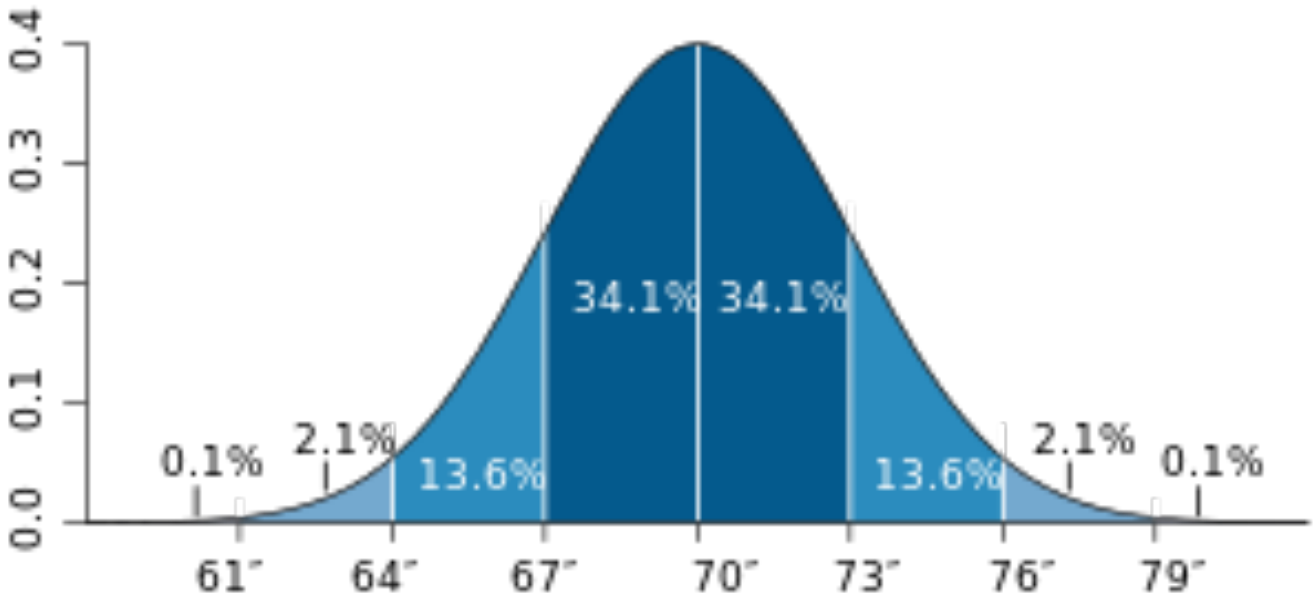
Outcomes	Frequency	Relative frequency
2 heads	1	0.25
1 head, 1 tail	2	0.5
2 tails	1	0.25

In real-world data – 2 dice rolls

Outcome (sum)	Frequency	Relative Frequency
2	1	2.778%
3	2	5.556%
4	3	8.333%
5	4	11.111%
6	5	13.889%
7	6	16.667%
8	5	13.889%
9	4	11.111%
10	3	8.333%
11	2	5.556%
12	1	2.778%



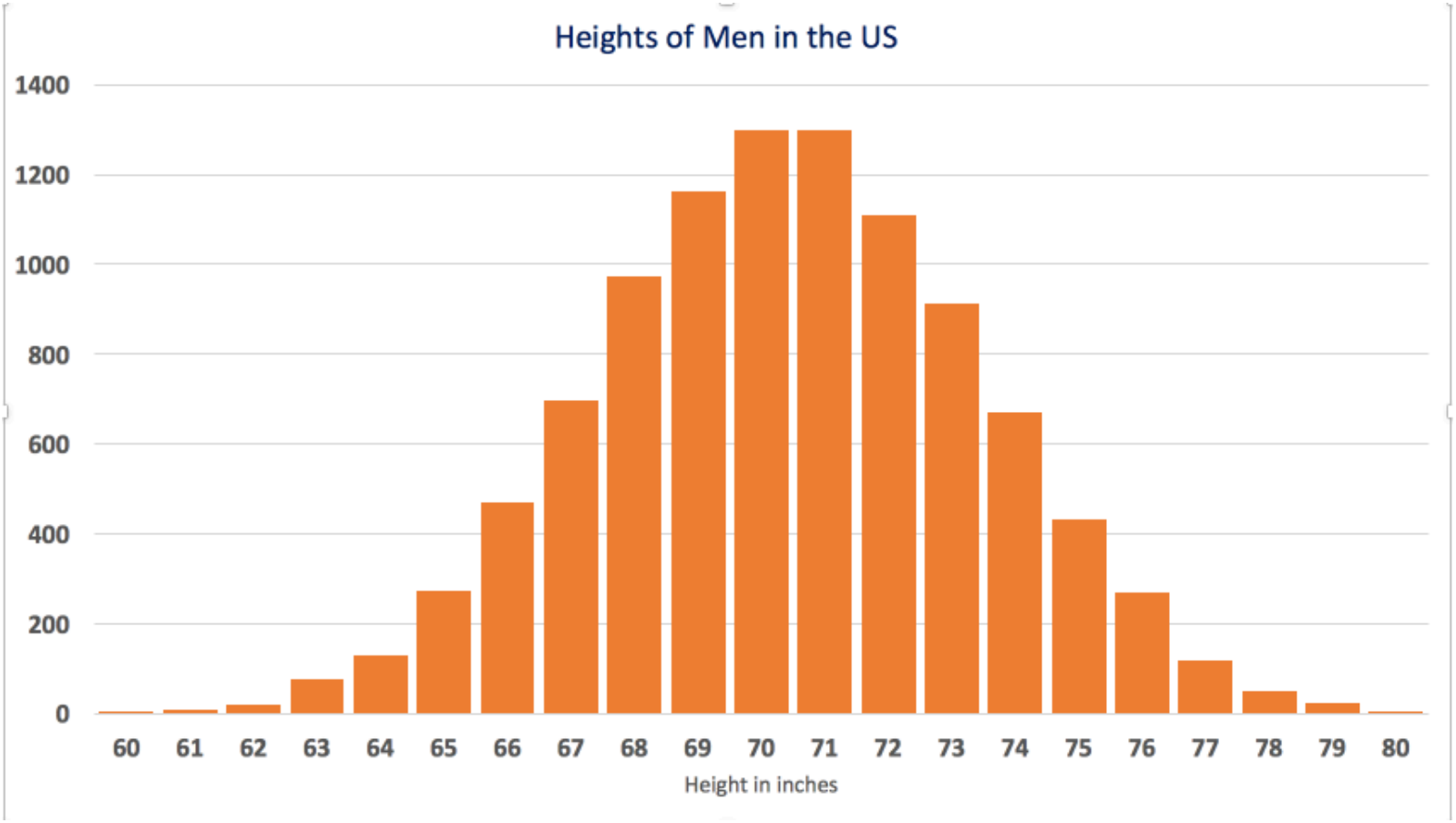
Estimated Heights of Men



Estimated probability density function of the height of adult men in the United States

Image: en.wikipedia.org/wiki/File:Visualisation_mode_median_mean.svg

Estimated Heights of Men



Probability mass function (discrete 10,000 men)

Normal Distribution

CLASSROOM WORK

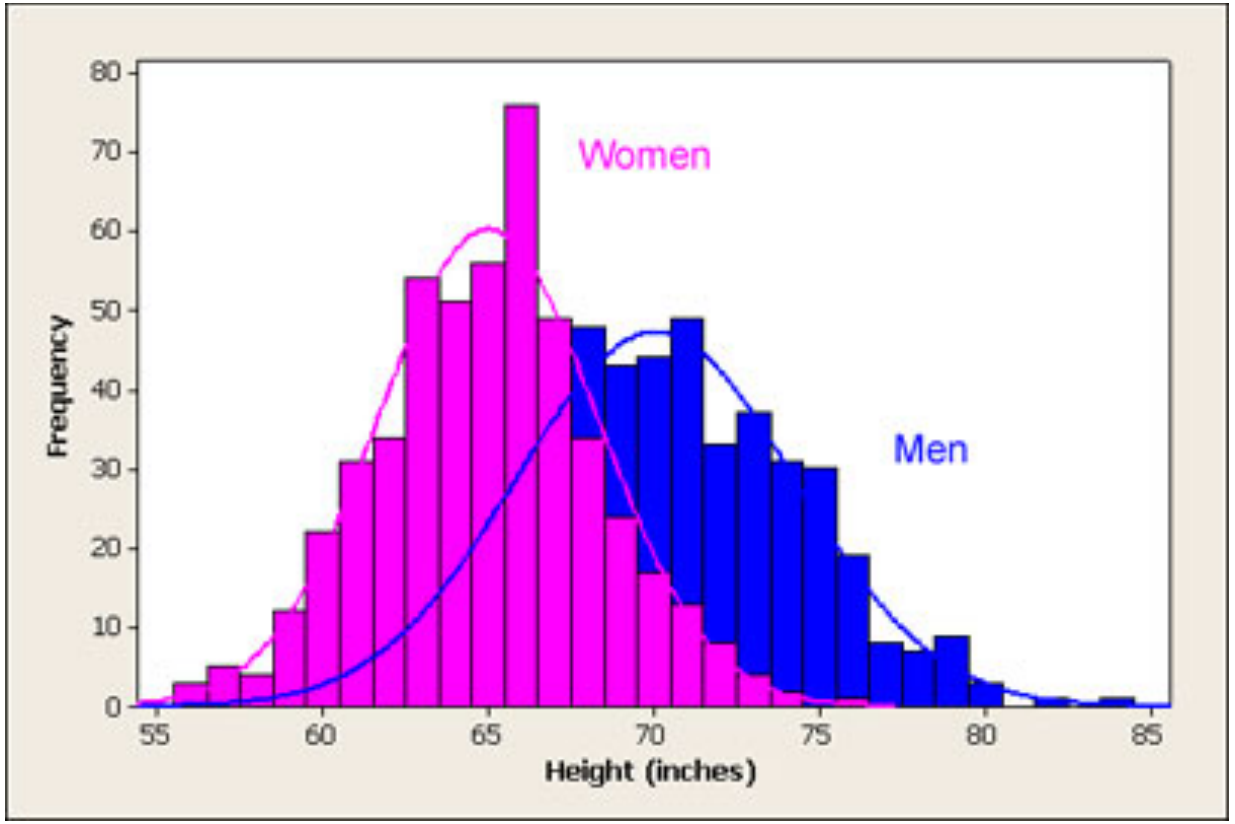
30 minutes

Use “**heights of men.xlsx**” . It has real data of 10,000 heights of discrete men

Find the following

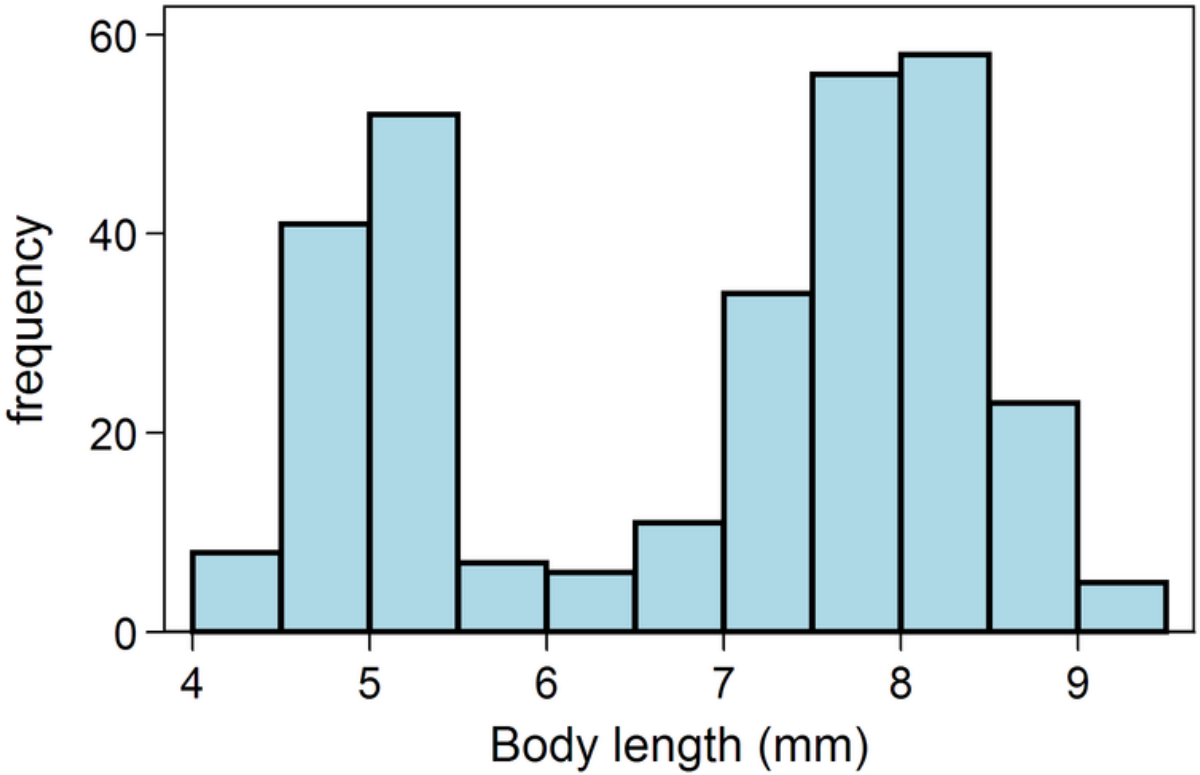
- Average
- Standard deviation
- Plot Frequency distribution

Men and Women



Men and women create two normal distros, or a **bimodal distribution**

Bimodal distribution



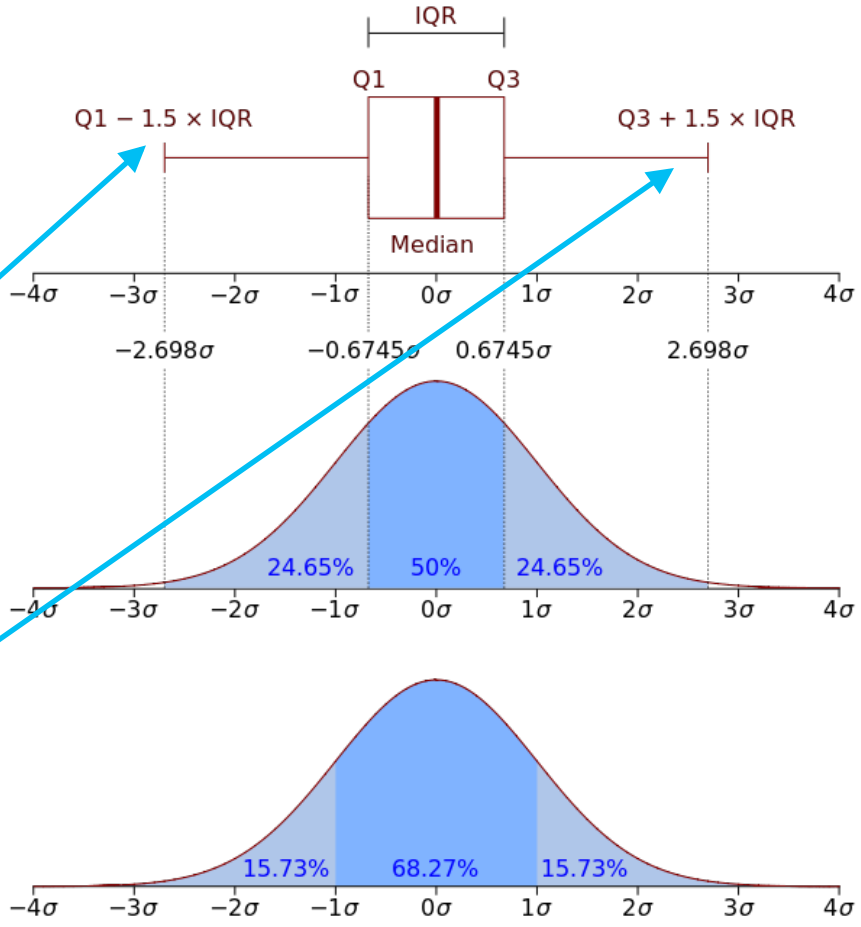
Someone measured the body length of 300 weaver ants and got a bimodal dist as well.

Image: https://en.wikipedia.org/wiki/File:BimodalAnts.png

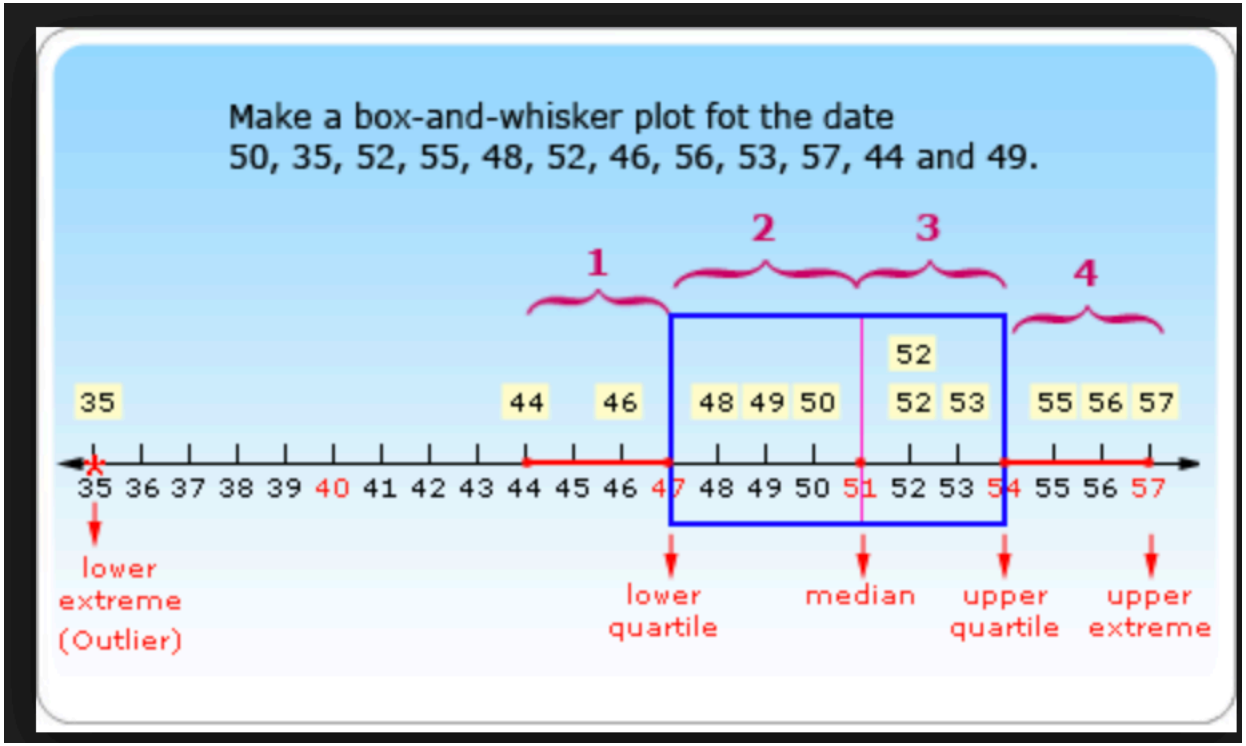
Quartiles and the Normal

Interquartile Range (IQR)
Amount of spread in the middle 50% of a dataset.
OR distance between Q3 and Q1.
Thus IQR = (Q3-Q1)
How to find Q1?
How to find Q3?
IQR from Quartiles slide

1.5 is variable value based on field of study

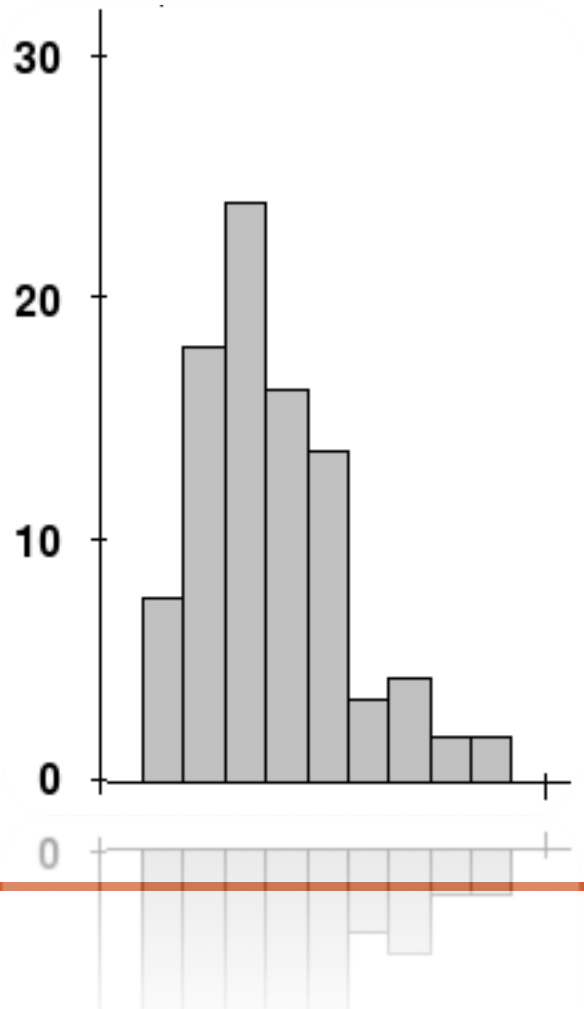


IQR and outliers



x is a moderate outlier if, x<**Q1**-**1.5*IQR** or x>Q3+**1.5*IQR**
x is an extreme outlier if, x<**Q1**-3***IQR** or x>Q3+3***IQR**

Normal but Slightly Skew



- Example distribution with non-zero (positive) skewness.
- Discrete variable
- Long tail to the right

Image: en.wikipedia.org/wiki/File:SkewedDistribution.png

Positive vs. negative skew

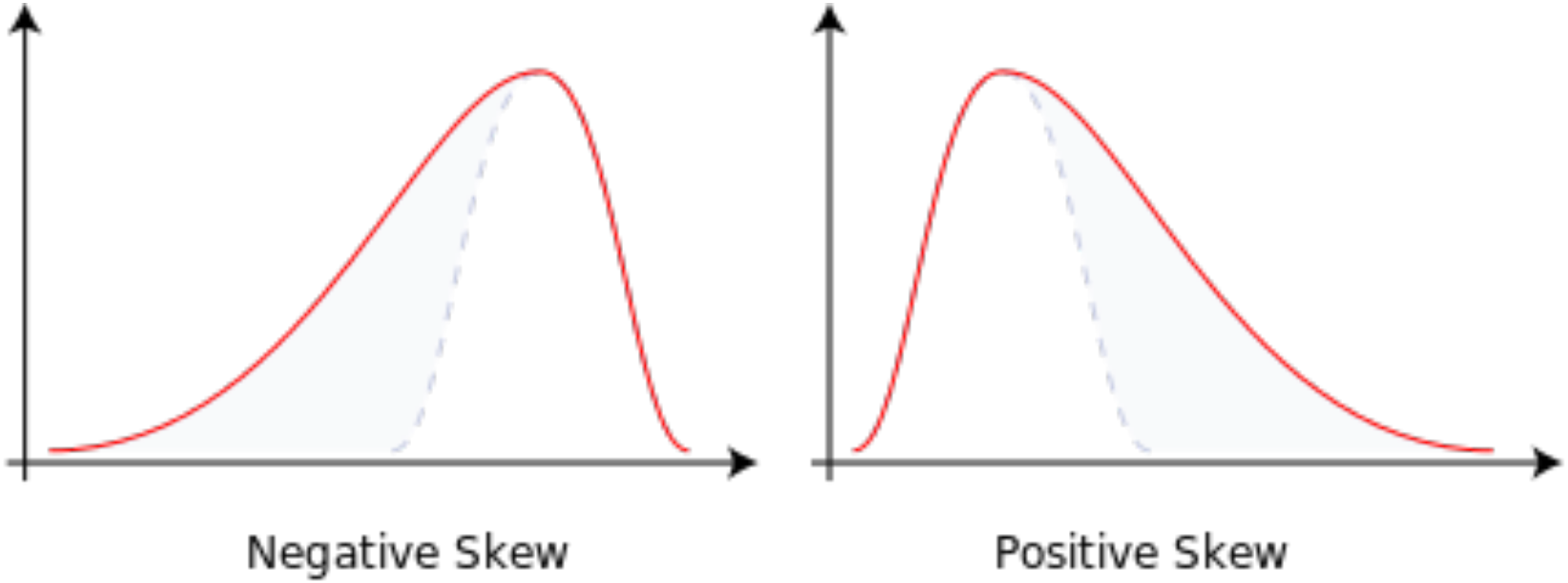
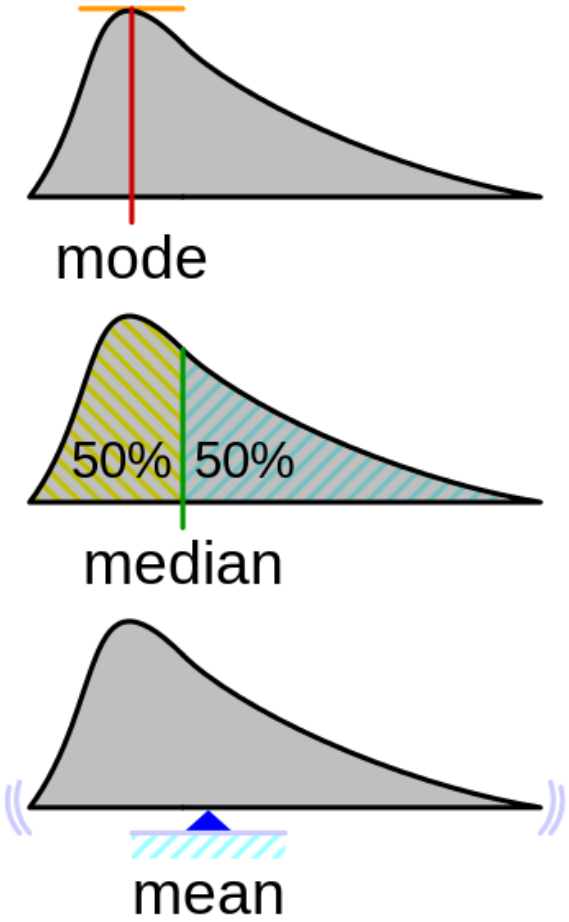


Image: https://en.wikipedia.org/wiki/File:SkewedDistribution.png

So?



Positive skew:
Mode < median < mean

Negative skew:
Mode > median > mean

https://en.wikipedia.org/wiki/Skewness

How to calculate Skewness

1. Pearson's Coefficient of Skewness #1 uses the mode. The formula is:

$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

Where $\bar{X}$ = the mean, Mo = the mode and s = the standard deviation for the sample.

2. Pearson's Coefficient of Skewness #2 uses the median. The formula is:

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

Where $\bar{X}$ = the mean, Mo = the mode and s = the standard deviation for the sample.

It is generally used when you don't know the mode.

- The direction of skewness is given by the sign.
- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.
- A value of zero means no skewness at all.
- A large negative value means the distribution is negatively skewed.
- A large positive value means the distribution is positively skewed.

Z-score

Z-score is the measure of how many standard deviation a number is from it's distribution mean. Z-score of the center of the bell curve is zero.

In the heights of men example,
Mean = 70 inches ; STD = 3.0120

Let's find the z-score of 78 inches.
Z-score = (78-70)/3.012 = **2.65**

The basic z score formula for a sample is:

$$z = (x - \mu) / \sigma$$

The **Z score** is a test of statistical **significance** that helps you decide whether or not to reject the null hypothesis. OR simply
This measure how far an observation is from the mean.

Other distributions

So far we discussed **Normal distribution**, but there are lot of others.

Why do we study distribution statistics

- You want to know the data distribution and accordingly make decisions
- You want to tweak your loss function as per the data distribution for any prediction

# of beds	# of baths	Sq Feet	# floor	Price

Suppose this is Lognormal distribution,

Then you will want to choose lognormal loss function in order to converge the decision boundary

Binomial Distribution (Discrete)

Bernoulli Trial – ‘s’ trials are made of an event, with probability ‘p’ of success in any given trial

Binomial :- gives the discrete probability distribution of obtaining exactly ‘n’ successes out of N Bernoulli trials.

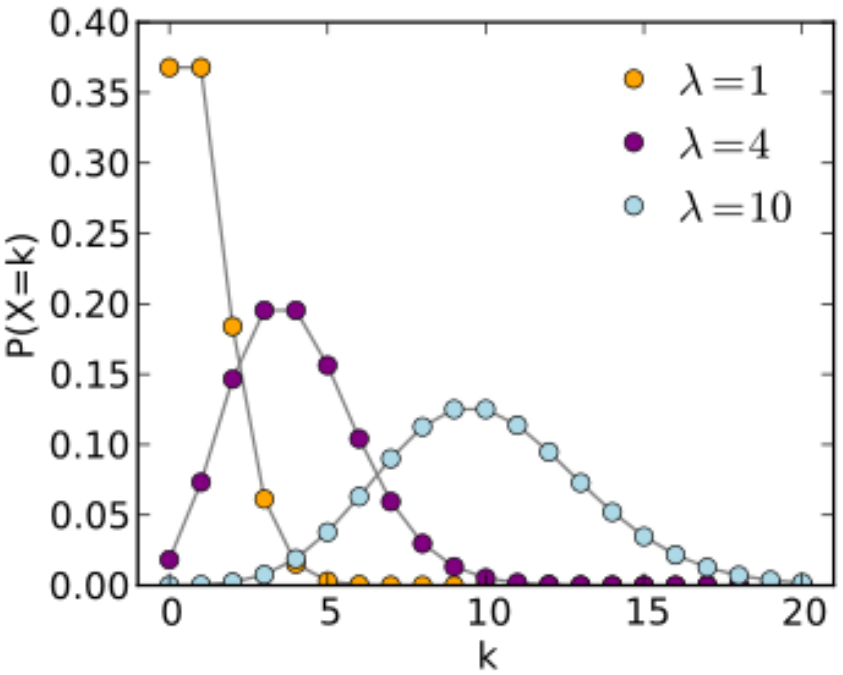
$$P_p(n \mid N) = \binom{N}{n} p^n q^{N-n}$$
$$= \frac{N!}{n!\,(N-n)!} p^n (1-p)^{N-n},$$

Poisson Distribution (Discrete)

Independent Events with a known average rate

Common Example

- The number of meteorites greater than 1 meter diameter that strike Earth in a year
- The number of patients arriving in an emergency room between 10 and 11 pm



$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

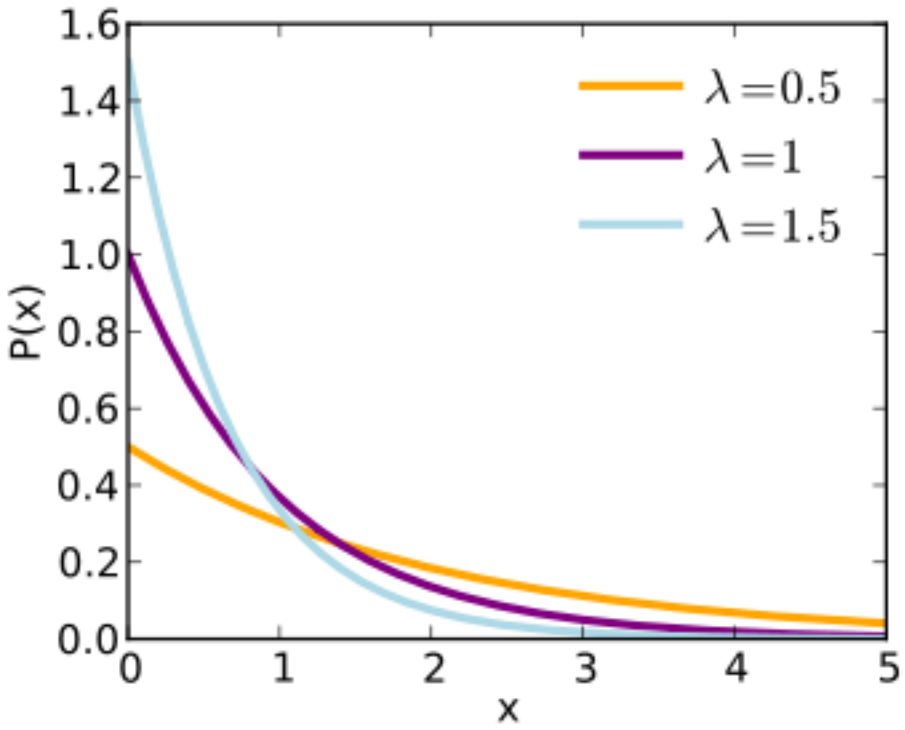
Where
Lambda is the avg number of events per interval
K = number of trials 0,1,2,3,…

Exponential Distribution (Continuous)

- Shows decaying probability

Common Example:

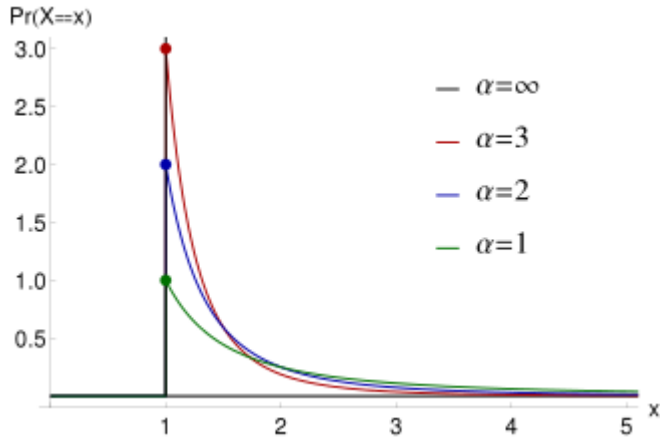
- Time until the next call at a call center



The probability density function (pdf) of an exponential distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Pareto Distribution ("80/20")



- Special type of **Power Law distribution**

Common Examples

- Wealth distribution
- Size of meteorites

Log Normal Distribution (Continuous)

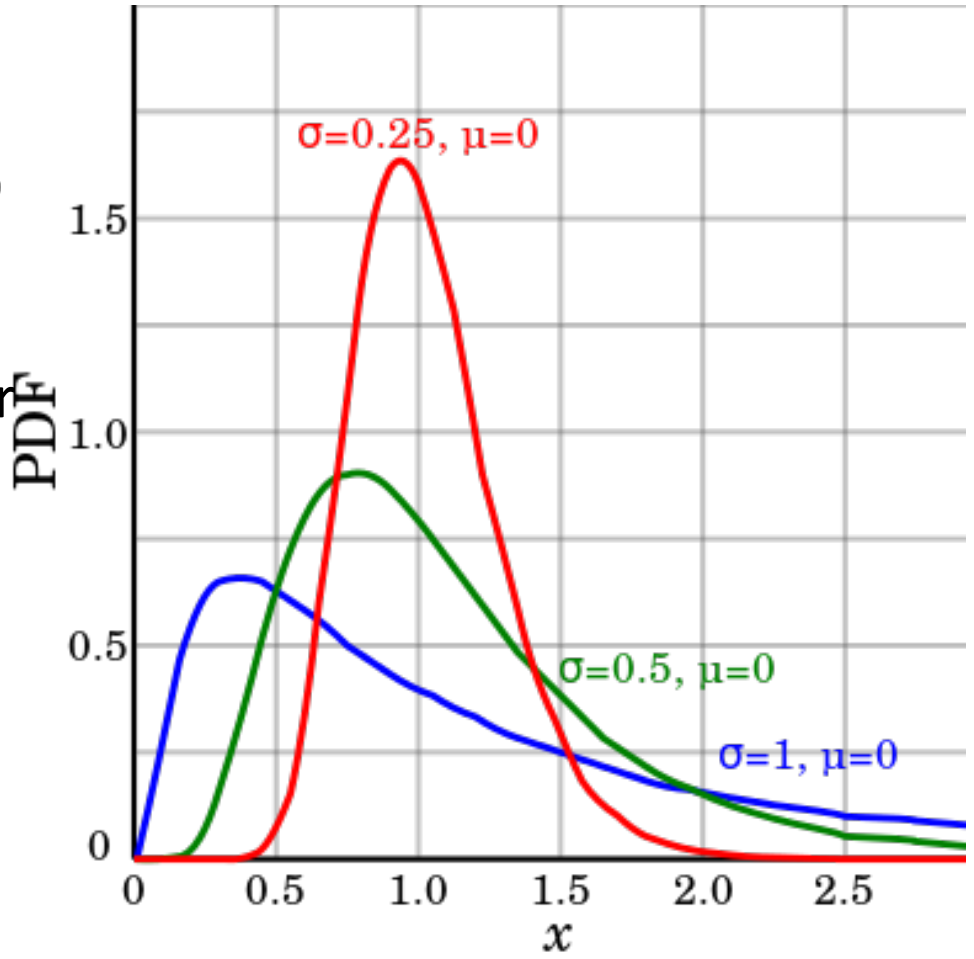
Variable whose log is normal distributed

- If X is Log Normal, Y=ln(X) is Normal

Common Example: (wherever there is skewed)

- Amount of rainfall
- Repair times

Next common to Normal



Distributions in Excel

Excel has many functions available to work with distributions. Using them, you can determine the likelihood of getting:

- exactly a value
- less than or equal to a value
- greater or equal to a value

Distributions in Excel

As an example, consider the normal distribution of the heights of men in the US…

We will use the normal distribution and the Excel function

NORM.DIST(x, mean ,std_dev, cumulative)

Distributions in Excel

What's the probability of measuring a random man **at least 67 inches**

NORM.DIST(67, 70, 3, **TRUE**)

What's the probability of measuring a random man **at exactly 67 inches**

NORM.DIST(67, 70, 3, **FALSE**)

Distributions in Excel

Normal is not the only one. Other distributions are available.

- BINOM
- POISSON
- LOGNORM
- T
- CHI

Poisson Distribution in Excel

CLASSROOM WORK

30 minutes

Use “Call Center Analysis.xlsx” . 5000 log measured for number of calls per hour.

A call center manager has decided to motivate the employees by offering a cash incentive for handling a higher call-per-hour at the center. In order to determine what the incentive number (calls per hour) should be.

>> The analyst needs to profile the current performance of the floor and the manager needs to evaluate the likelihood of achieving certain milestones in order to set a high bar which is still accessible by people on the floor.

1-POISSON.DIST(6,Mean,TRUE)

Find the following

- Average, STDEV, Min, Max, Frequency Plot
- P(calls/hour > 6)
- P(calls/hour > 8)

- # Day 2

Fitting Data

We discuss how to fit data to data sets and discuss how these fits become our models

Section 7

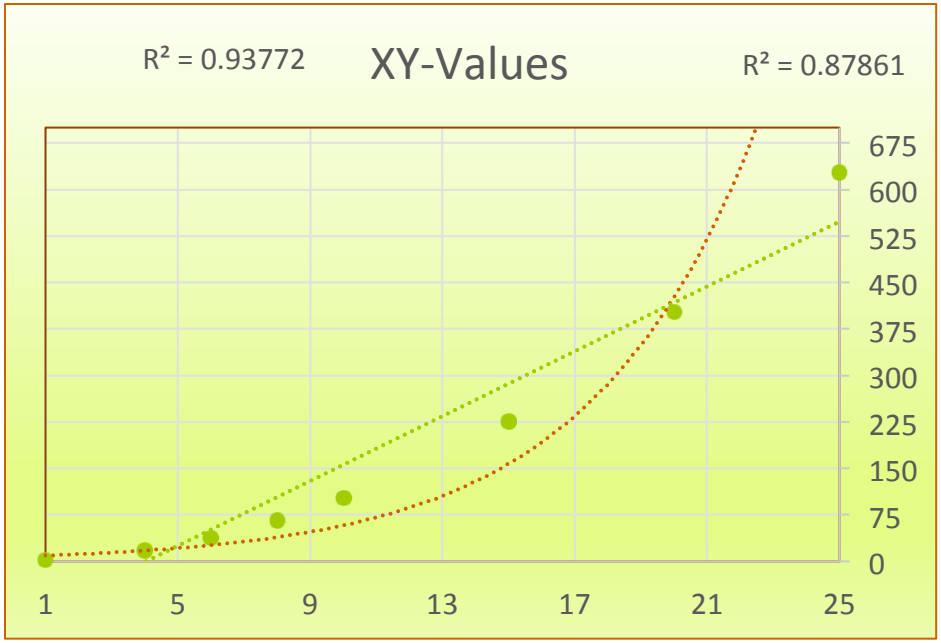
Bivariate Data

- Scatter plots
- Covariance and Correlation
- Regression
- R^2

Bivariate data – Scatter Plot

- Bivariate – data has 2 attributes (X, Y)
- Relation between X Y ?

X-Values	Y-Values
1	3
4	18
6	38
8	66
10	102
15	227
20	402
25	627



Covariance and Correlation

Covariance = $E[(X-E(X))(Y-E(Y))]$

- Units are Units of X time Units of Y
- Not normalized
- May be hard to interpret

Correlation = $Cov/(\sigma(X)\sigma(Y))$ **--- Important !!**

- Normalized by standard deviations
- Unit less
- Between -1 and 1
- How X and Y are related, Corr(x,y) = 0.9 is highly correlated

Correlation in Excel

	Correlation part 1			Correlation part 2			Correlation part 3	
	x1	y1		x2	y2		x3	y3
	1	3		1	2		1	12
	2	7		2	4		2	11
	3	11		3	6		3	9
	4	12		4	8		4	7
	5	8		5	12		5	5
CORREL	0.665517382			0.986393924			-0.993883735	

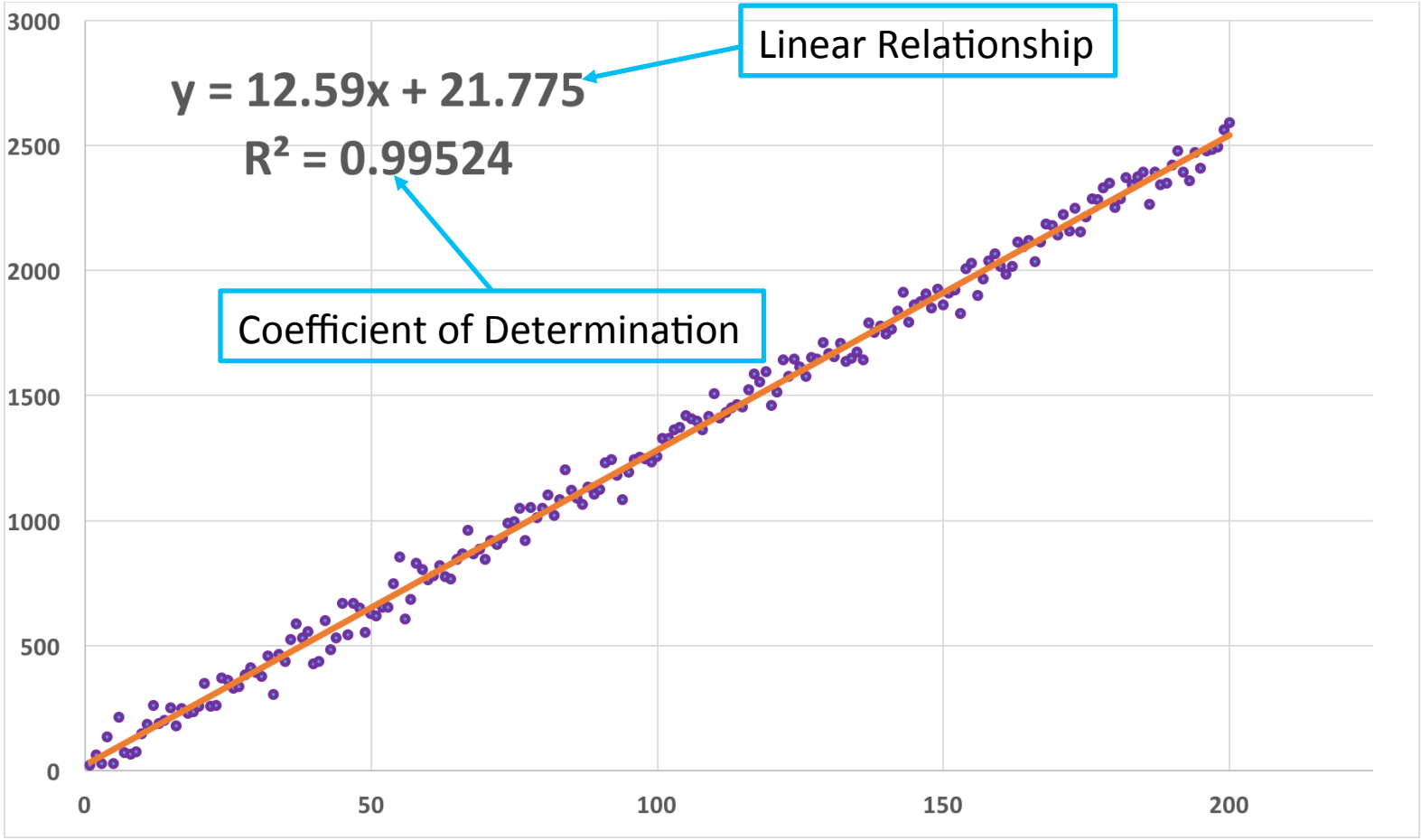
EXCEL FUNCTION: **=CORREL(arr1, arr2)**

Simple Linear Regression

- Finding a linear relationship which can predict the (average) value of Y for a given value of X.
  - Y = (mx + c)
  - Fit a line with given values of (X,Y), so that Y can be estimated using any value of x

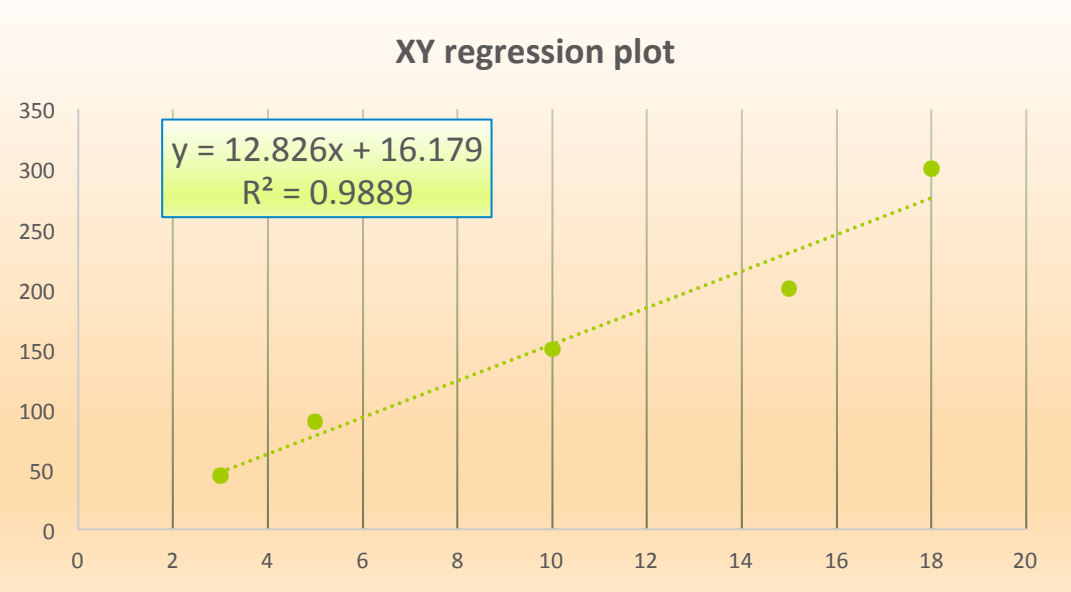
- Coefficient of Determination (R^2)
  - is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
  - Simply – this is the average Euclidean distance between (xi, yi) from the fitted line

Linear Regression



Regression in Excel

X	Y
10	150
5	90
3	45
15	200
18	250



CORREL	0.978669933

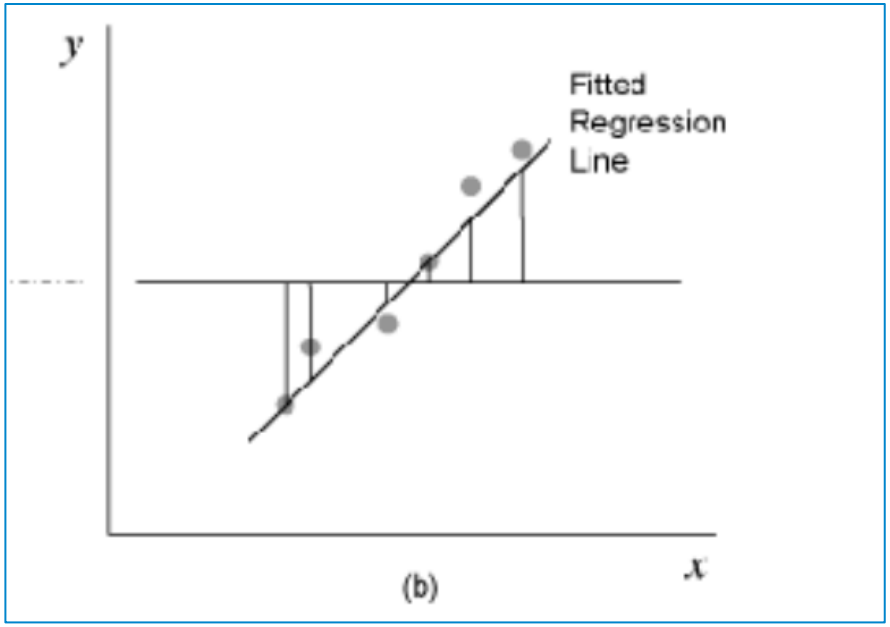
How to:

- Plot a scatter plot
- Add a trend line (LINEAR)
- Add R^2 co-efficient

Observation:-

- X and Y are monotonic increasing function that can be verified by the equation on the chart
- Corr(X, Y) is very high

R^2



Squared sum of vertical distance between the points and the fitted line

Higher the value, better the fit is

Square_Root{
(d1^2 + d2^2 + … + dn^2)
} / n

Linear Regression

Beware: Some people try to “force” a linear relationship where one does not exist. Plotting linear regression is only a valid tool IF:

- The scatterplot forms a linear pattern (you must use a scatterplot to detect a linear relationship in the first place)
- The correlation, r, is moderate to strong (typically beyond 0.50 or –0.50).

Other Fitting Functions

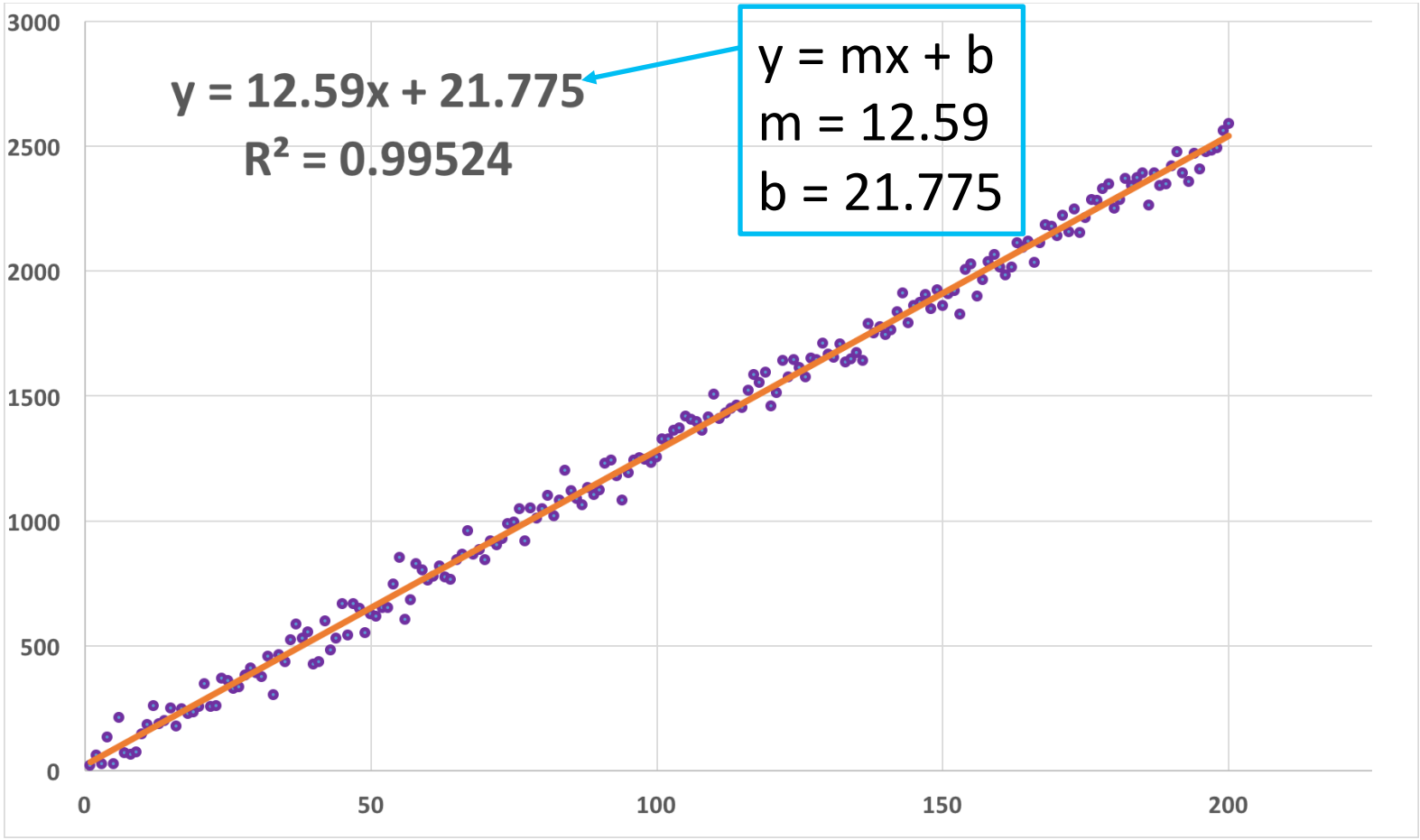
- Linear – $y = mx + b$
- Polynomial - $y = a + b*x + c*x^2 + \ldots$
- Power Law - $y = a*x^b$
- Exponential – $y = a*b^x$ $(a*e^x)$
- Sinusoidal – $y = a*\sin(b*x + c)$
- Logarithmic, etc.

Linear Fit

- Simplest fitting
- Everything is linear if inspected closely enough
- m is the slope; b is the y-intercept

$$y = mx + b$$

Linear Fit

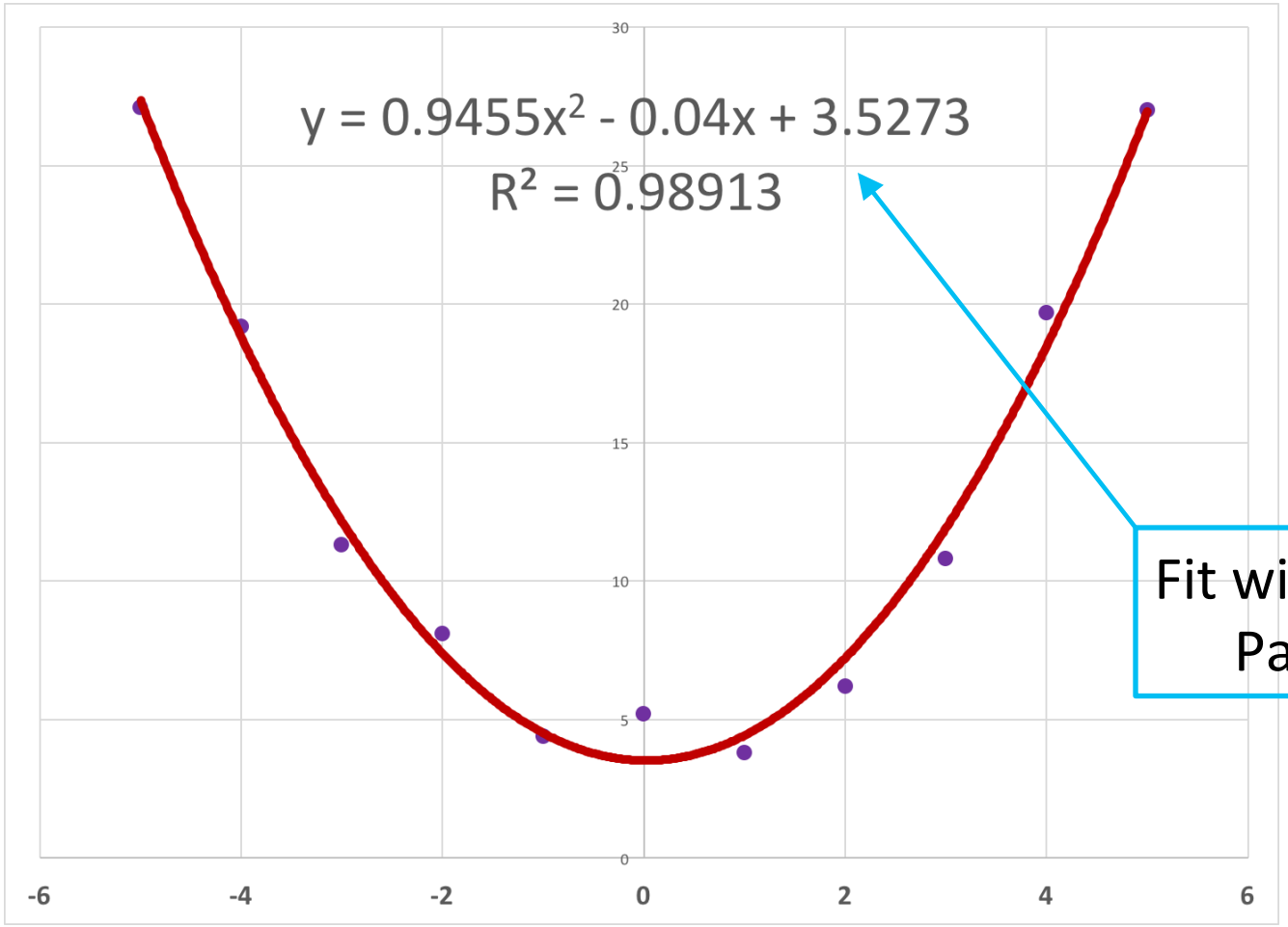

y = 12.59x + 21.775
R² = 0.99524
y = mx + b
m = 12.59
b = 21.775
3000
2500
2000
1500
1000
500
0
0
50
100
150
200

Polynomial Fit

- Can fit anything with enough parameters
- Be careful with extrapolation

$$y = a + bx + cx^2 + dx^3 \ldots$$

Polynomial Fit



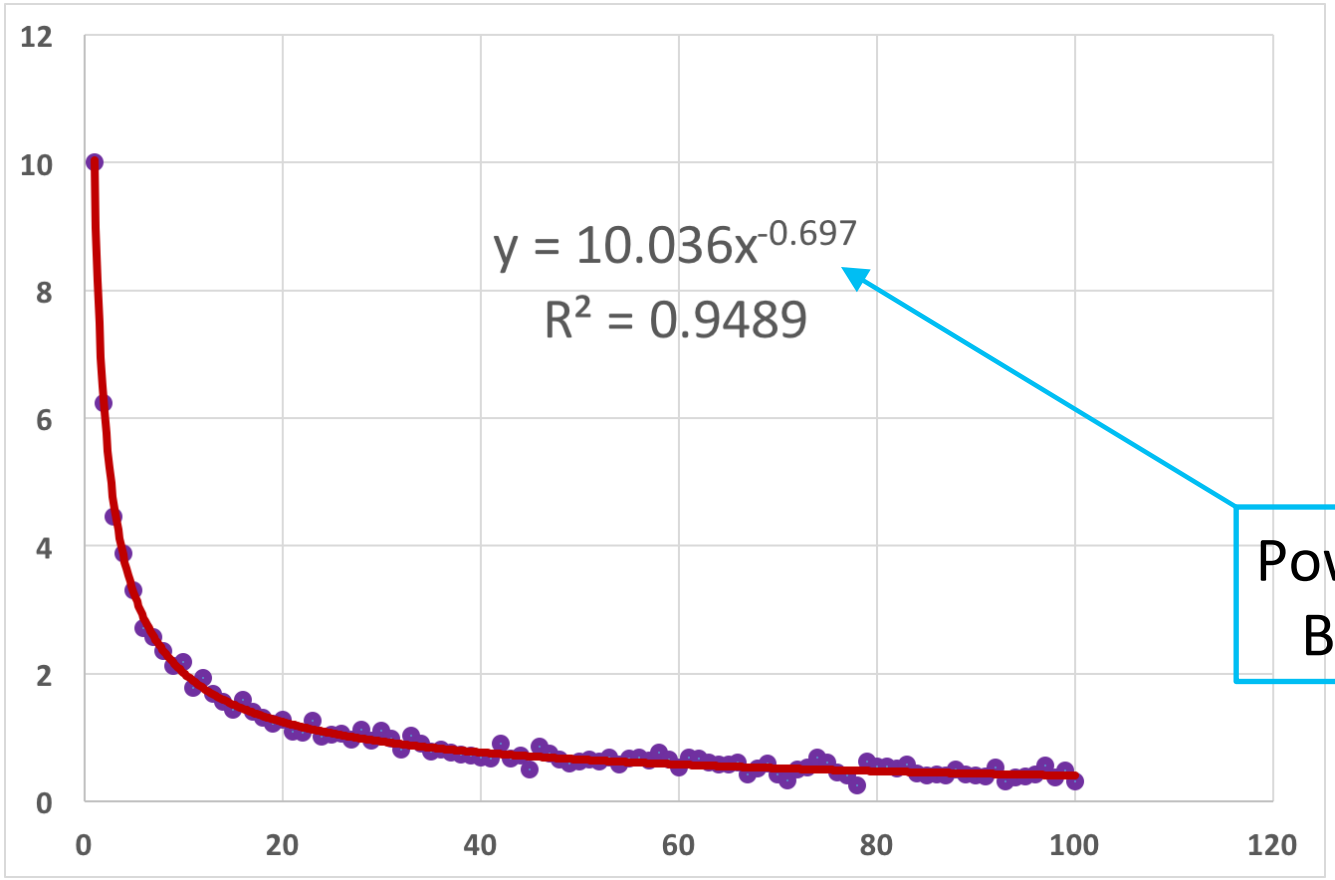
Fit with order 2 Parabolic

Power-Law Fit

- Income counts histogram
- Frequency of words in Text

$$y = Ax^{\uparrow}B$$

Power Law Fit

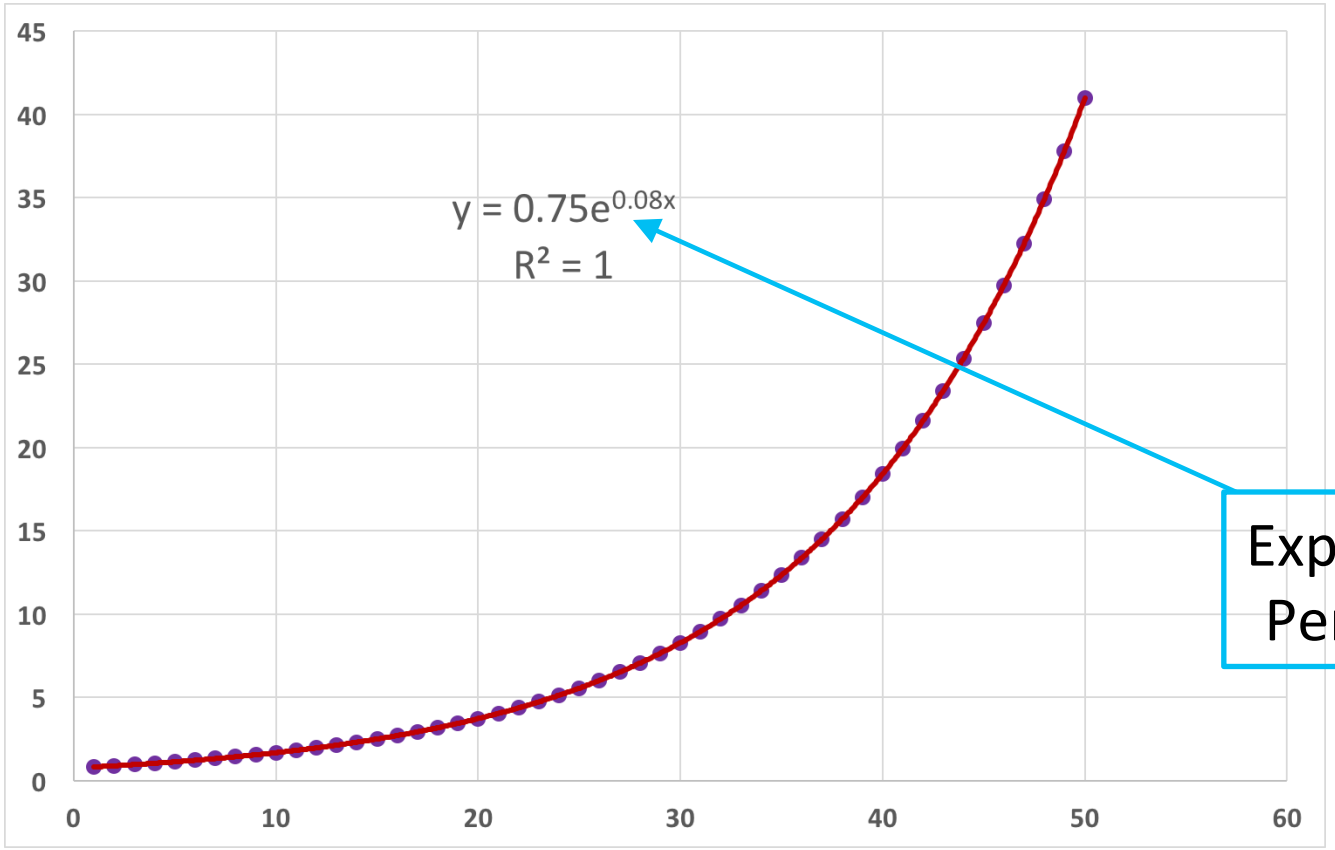


Exponential Fit

- Continuously Compounded Interest
- Radioactive Decay (Half-life)
- Amount of a drug in the blood

$$y=Ae^{Bx}$$

Exponential Fit



Data Fitting

- **Choosing which fitting function to use should be based on an underlying understanding of the data.**
- **Extrapolation beyond the bounds of the data must be done with care.**

Data Fitting in Excel

CLASSROOM WORK

45 minutes

Use "**Customer Retention.xlsx**"

- Plot scatter plot
- Add trendline – Linear, polynomial, exponential
- Check whose R^2 is least
- determine what kind of fit is best

Repeat the same process for "**Candy factory.xlsx**", "**Nurf Foam Factory.xlsx**"

Data Fitting in Excel

DISCUSSION

15 minutes

"**Customer Retention.xlsx**" – Should be exponential

"**Candy Factory.xlsx**" – Should be linear

"**Nurf Foam Factory.xlsx**" – Should be Polynomial

Cluster Analysis

We look at another way of drawing conclusions from visual data

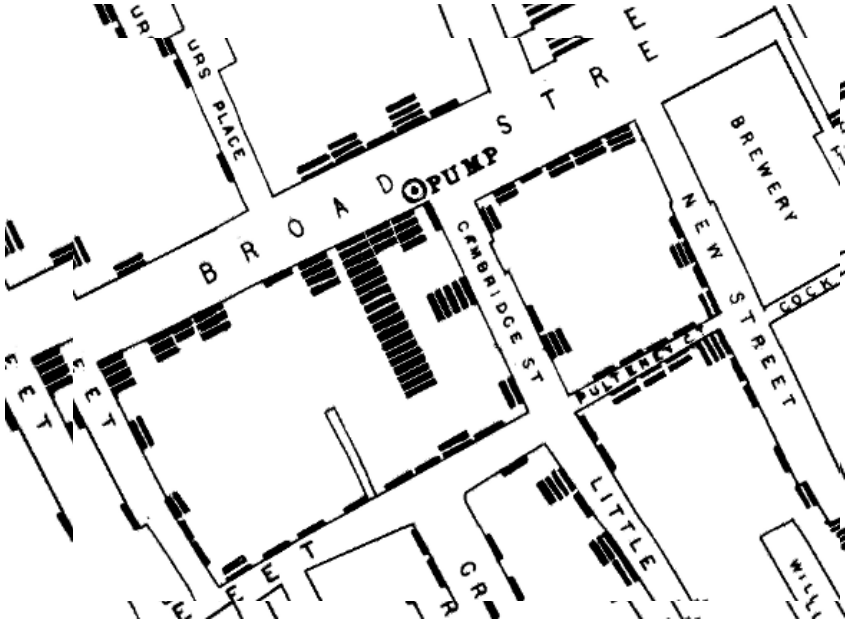
Section 8

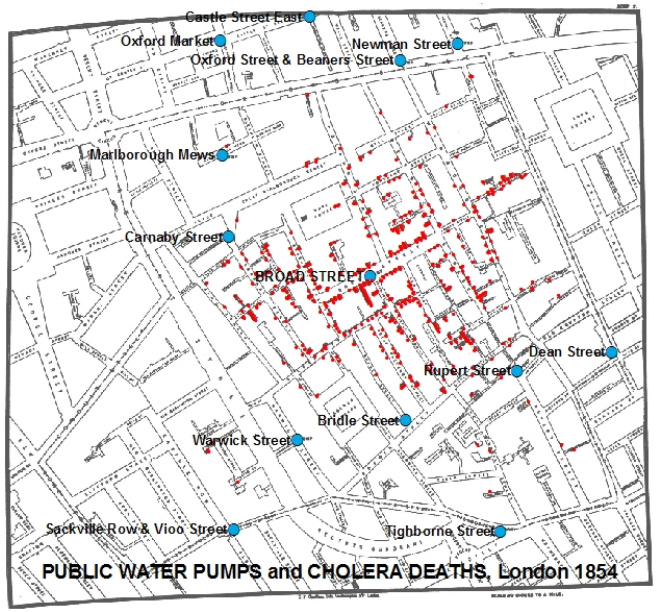
Clustering

- Finding common groupings in a multi-dimensional dataset (segmentation)
- Can be used to find patterns that are difficult to identify by eye
- Known as unsupervised learning in machine learning

Cholera in London

- An early example of clustering
- John Snow found the cause of an outbreak and help prove the germ theory of disease

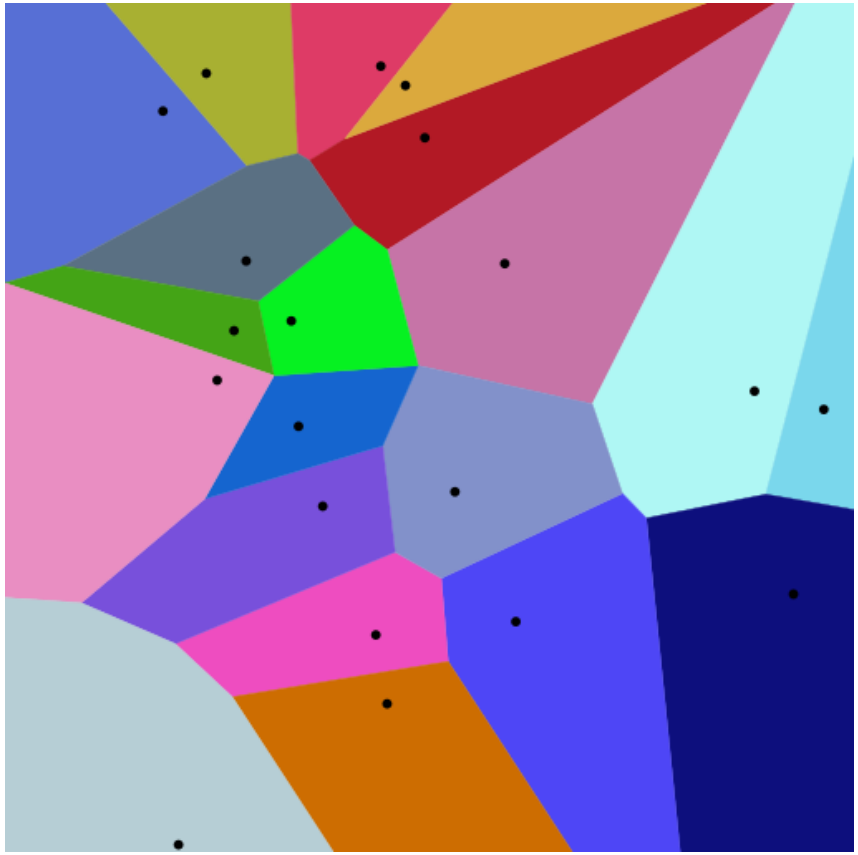




https://www1.udel.edu/johnmack/frec682/cholera/cholera2.html

Voronoi Diagrams

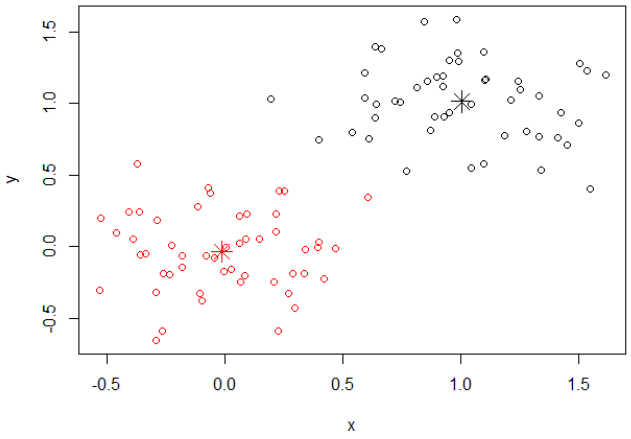
- Show the group of points that are closest to certain centers
- Centers can be chosen by hand, or determined by and algorithm (e.g. K-means)

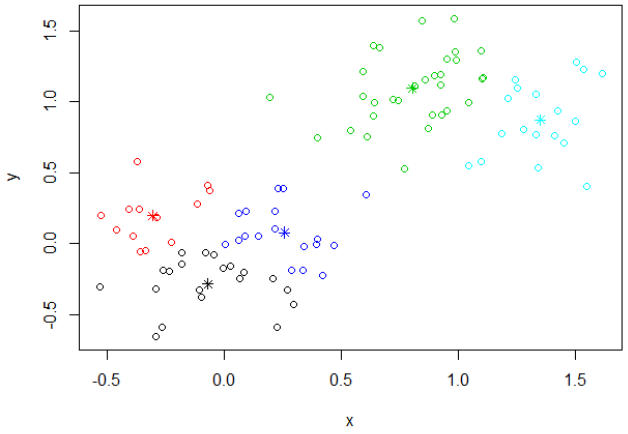


https://en.wikipedia.org/wiki/Voronoi_diagram

K-Means

- Algorithm that iteratively moves center points to define clusters
- Initial centers are either chosen or placed randomly





Clustering in Python

- http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

- http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py

Monte Carlo

We discuss a few things that take us into advanced prediction

Section 9

Monte Carlo Method

- Random Inputs into Models to Sample possible Outcomes
- Works on models that are too complicated to calculate analytically
- Used routinely in Business, Finance, Science, and elsewhere

Monte Carlo Method

How does it work?

- **Build a model**
- **Use a random number generator (RAND() in Excel) for input to the model**
- **Do this thousands, if not millions, of times to build a probability density function**
- **Query the PDF to find answers to questions**

Monte Carlo Method

- Data set is 300 previous jobs from DB records

- Analysis shows
  - Average = 20 days
  - Standard Dev. = 3 days
  - Distribution = Normal

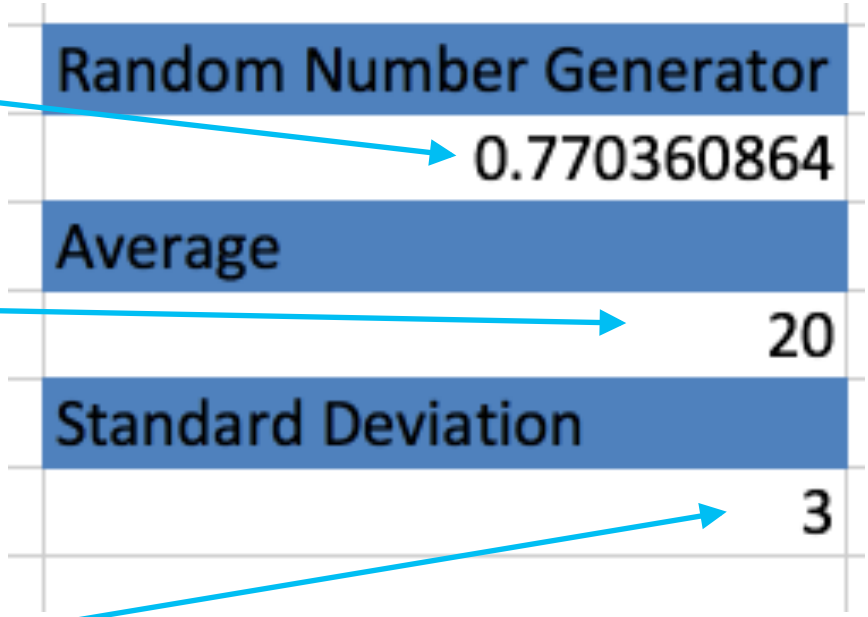
Job Number	Days To Deliver
1	18
2	20
3	22
4	17
5	20
6	19
7	20
8	19
9	15
10	23
11	21
12	20

Monte Carlo

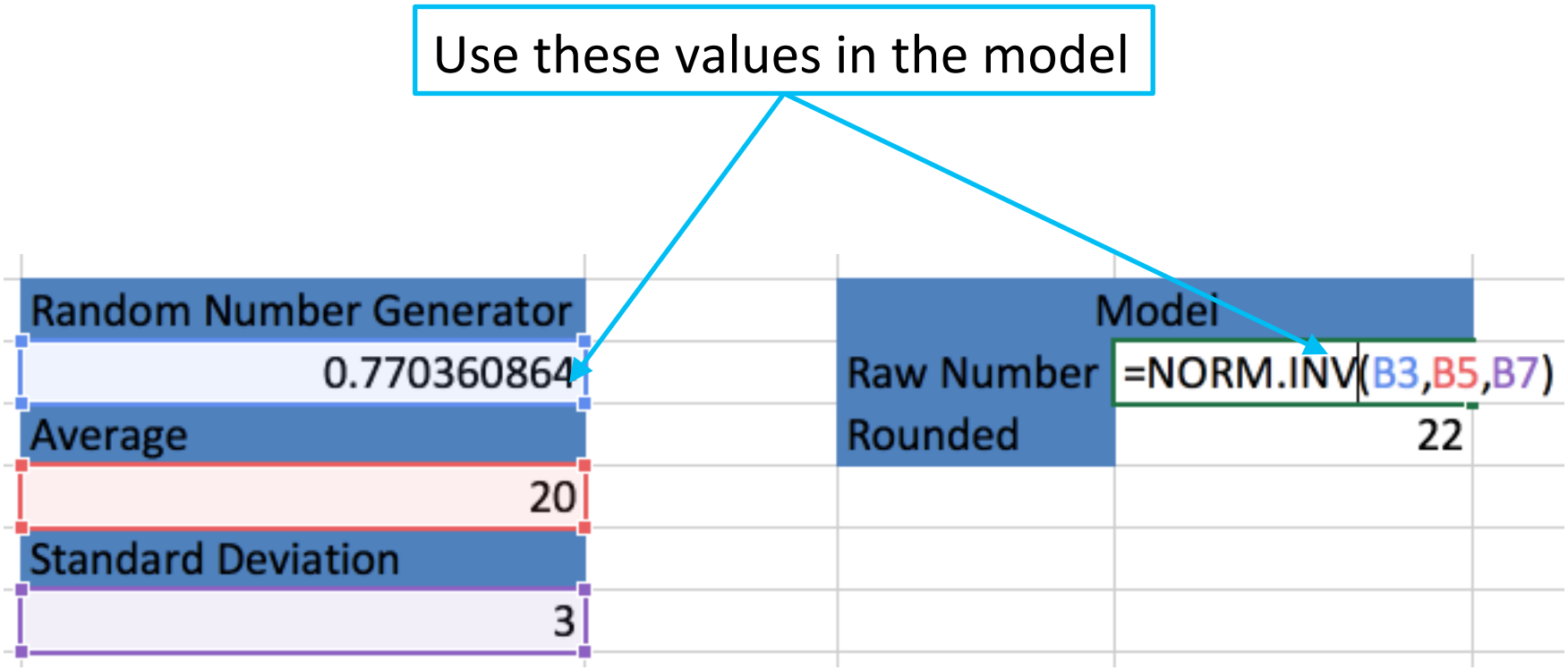
Make call to RAND()

Fill in the Average from the data analysis

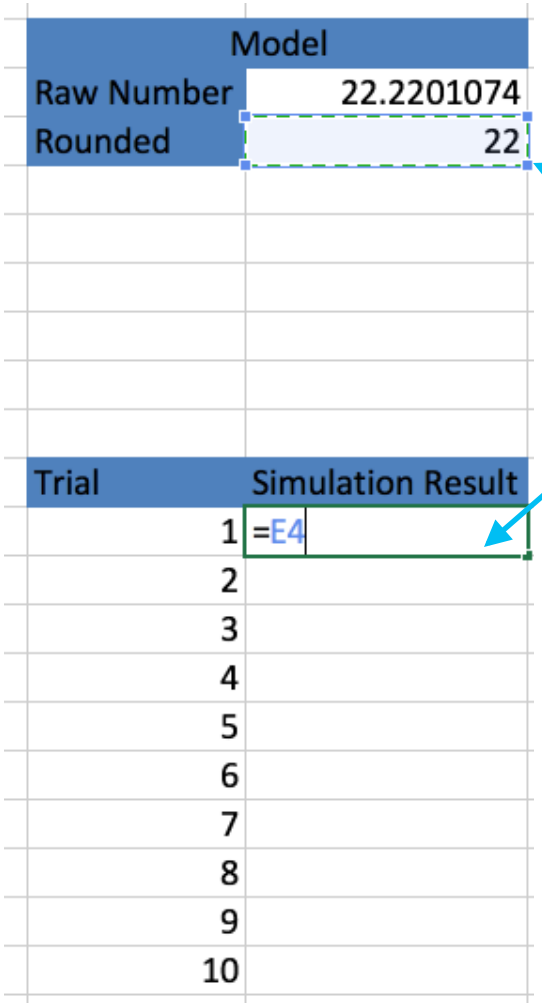
Fill in the Std Dev from the data analysis



Monte Carlo


Use these values in the model
Random Number Generator
0.770360864
Average
20
Standard Deviation
3
Model
Raw Number
=NORM.INV(B3,B5,B7)
Rounded
22

Monte Carlo

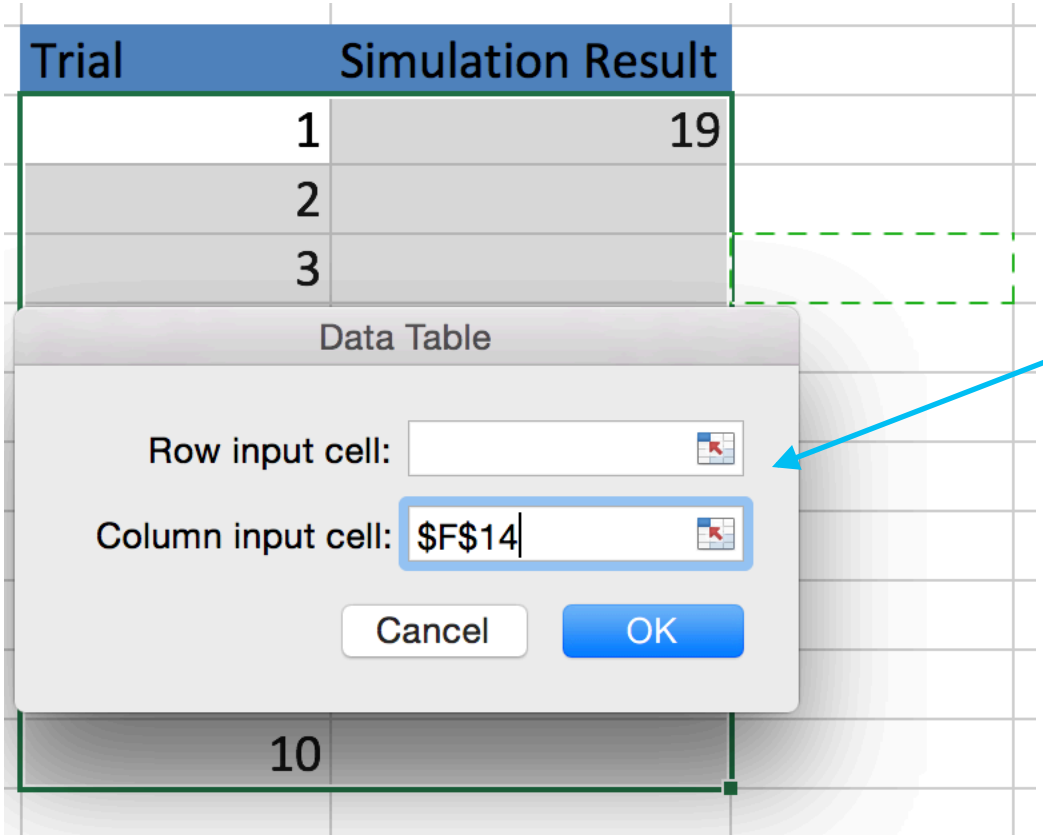


Model	
Raw Number	22.2201074
Rounded	22

Trial	Simulation Result
1	=E4
2	
3	
4	
5	
6	
7	
8	
9	
10	

Build a simulation table and reference the model

Monte Carlo



Create a What If? Data Table

Monte Carlo

Now there are 10 independent trials which we can use as a distribution

Trial	Simulation Result
1	18
2	21
3	19
4	20
5	18
6	22
7	16
8	18
9	17
10	21

Monte Carlo

Query the results by finding relative frequencies

Trial	Simulation Result
1	19
2	18
3	17
4	22
5	23
6	12
7	15
8	25
9	18
10	19

Questions	
P(Days > 23)	=COUNTIF(Simulation_Result, ">23")/10
P(Days > 21)	3
P(Days < 18)	3

Using COUNTIF() and a Named Range

Monte Carlo

CLASSROOM WORK

30 minutes

Use "**Concrete Slab Co Workbook.xlsx**"

- Find average, standard deviation
- Create a model –
  - =NORM.INV(RAND(), Mean, STDEV)
- Create N values from this model
- Find probability for
  - P(> 23) = ?
  - P(> 24) = ?

  Using =COUNTIF(RANGE,">23")/N

R Programming

We go back over the class material, this time in R

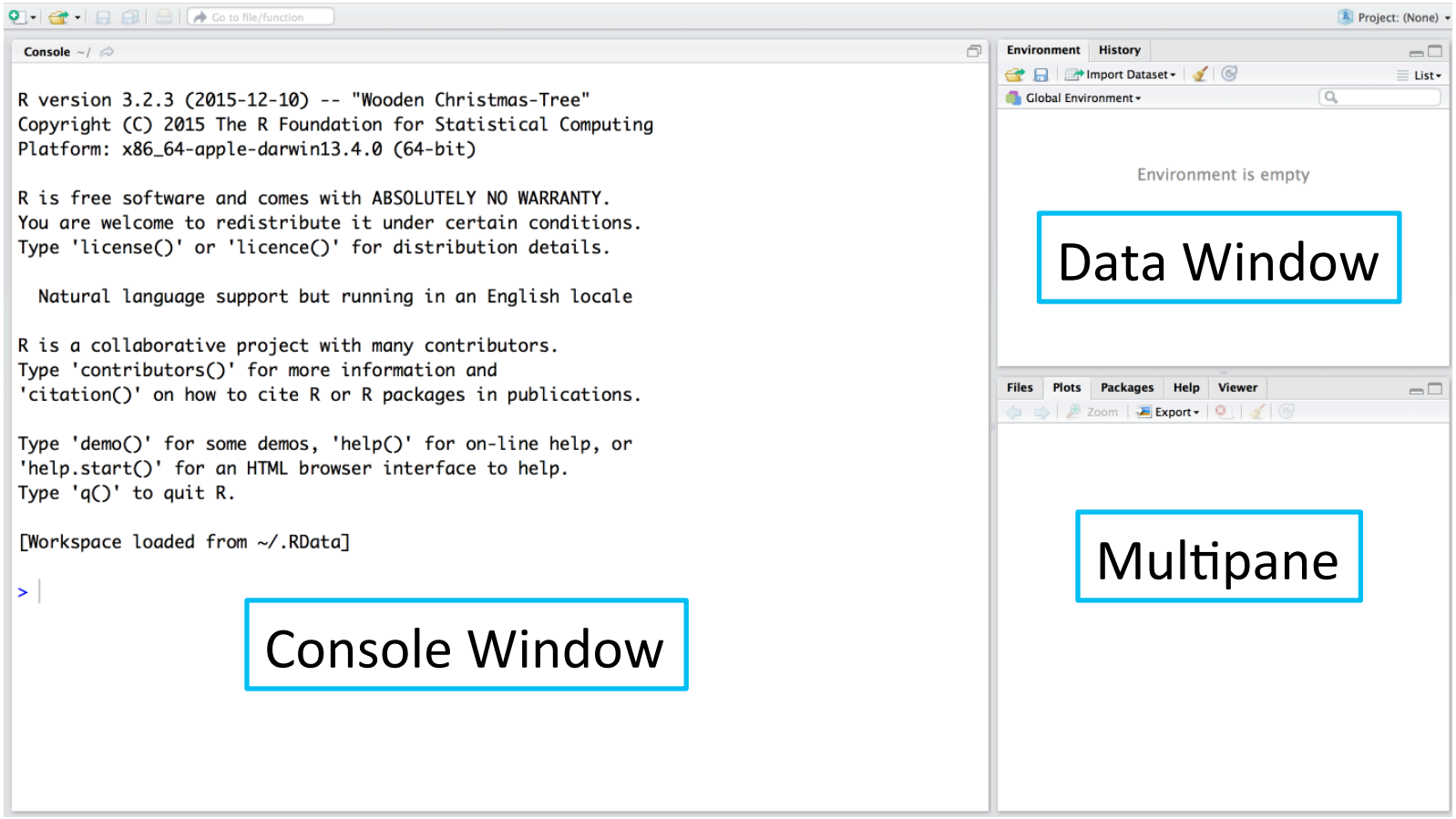
Section 10

What is R?

R is a programming language for

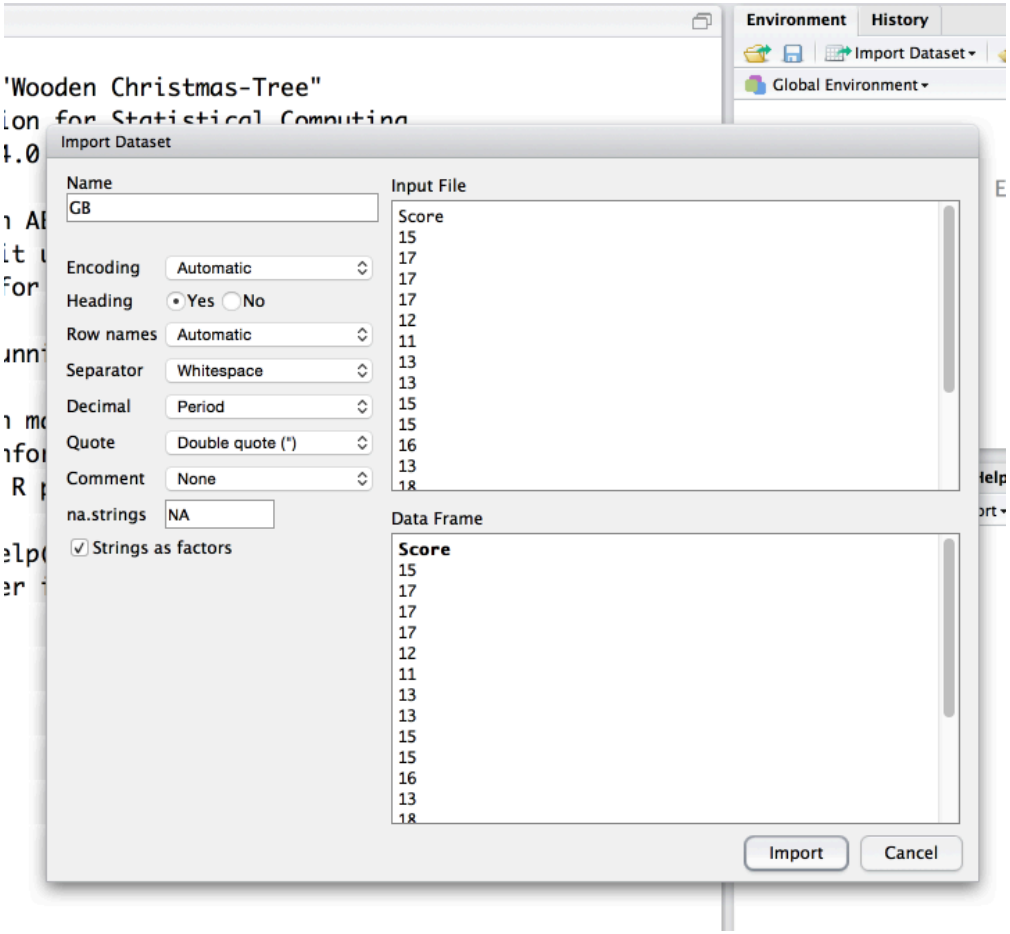
- **Statistical Computing**
- **Data Analysis**
- **Data Charting**
- **Data Modelling**

R Studio


Console ~/
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Console Window
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Data Window
Files
Plots
Packages
Help
Viewer
Zoom
Export
Multipane

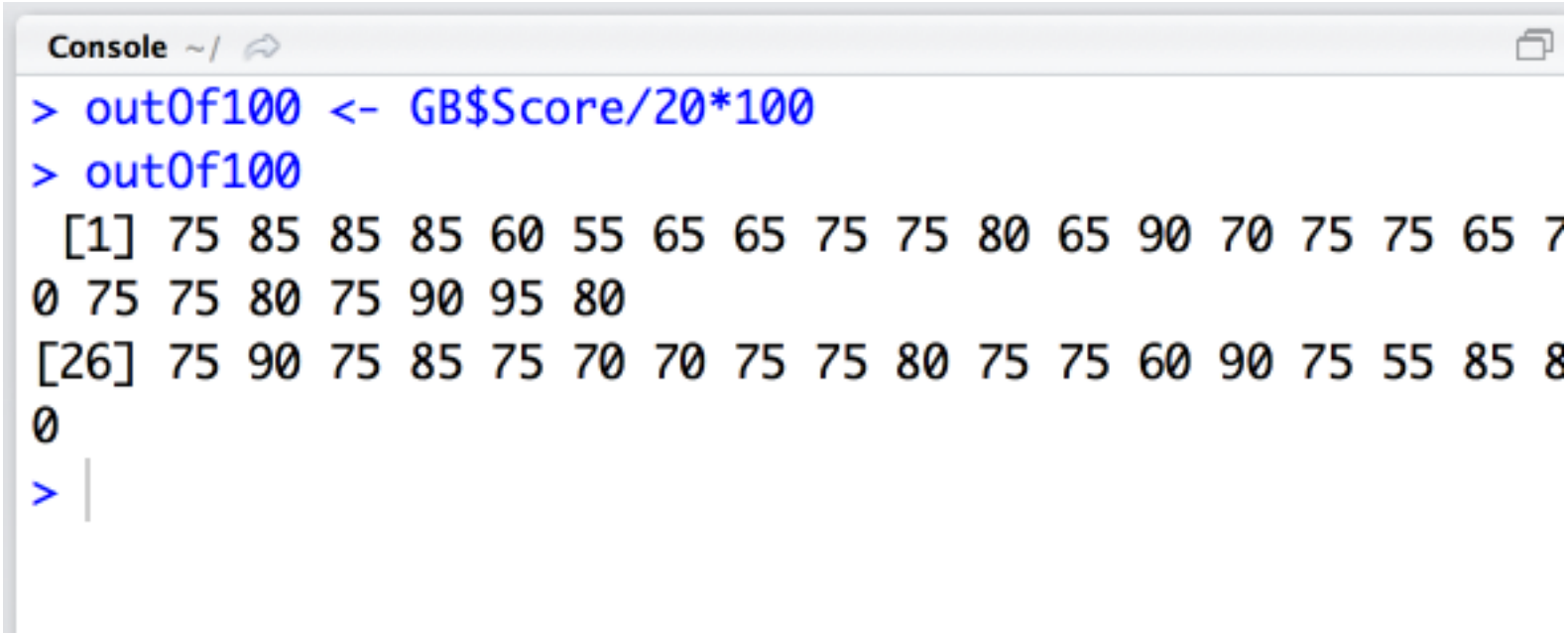
Basic Analysis in R

- Import Grades
- Give data set a name
- Will access with heading as variable



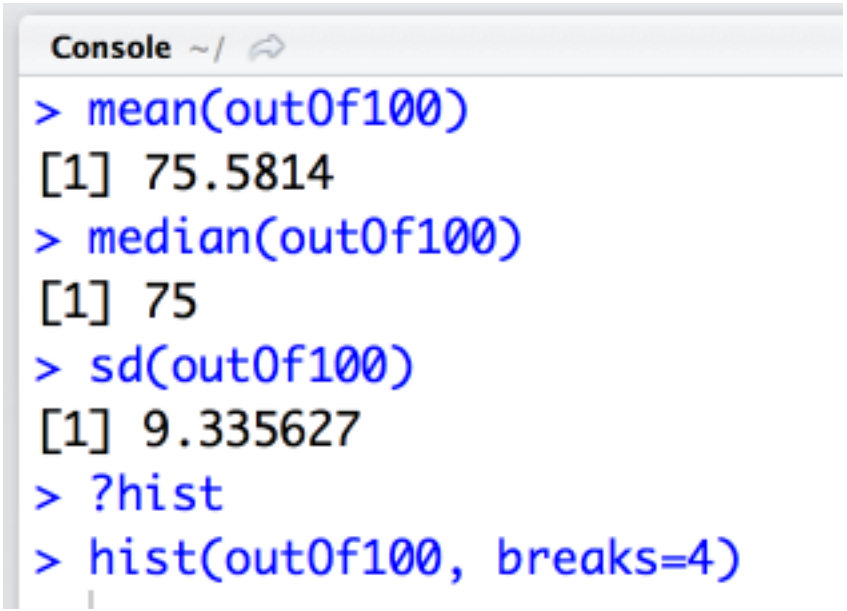
Basic Analysis in R

Convert scores to percentages
Converts all elements in data set at once

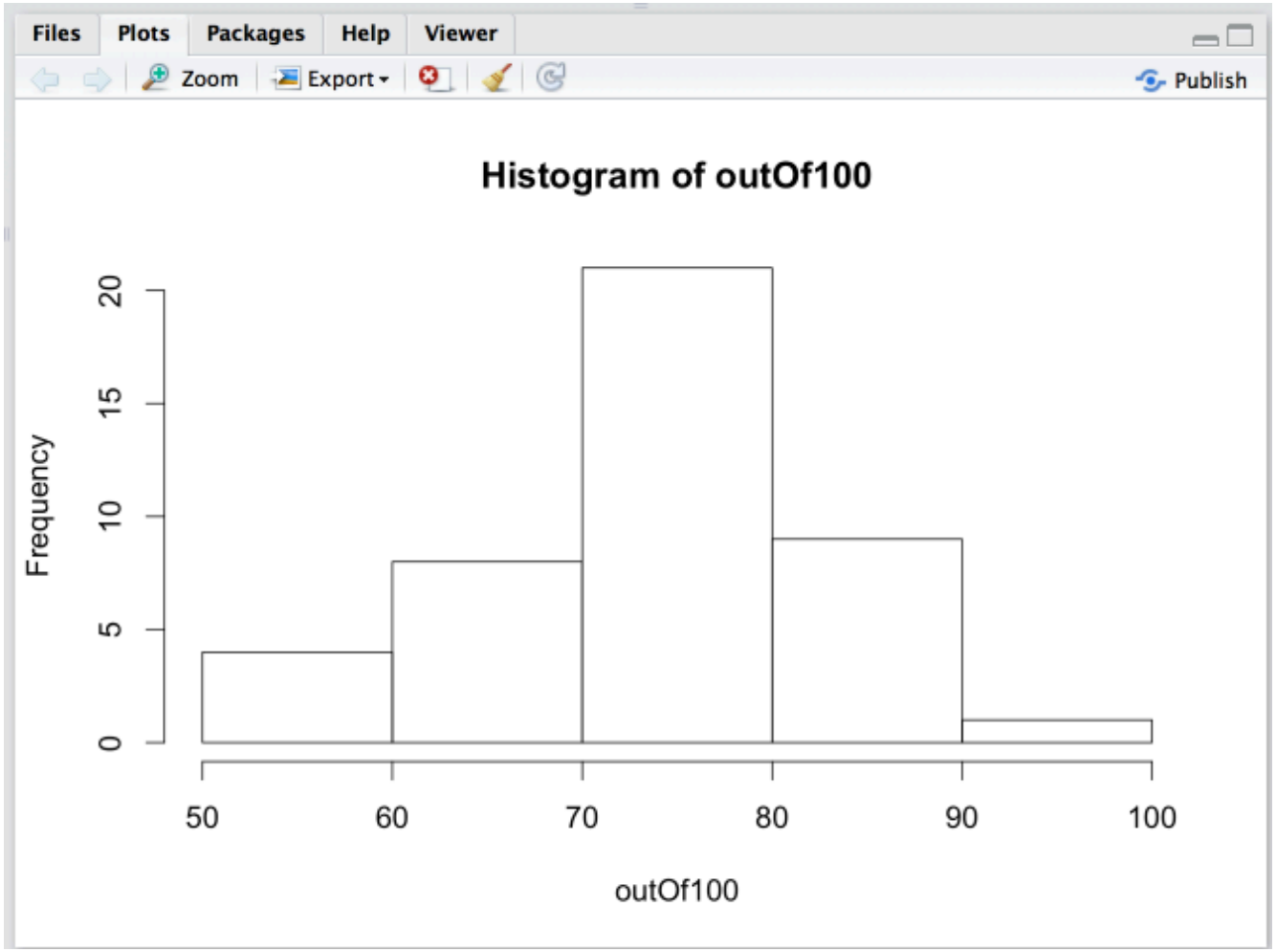


Basic Analysis in R

- Calculate central tendency
- Ask for help
- Plot histogram

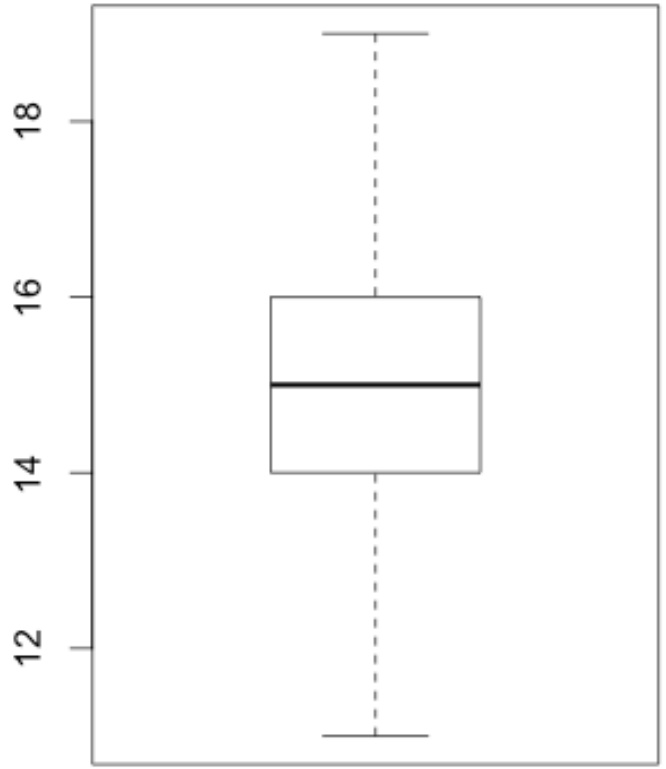


Basic Analysis in R



Basic Analysis in R

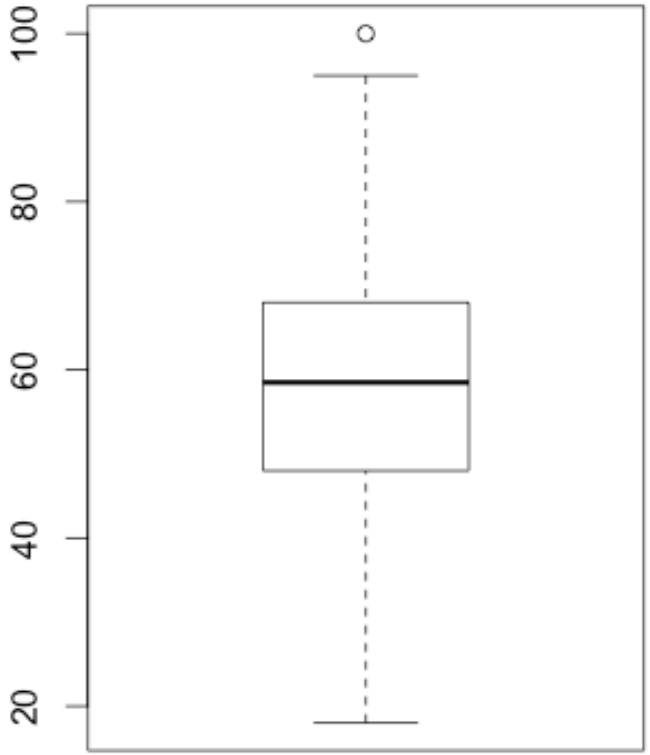
- Create a boxplot with boxplot(GB$Scores)



Basic Analysis in R

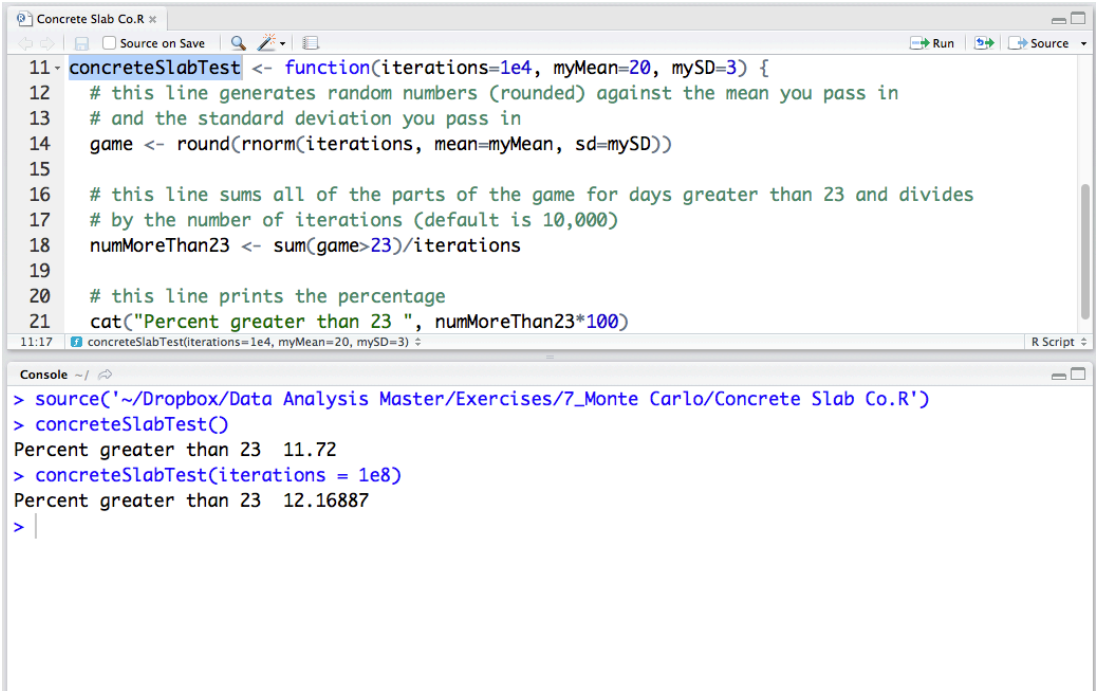
- A different set has an outlier

outlier(x, plot=TRUE, bad=5, na.rm=TRUE)



Monte Carlo in R

- Create a function
- Call the function multiple ways
- Note operations on arrays as variables



R Summary

- Open-source, well-supported tool for data analysis
- Often more efficient than Excel for repetitive tasks
- Numerous packages are available for complicated analysis techniques and data visualization

Data Visualization

We talk about how to tell the story

Section 11

Time to Tell the Story

A Few Simple Rules Go a Long Way

- Remember what you're accomplishing
- Really know your user
- Keep it simple
- Use space judiciously
- Focus on the design as much as the data

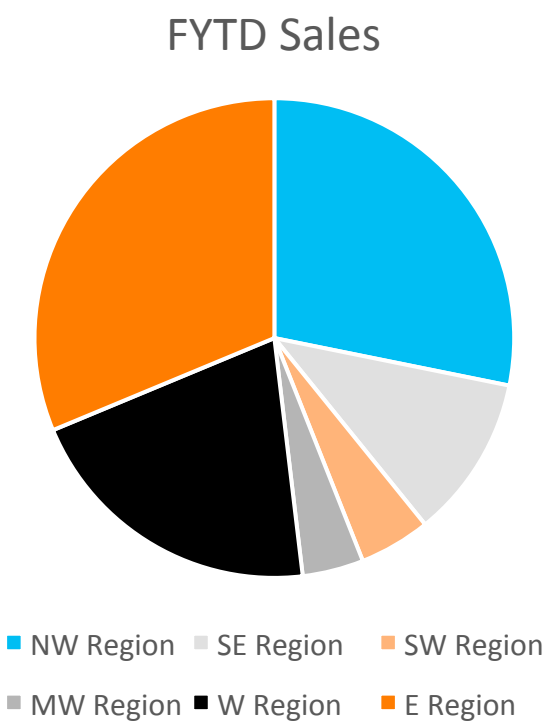
THE GOAL OF VISUALIZATION

What you're really trying to accomplish

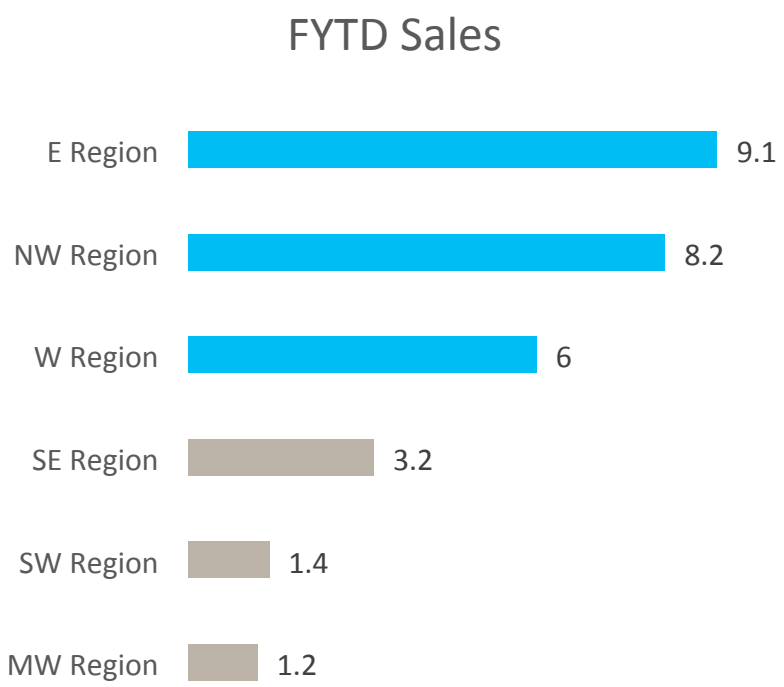
- **Efficiently communicate the right information to the right person**
- **Tell the story**
- **Critical Characteristics**
  - **Has a Beginning, Middle, End**
  - **Has a punchline**
  - **Doesn't lose the reader (stakeholder)**
- **Enable better decision making**

Two Tales of The Same Information

Common go-to Object



“Bolder,” more efficient choice



KNOW THE USER

- Understand what types of decisions are being made
- Timeliness of decisions affects design
- Exception vs. Total Picture
- Some people like more numbers

  …but you can still help them be more efficient

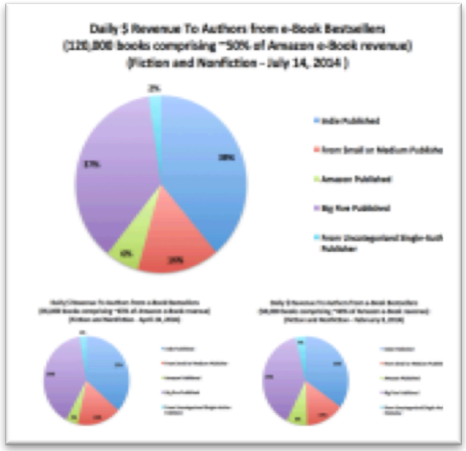
Stakeholder Cheat Sheets*

**Exceptions will apply*

THE VERIFIER

- Likes Detail
- Tends to Question
- Methodical
- Decision Maker
- Resists Change

BEST BETS



Stakeholder Cheat Sheets*

**Exceptions will apply*

THE BIG THINKER

- Shuts Down with Detail
- Tends to trust
- Moves first, Plans In-Route
- Thrives on Change

BEST BET

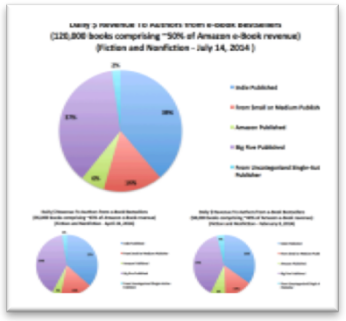
Stakeholder Cheat Sheets*

**Exceptions will apply*

THE INTERPRETER

- Moves quickly to change
- Stops to question logic
- Diverse
- Wants enough detail to feel comfortable

BEST BETS



Common Presentation Missteps

- Data chucking
- Overly graphic
- Lack of graphical insight
- “Just get it out” mentality
- Lacking clear message

Choosing the Format

- Based on Context
  - High-level, At-a-glance = Dashboard
  - Middle-level, Summarized = Report
  - Low-level, Aggregation-ready = Excel
- Based on Usage
  - Strategic Direction = Report
  - Tactical Course = Dashboard/Report
  - Analysis/Collaboration = Excel

Reports

- Mix of Graphs and Tabular
- Snapshot (cached and versioned)
- Interactive
  - Drill-down
  - Filter
  - Slice
- Single purpose
- Easier to create

Dashboards

- Best are purely Graphical
- Often real-time or close to real-time
- Some Interactive
  - Drill to detail
  - Some filtering
- Can be multi-use
- Harder to get right

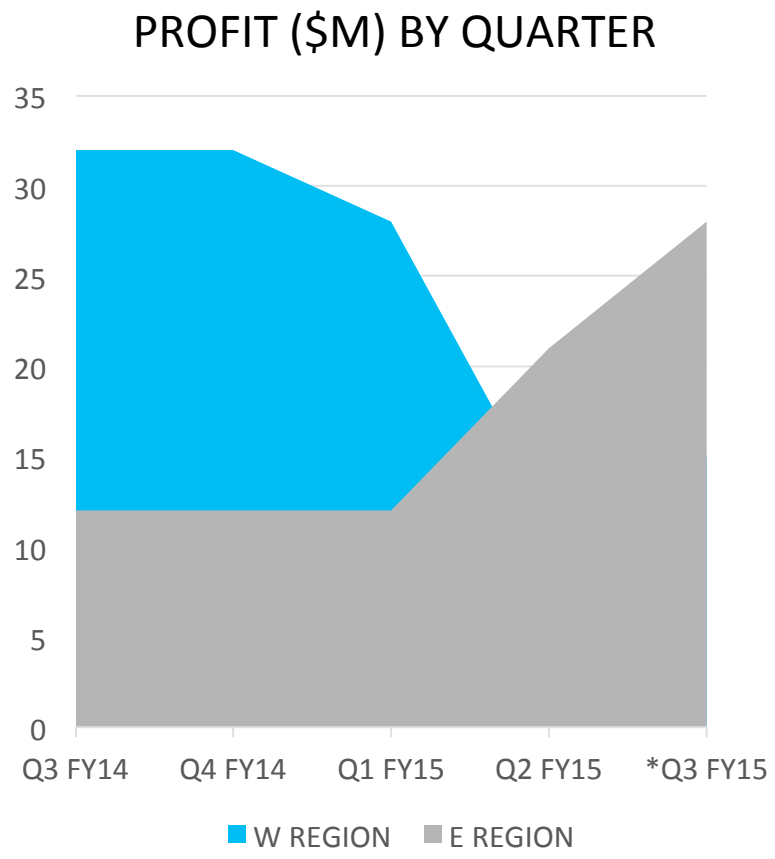
Excel

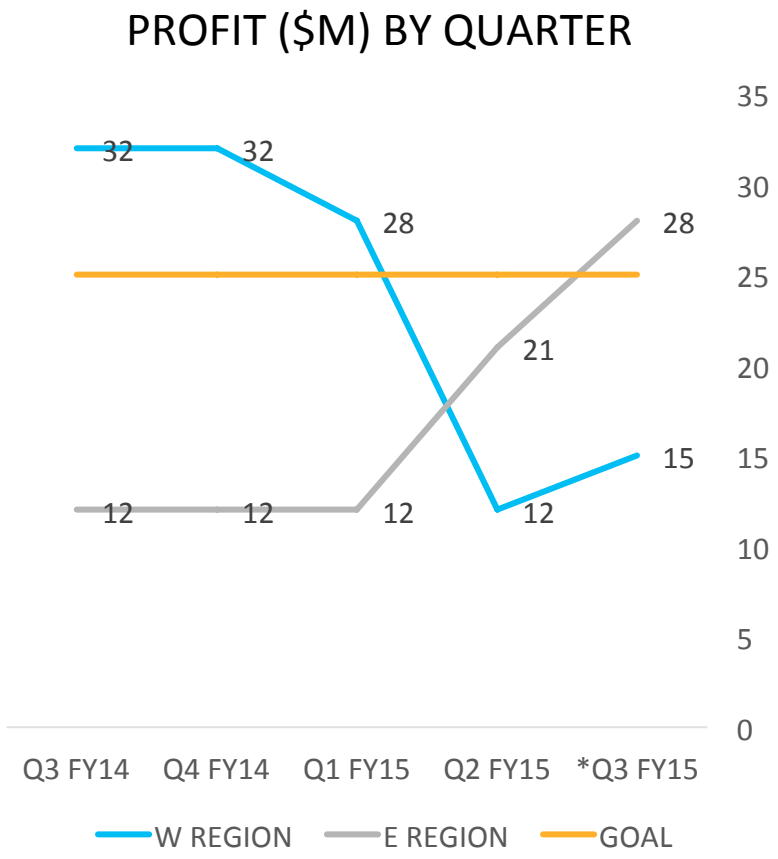
- Detail data
- Give a push with PivotTable or Chart
- Best include Dynamic Data

KEEP IT SIMPLE

- Ask yourself the question: “Will my user ‘get it’ in less than 5 seconds?”
- Classic chart types are classics for a reason
- Don’t go crazy with color
- Take a minimalist attitude
- Remember the goal

Which Graph Conveys Information Better?

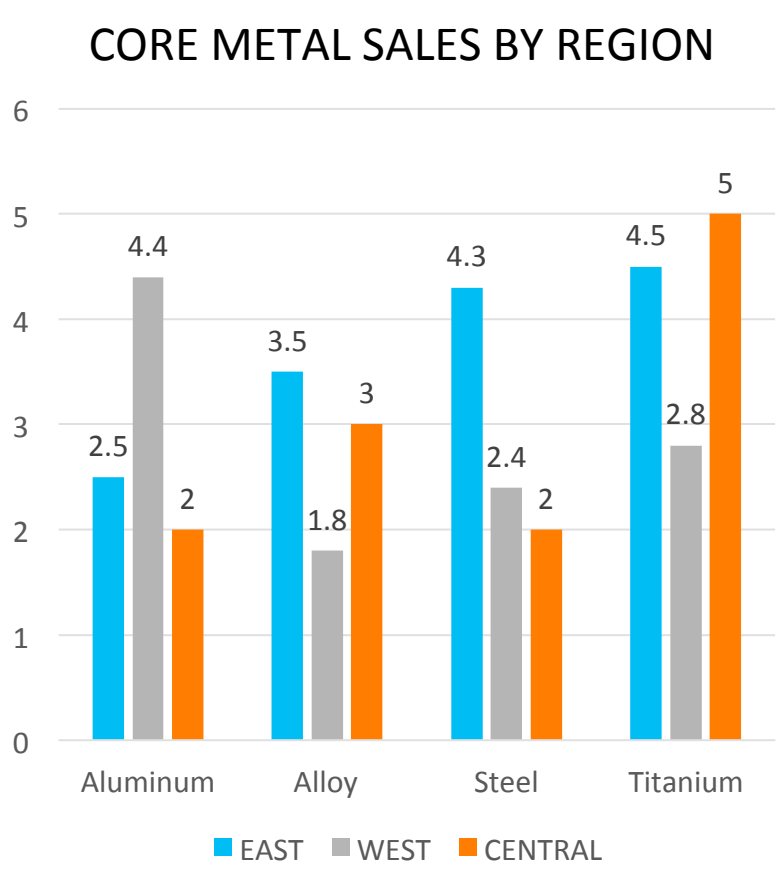


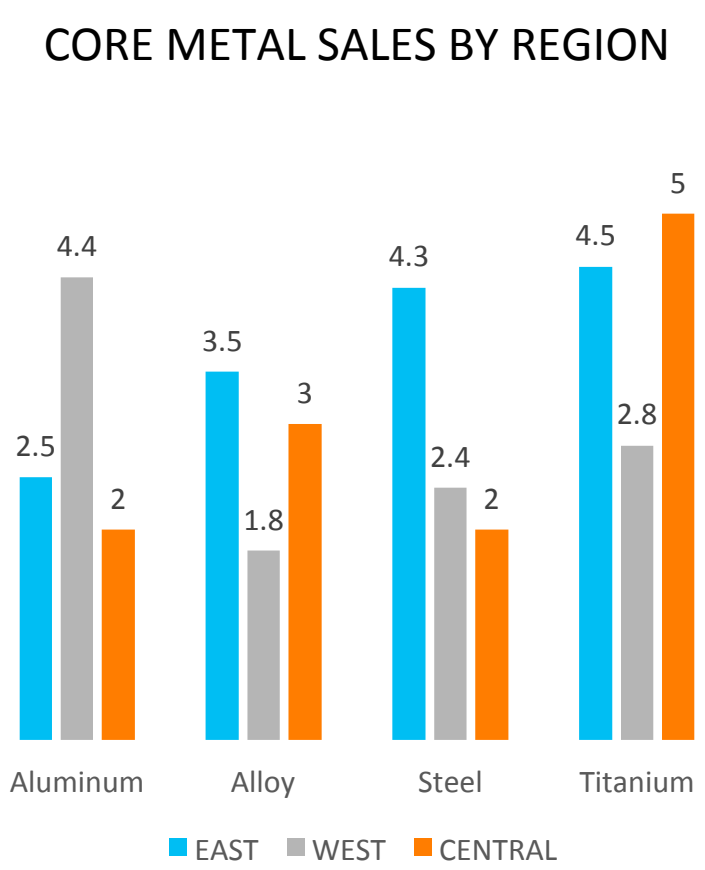


USE SPACE JUDICIOUSLY

- Reduce noise
- Continually ask, “How can I make it cleaner?”
- If it doesn’t help tell the story, it’s out

Zebras Don’t Belong in Graphs





Grid-based can be more challenging

COUNTRY	FYTD Revenue	Profit Margin	Consistency of Execution	Pipeline Growth
Germany	15.0	22.5%	85.0%	25.0%
UK	14.0	18.5%	75.0%	-15.0%
Denmark	8.0	20.0%	80.0%	10.0%
Belgium	5.0	22.0%	75.0%	2.5%
France	5.0	32.0%	92.0%	18.0%
Ireland	5.0	19.0%	64.0%	-16.0%
Norway	2.0	25.0%	98.0%	-25.0%

COUNTRY	Health Indicator	FYTD Revenue	Profit Margin	Consistency of Execution	Pipeline Growth
Germany	▲	15	23%	85%	25%
UK	▼	14	19%	75%	-15%
Denmark	▲	8	20%	80%	10%
Belgium	–	5	22%	75%	3%
France	▲	5	32%	92%	18%
Ireland	▼	5	19%	64%	-16%
Norway	▼	2	25%	98%	-25%

FOCUS ON DESIGN

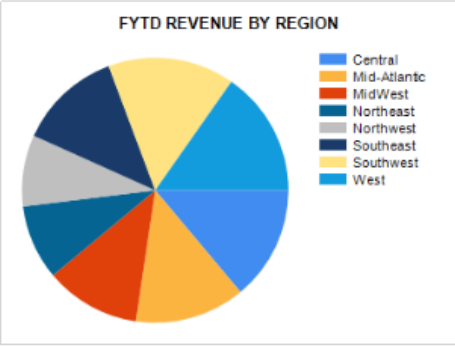
- Color palette matters
- Go for a clean look
- Don’t clutter with icons and FAQ
- Drill-down starting from high level
- Reduce or eliminate non-essential features
- Be consistent

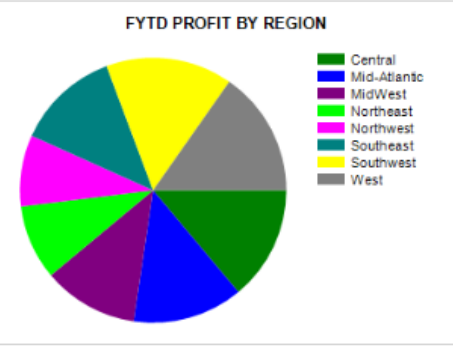
An Example of Poor Design

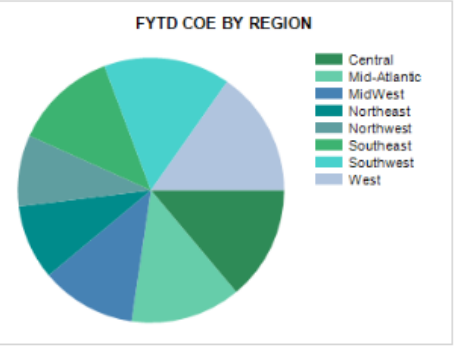
Acme Corporation

Today's Date: November 15, 2014

Welcome: John Henson

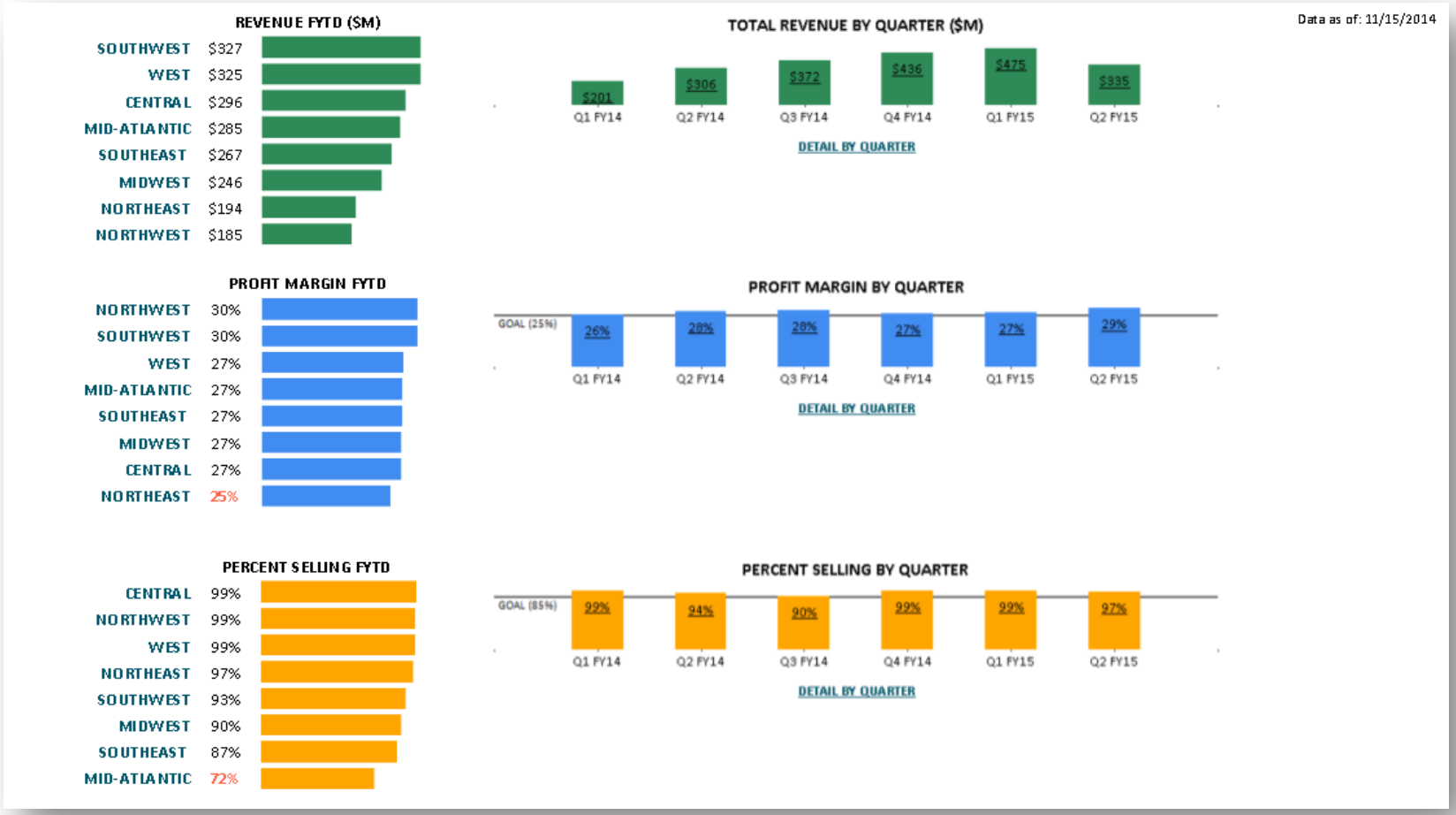






	Q1 FY14			Q1 FY15			Q2 FY14			Q2 FY15		
REGION	REVENUE	PROFIT %	% SELLING	REVENUE	PROFIT %	% SELLING	REVENUE	PROFIT %	% SELLING	REVENUE	PROFIT %	% SELLING
Central	$39.00	27.90%	72.52%	$90.00	28.00%	99.03%	$38.00	28.00%	45.41%	$12.00	25.61%	5.56%
Mid-Atlantic	$64.00	28.00%	28.84%	$25.00	25.80%	72.42%	$32.00	28.00%	9.05%	$39.00	28.00%	36.02%
MidWest	$3.00	18.62%	49.44%	$76.00	25.00%	29.80%	$1.00	28.00%	18.76%	$96.00	28.00%	46.34%
Northeast	$19.00	28.00%	69.62%	$47.00	28.00%	35.81%	$45.00	17.55%	32.86%	$3.00	28.00%	97.13%
Northwest	$9.00	28.00%	98.87%	$50.00	28.00%	33.24%	$28.00	42.62%	93.79%	$18.00	28.00%	5.80%
Southeast	$2.00	45.05%	21.83%	$87.00	28.00%	39.07%	$71.00	29.00%	63.49%	$44.00	23.00%	50.37%
Southwest	$1.00	28.00%	53.69%	$93.00	28.00%	7.73%	$49.00	28.00%	92.73%	$44.00	43.96%	89.82%
West	$64.00	21.66%	84.87%	$7.00	24.00%	29.70%	$42.00	28.00%	12.30%	$79.00	28.00%	16.16%

Good Design Techniques Applied



Some Final Thoughts…

- Mock up in Excel first
- Design for the long-run
- Fight the temptation to use flashy graphs
- Constantly edit down to the basics
- Mix Qualitative and Quantitative for maximum impact

Course Evaluation

Please take a moment to complete your course evaluation

- **Go to http://www.metricsthatmatter.com/ASPE**
- **Choose the class name listed with the date of the class and your instructor's name**
- **Fill out and submit the form**
- **It is available for 10 calendar days**

Thank You!

The End of the Tale

We take a deep breath and relax before we go and analyze data