

Overview & Context

Descriptive Statistics

Business Analytics - *Week 1*

Meet the Instructors

Nil Simsek

VP, Customer Acquisitions and Line
Management Analytics at Citi & Adjunct
Professor at NYU



[LinkedIn](#) | ns1254@stern.nyu.edu

JeanCarlo Bonilla

Director of Insights at DataKind & Adjunct
Professor at NYU



[LinkedIn](#) | jb3379@nyu.edu



NYU

TANDON SCHOOL
OF ENGINEERING

Course Description

Business analytics is a set of data analysis and modeling techniques for understanding business situations and improving business decisions. This course provides an introduction to business analytics concepts, methods and tools with concrete examples from industry applications.

Throughout the course, we explore the challenges that can arise in implementing analytical approaches within an organization. The course emphasizes that business analytics is not a theoretical discipline: these techniques are only interesting and important to the extent that they can be used to provide real insights and improve the speed, reliability, and quality of decisions.

Course Description

- In the first part of the course, we will focus on descriptive analytics and exploratory data analysis concepts with a refresher on basic probability and statistics.
- In the second part, we will cover principles, techniques, and techniques for spatial data, time series, and text as data.
- The final part of the course will introduce a project that links business impact and modern data analytics techniques for managerial decision making in functional areas, including finance, marketing, and operations.

Required Materials

Required Textbook

None.

Suggested Textbook by Topic

- Essentials of Business Analytics. Jeffrey D. Camm, James J. Cochran, Michael J. Fry, Jeffrey W. Ohlmann, and David R. Anderson. Cengage Learning, 2014.
- *Excell: Management Science: The Art of Modeling with Spreadsheets*, Powell and Baker. Wiley
- *R: Data Mining and Business Analytics with R*, Johannes Ledolter, 1st Edition
- *Business Analytics: Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*, Thomas H. Davenport & Jinho Kim, 2013
- *Decision Models: Spreadsheet Modeling & Decision Analysis: A Practical Introduction to Management Science*, Cliff Ragsdale, 6th edition.
- *Data Visualization: The visual display of quantitative information*, Edward R. Tufte, 2001
- *Scoping: Thinking with Data How to Turn Information into Insights*, Max Shron, 2014

Letter Grade	100% Scale	Grade Point Value
A	100-95	4.0
A-	94-90	3.7
B+	89-85	3.3
B	84-80	3.0
B-	79-75	2.7
C+	74-70	2.3
C	69-65	2.0
F	64-0	0

Letter	Value
A	
A-	
B+	
B	
B-	
C+	
C	
F	

**NO GRADE
NEGOTIATION IN THIS
COURSE!**



Course Web Page

You must have access to the **NYUClasses** site (<http://classes.nyu.edu/>). All announcements and class-related documents (supplemental and suggested readings, discussion questions, etc.) will be posted there.

Some class announcements will be distributed via NYU e-mail. Thus, it is important that you actively use your NYU e-mail account, or have appropriate forwarding set up on NYU Home (<https://home.nyu.edu/>)

In addition, lecture, data, and code will be posted in the following github repository <https://github.com/jcbonilla/BusinessAnalytics>

Statement of Academic Integrity

Students are expected to follow standards of excellence set forth by New York University. Such standards include respect, honesty, and responsibility. This class does not tolerate violations to academic integrity including:

- Plagiarism
- Cheating on an exam
- Submitting your own work toward requirements in more than one course without prior approval from the instructor
- Collaborating with other students for work expected to be completed individually
- Giving your work to another student to submit as his/her own
- Purchasing or using solutions or work online or from a commercial firm and presenting it as your own work

Course Schedule & Grading

See NYUClasses



What is Business Analytics?



Business Analytics Definition

Business analytics refers to the application of data analysis and modeling techniques for understanding business situations and improving business decisions.

IMPLICATIONS:

- data → past business performance
- methods → statistics + mathematics + computational methods
- business decisions → actionable insight

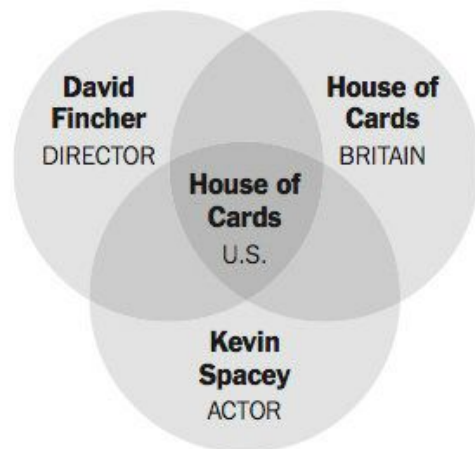


What analytics examples can you think of?



Circles of Proven Success

Netflix determined that the overlap of these three areas would make “House of Cards” a successful entry into original programming.



THE NEW YORK TIMES



NYU

TANDON SCHOOL
OF ENGINEERING



guests who viewed this item ultimately bought



\$199.99

reg: \$299.99

**Nikon Coolpix L840
16.1MP Digital Camera...**

Nikon

💎 spend \$25, get free shipping

★★★★☆ (11)



\$249.99

**Canon PowerShot SX400
IS Digital Camera ...**

Canon

💎 spend \$25, get free shipping

★★★★☆ (5)



see low price

reg: \$219.99

**Sony DSCW830/B 20MP
Digital Camera w...**

Sony

💎 spend \$25, get free shipping

★★★★☆ (25)



\$129.99

List: \$139.99

**Sony CyberShot
DSCW830 20.1MP Digital
Ca...**

Sony

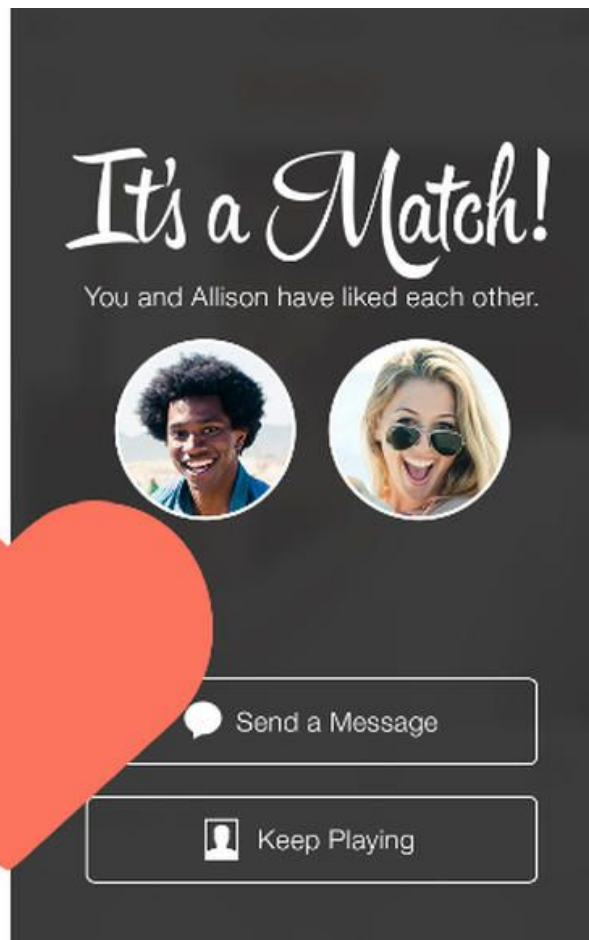
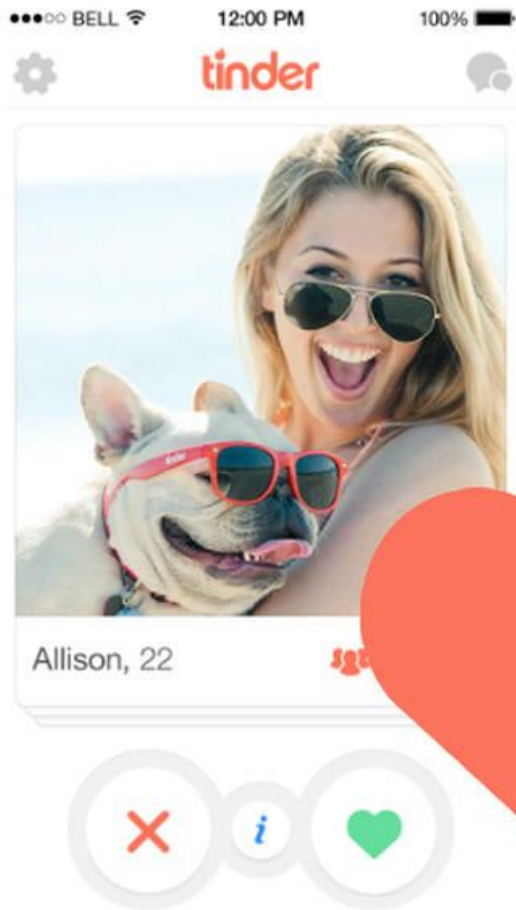
💎 spend \$25, get free shipping

★★★★☆ (139)



NYU

TANDON SCHOOL
OF ENGINEERING





NYU

TANDON SCHOOL
OF ENGINEERING

Analytics can also do harm!

Two Petty Theft Arrests

	
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.



Types of Analytics



Descriptive Analytics

Descriptive Analytics:

Encompasses the set of techniques that describe what has happened in the past whether that is one minute or one year ago. The vast majority of the statistics we use like sums, averages, percent changes fall into this category.

About 35% of companies surveyed say they do this consistently.

Use Descriptive statistics when you need to understand at an aggregate level what is going on in your company, and when you want to summarize and describe different aspects of your business.



Types of Analytics



Descriptive Analytics



Predictive Analytics

Predictive Analytics:

This type of analytics are about understanding the future and providing estimates about the likelihood of a future outcome. They combine historical data found in ERP, CRM, HR and POS systems to identify patterns and apply models and algorithms to capture relationships between various data sets.

However, less than 1% of companies surveyed have tried this yet.

Use Predictive analysis any time you need to know something about the future or fill in the information that you do not have.



NYU

TANDON SCHOOL
OF ENGINEERING

Types of Analytics



Descriptive Analytics



Predictive Analytics



Prescriptive Analytics

Prescriptive Analytics:

Prescriptive analytics attempt to quantify the effect of future decisions in order to advise on possible outcomes before the decisions are actually made. Prescriptive analytics are relatively complex to administer, and most companies are not yet using them in their daily course of business.

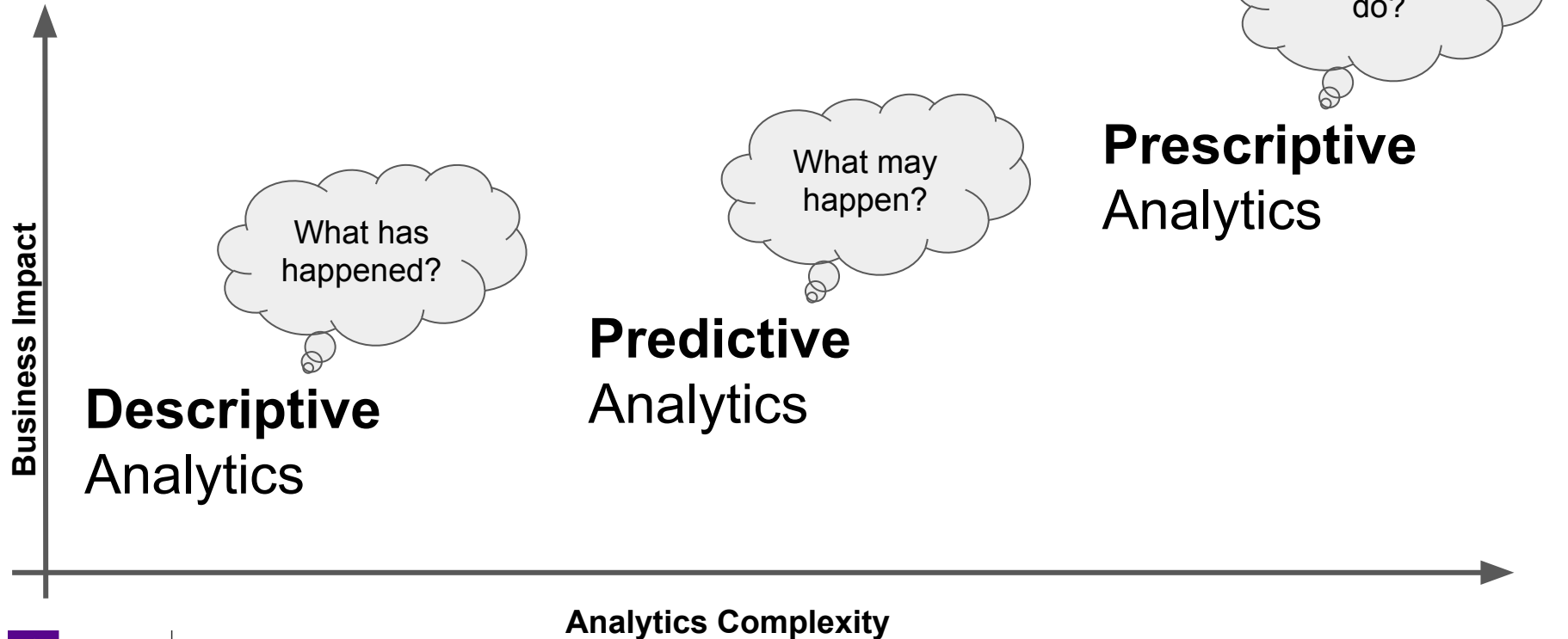
Use prescriptive statistics anytime you need to provide users with advice on what action to take.



NYU

TANDON SCHOOL
OF ENGINEERING

Types of Analytics



Analytics vs BI vs Data Science

Business Intelligence (BI): tools and systems to gather, store, access and analyze an organization's raw data. BI actually presents the insights determined by Analytics in reports, dashboards, or interactive visualizations

- querying and reporting tools, dashboards
- traditionally used to determine trends in historical data

Data Science: involves using automated methods to extract knowledge or insights from structured or unstructured data.

- employs techniques and theories drawn from mathematics, statistics, information science, machine learning, AI, and others.

Data Mining vs...

- Data querying (SQL, QBE, other GUI-based querying)
 - Very flexible interface to ask factual questions about data
 - No modeling or sophisticated pattern finding
- Traditional statistical analysis
 - Mainly based on hypothesis testing or estimation/quantification of uncertainty
 - Should be used to follow-up on data mining's hypothesis generation
- Automated statistical modeling(e.g., advanced regression)
 - This is data mining, one type -- usually based on linear models
 - Massive databases allow non-linear alternatives

Analytics is more a catch-all term that encompasses both BI and Data Science as well data mining/modeling, and forecasting



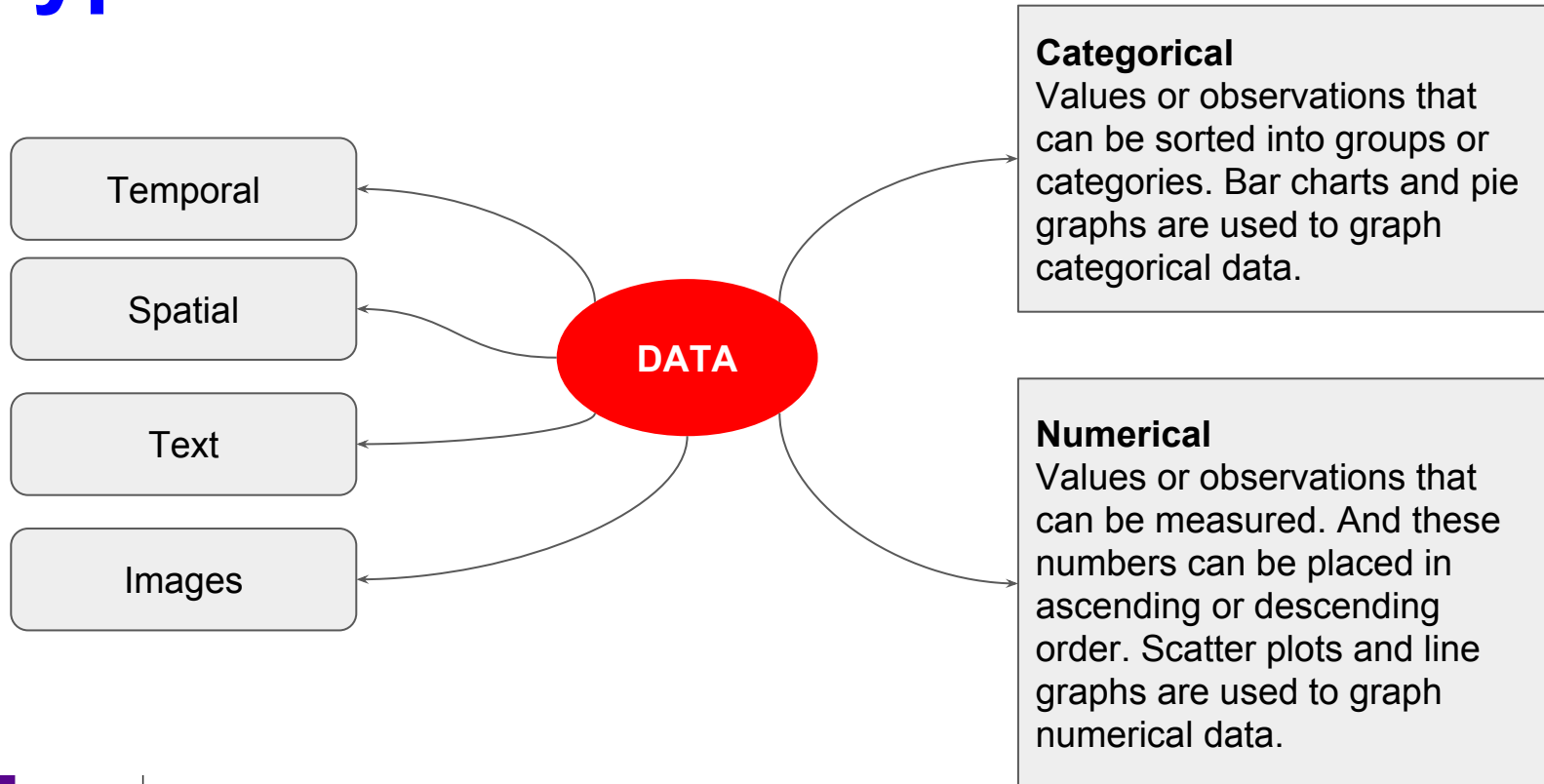
Business Analytics Definition

Business analytics refers to the application of data analysis and modeling techniques for understanding business situations and improving business decisions.

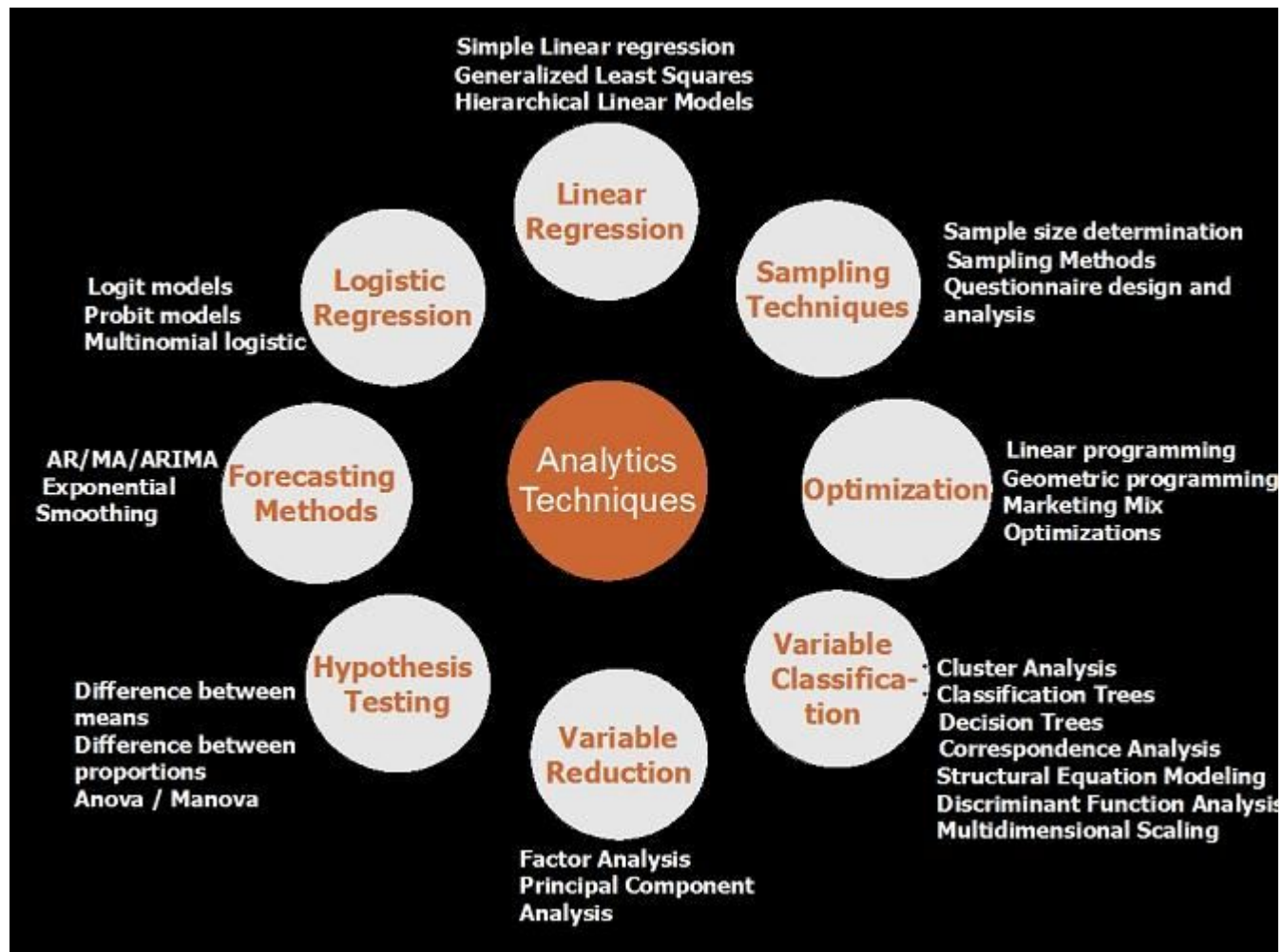
IMPLICATIONS:

- data → past business performance
- methods → statistics + mathematics + computational methods
- business decisions → actionable insight

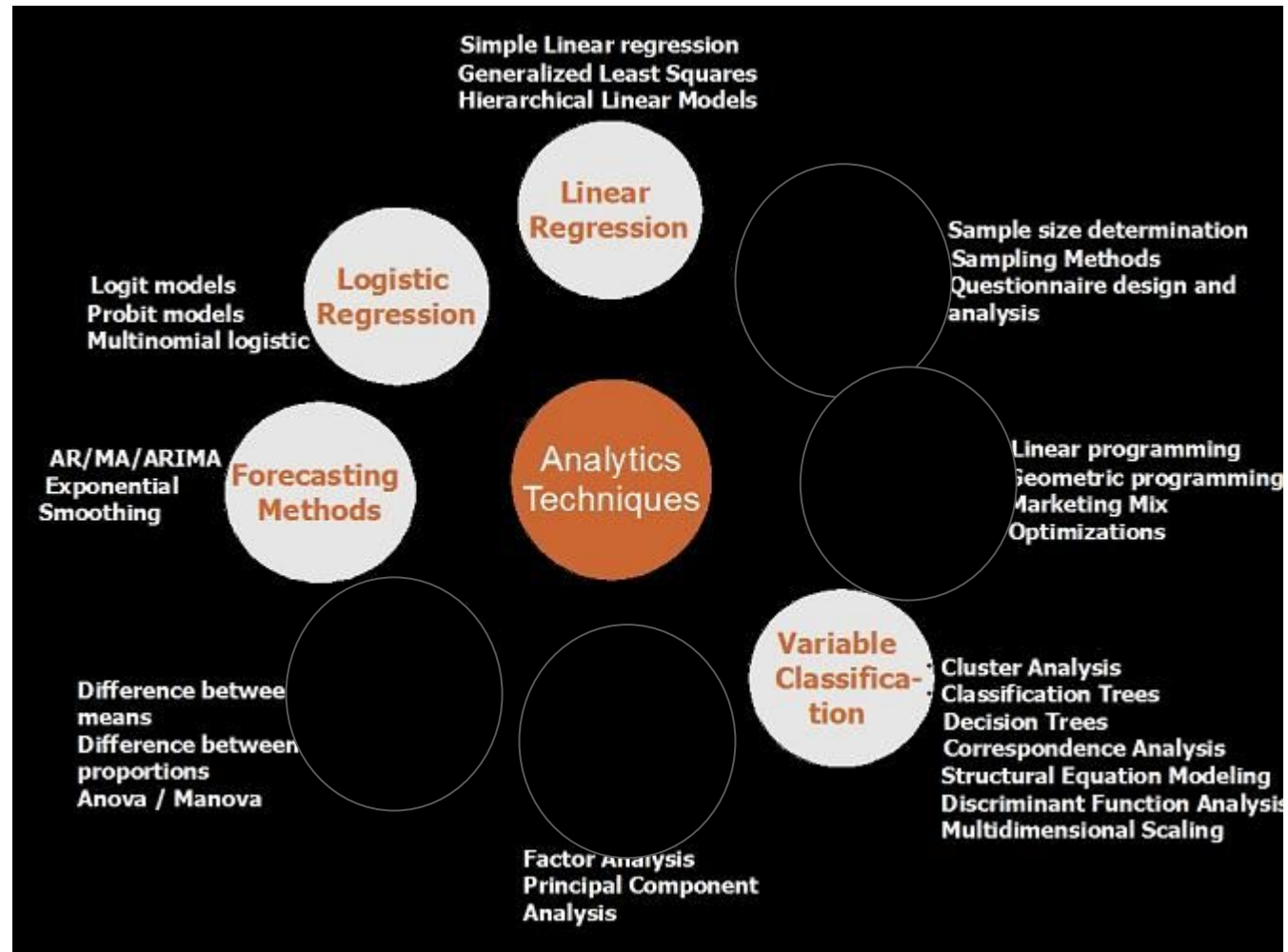
Types of data



Methods



Methods





Introduction to R & Descriptive Statistics



Lesson Objectives

1. Descriptive Statistics

- a. Measures of central tendency
- b. Measures of spread and variability
- c. Measures of association
- d. Shape of Distributions & Statistical Graphics
 - i. Histograms
 - ii. Scatterplots
 - iii. The Box Plot

2. Introduction to R

- a. Data variables & basic operations
- b. Loading data and reading data
- c. Summary stats

Descriptive & Summary Statistics

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way. Specifically it helps describe,

- Patterns that emerge from the data.
- Provide a way to describe our data
- Do not make conclusions or predictions

Descriptive statistics **do not** allow us to make conclusions or predictions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made.

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data → measures of central location.

1. Mean

- a. Arithmetic mean
- b. Geometric mean
- c. Harmonic mean
- d. Weighted mean
- e. Truncated mean

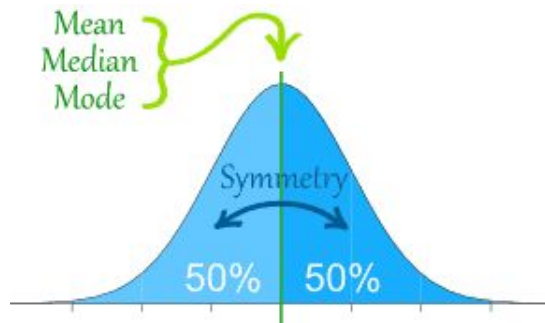
2. Median

3. Mode

Measures of Central Tendency

For any symmetrically shaped distribution the mean is the point around which the symmetry holds.

What is the single most important distribution pattern occurring in many natural phenomena?



$$\text{Mean} = \frac{\text{sum of all values}}{\text{total number of values}}$$

$$\text{Median} = \text{middle value (when the data are arranged in order)}$$

$$\text{Mode} = \text{most common value}$$

Note!: There is a difference in notation between the population mean and the sample mean. For population mean we use the Greek lower case letter "mu", denoted as μ , and for sample mean we use the x with bar on top, pronounced x bar

Example

Given this dataset below, find the mean (arithmetic) , media, and mode:

10, 20, 30, 40, 40

$$\text{Mean} = 140/5 = 28$$

$$\text{Media} = 30$$

$$\text{Mode} = 40$$

Measures of Central Tendency

- The median is the middle value in the data set above (below) which there are 50% of the values.
 - Usually calculated by sorting the values and looking at the value in the middle position
 - Median household income in US families = \$51,939
 - Mean household income = \$72,641
- For skewed distributions or when there is concern about outliers, the median may be a better measure to use.
- The mode is the most frequently occurring value.
 - useful when describing whether a distribution has a single peak (unimodal) or two or more peaks (bimodal or multi-modal)



Example: Zagat

What measures of central tendency would you use?

What is the business problem you are trying to solve?

What does it mean to be above or below the mean?

Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55



Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55
Mean	23.8	17.6	20.7	53.0
Median	23.8	17.6	20.7	53.0
Mode	22.0	3.0	13.0	74.0



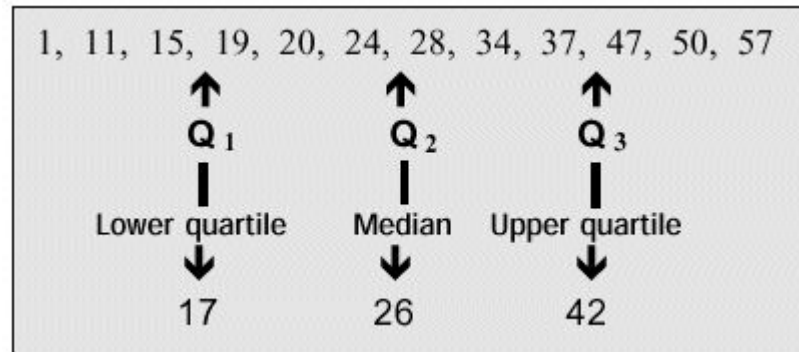
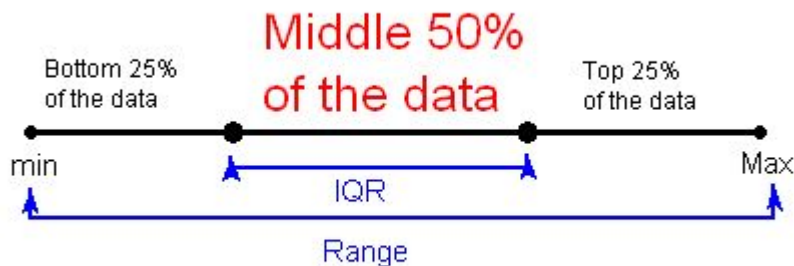
Measures of Spread or Variance

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population. It is usually used in conjunction with a measure of central tendency, such as the mean or median, to provide an overall description of a set of data.

1. Range
2. Quartiles and Interquartile Range
3. Variance & Standard Deviation
4. Coefficient of Variation

Measures of Spread

- The quartiles of a population divide the observed data into even fourths.
 - The 1st quartile (Q1) is the number below which 25% of the values fall.
 - The 3rd quartile (Q3) is the number below which 75% of the values fall.
 - Interquartile Range (IQR) = $Q3 - Q1$ → the middle half of the data fall within IQR
 - Useful in detecting outliers



Measures of Spread

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

- A frequently used formal definition for an outlier is any value outside the interval:

$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

Measures of Spread

- The variance of a population is defined as the mean squared deviation.
- This is a measure of spread, because the bigger the deviations from the mean, the bigger the variance gets.
- The standard deviation is simply the square root of the variance.

Variance → Standard Deviation

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \rightarrow \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Where

σ^2 = the variance

σ = the standard deviation

\sum = the sum of

X = a datapoint

μ = the mean of the data

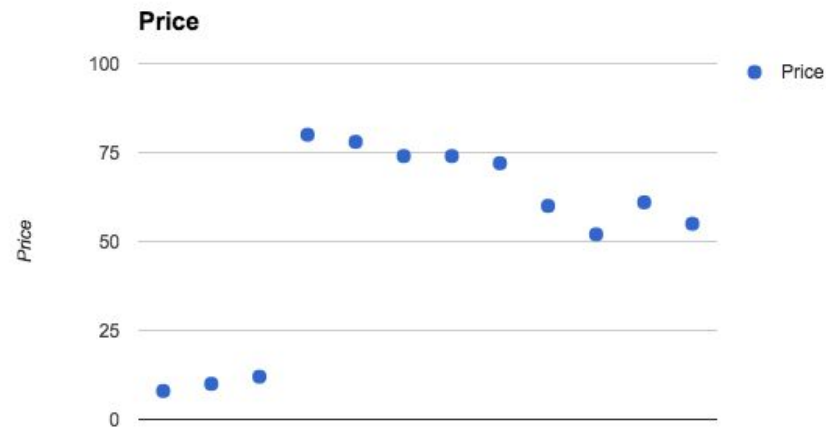
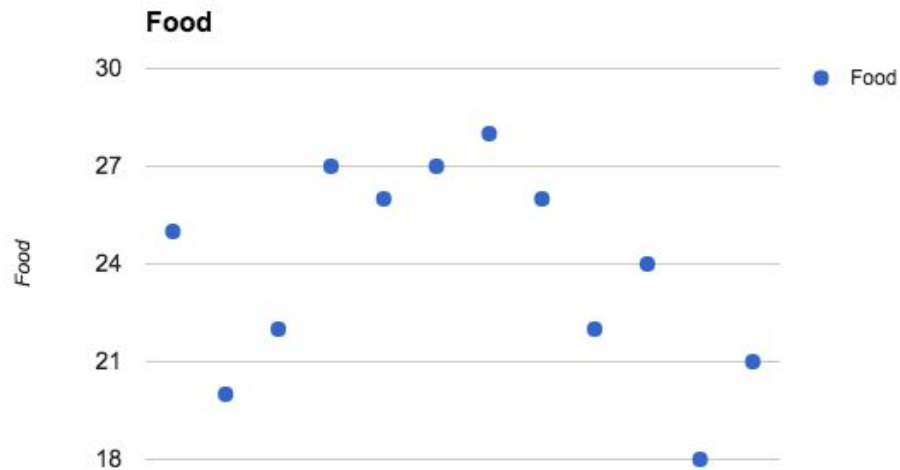
N = the number of datapoints in the set



Which rating has more variance?

Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55
Mean	23.8	17.6	20.7	53.0
Median	23.8	17.6	20.7	53.0
Mode	22.0	3.0	13.0	74.0





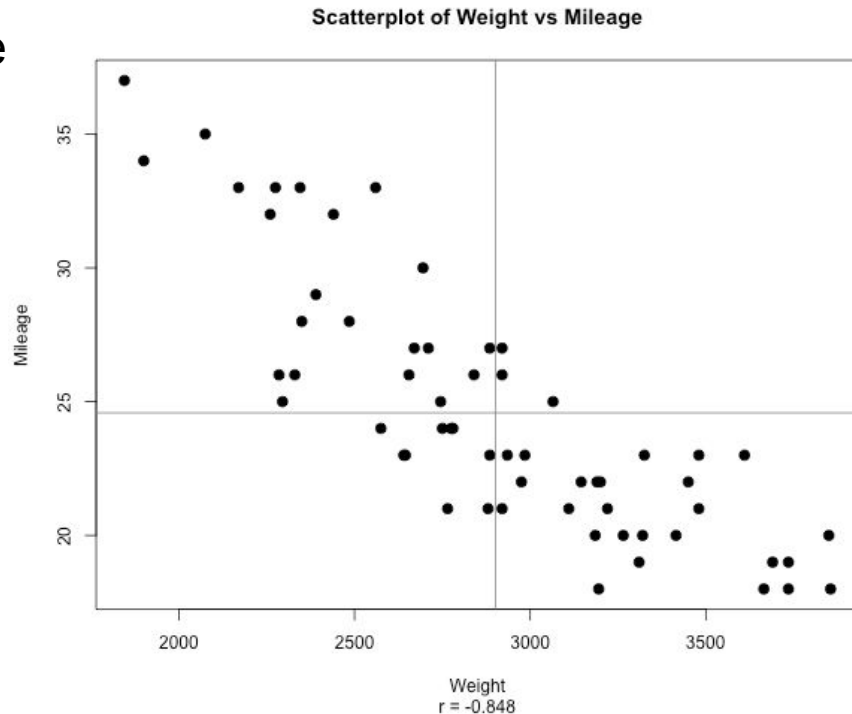
Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55
Mean	23.8	17.6	20.7	53.0
Median	23.8	17.6	20.7	53.0
Mode	22.0	3.0	13.0	74.0
Variance	10.2	71.4	28.4	751.8
Standard Deviation	3.2	8.4	5.3	27.4



Measures of Association

The measures of association refer to a wide variety of coefficients that measure the statistical strength of the relationship on the variables of interest.

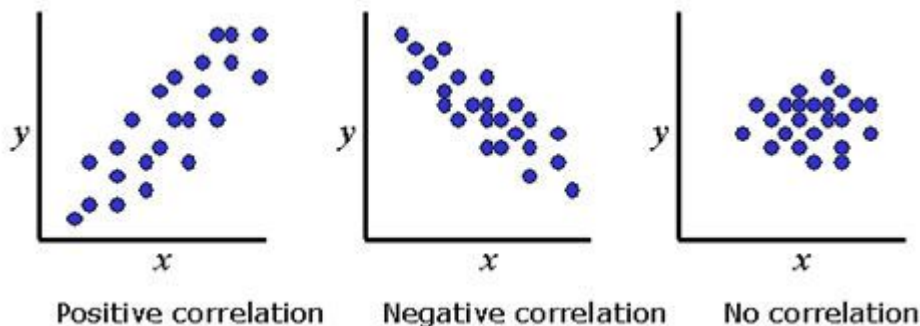
The term "association" is closely related to the term "correlation" → two or more variables vary according to some pattern.



Correlation

Correlation refers to any of a broad class of statistical relationships involving dependence.

Sometimes the pattern of association is a simple linear relationship → "the correlation coefficient"



Correlation Examples

Which of these are correlations in business

- Price & Revenue
- Salary & Education
- Gold & Oil

Classify as negative or positive correlation

- A student who has many absences has a decrease in grades
- As one increases in age, often one's agility decreases.
- When workers get a raise, morale improves.

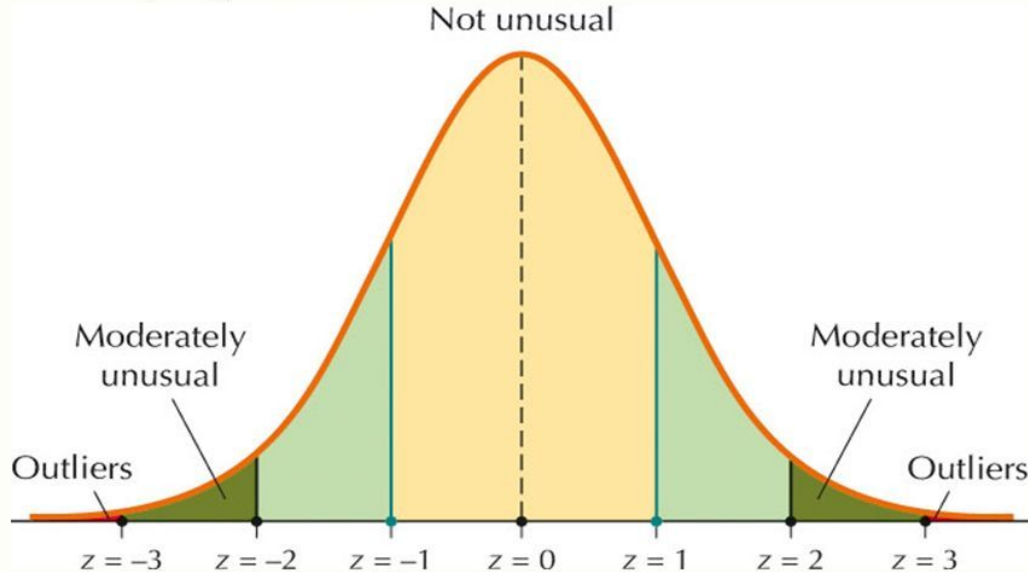
~ <http://www.tylervigen.com/spurious-correlations>

Outliners and Z-Scores

Detecting Outliers with z-Scores

28

An **outlier** is an extremely large or extremely small data value relative to the rest of the data set. It may represent a data entry error, or it may be genuine data.



$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

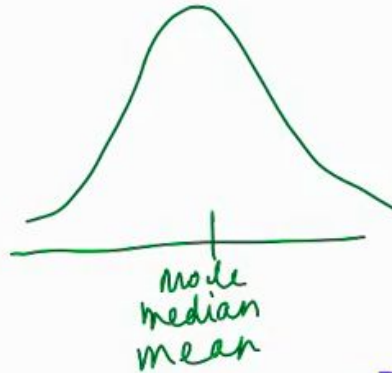
σ = Standard Deviation

- Z-score = 0 (observation is equal to the mean)
- Z-score = x (observation is x standard deviations away from mean)
- Z-scores between -3 and +3 (detects 99% of your information in dataset)
- Z-scores between -2 and +2 (detects 95% of your information in dataset)
- Z-scores between -1 and +1 (detects 68% of your information in dataset)

Shape of Distribution

The relative location of the mode, median, and mean in a **unimodal** distribution:

Symmetric



For a symmetric distribution, the mean, median, and mode are all approximately the same.

Left-skewed



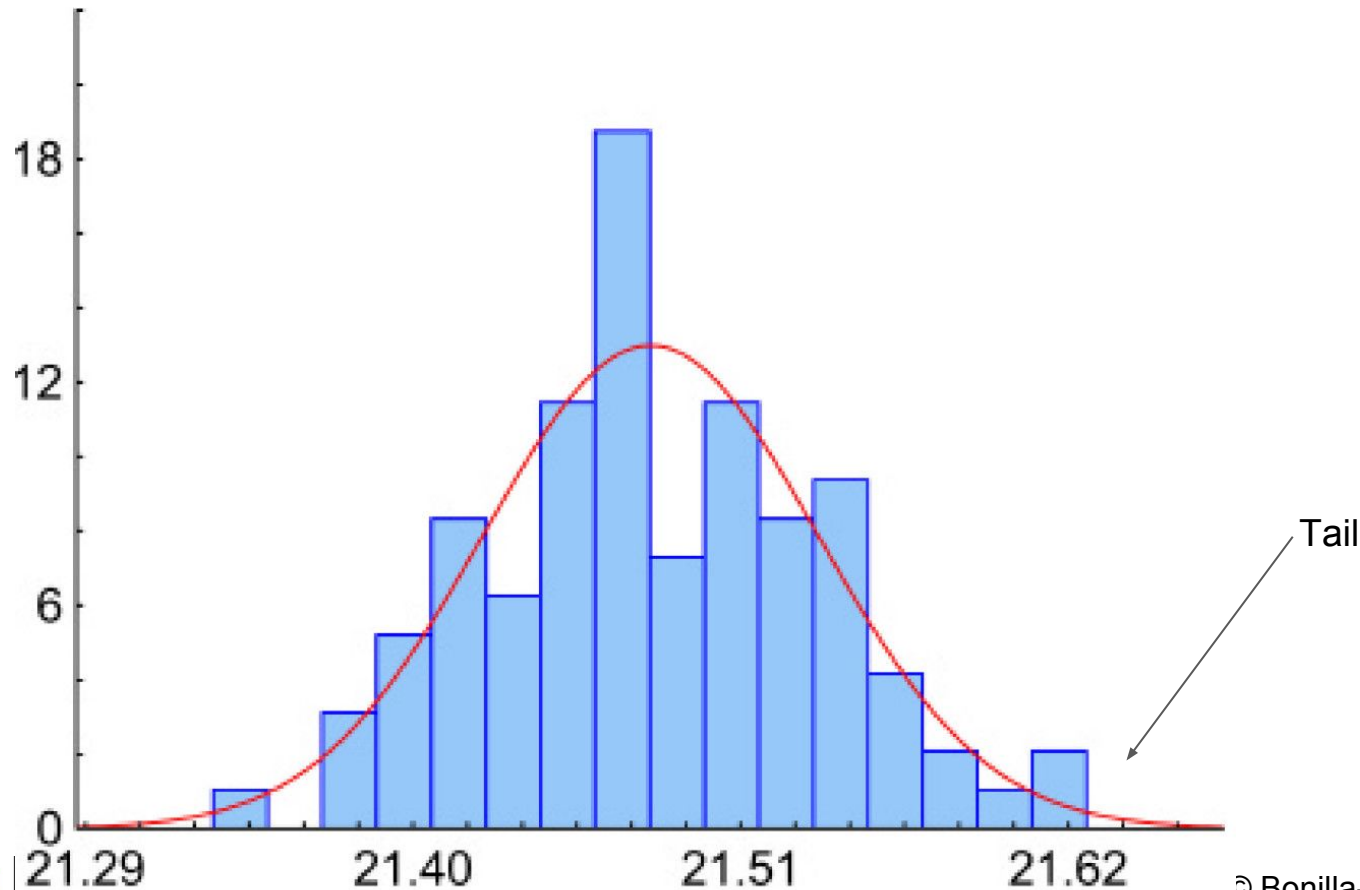
For a left-skewed distribution, the mode is larger than the median which is larger than the mean.

Right-skewed



For a right-skewed distribution, the mode is less than the median, which is less than the mean.

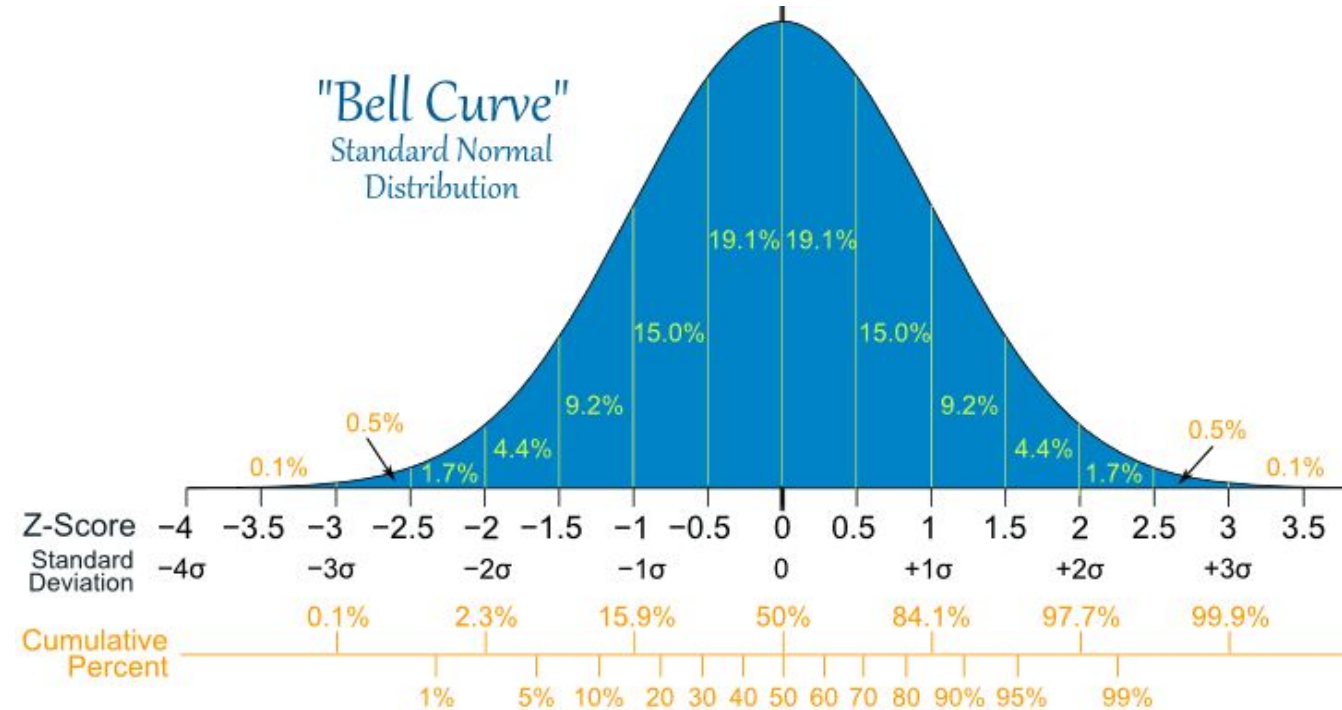
Histograms & Distributions



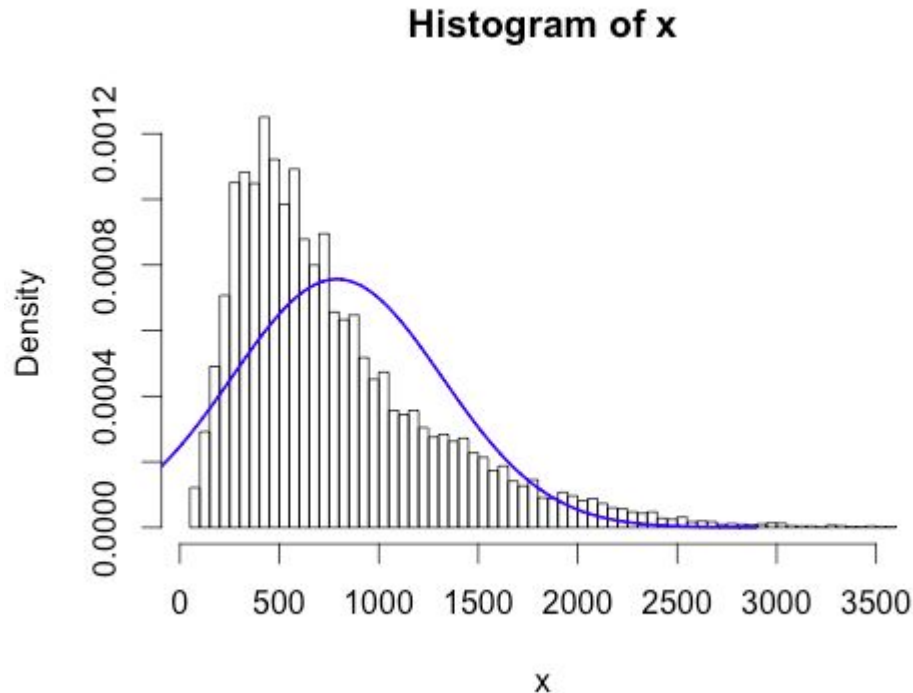
Normal Distribution

Properties:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean
- 68% of values are within 1 standard deviation of the mean
- 95% of values are within 2 standard deviation of the mean
- 99% of values are within 3 standard deviation of the mean



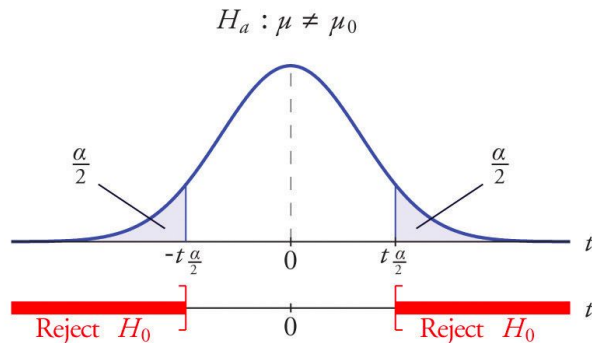
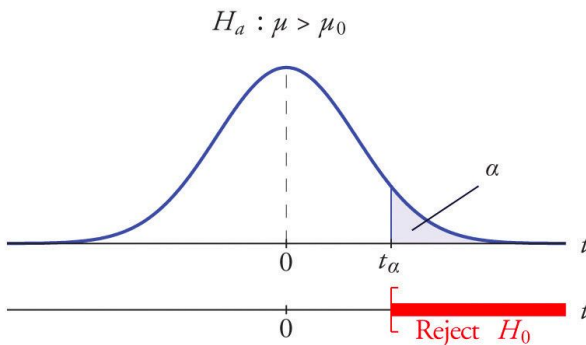
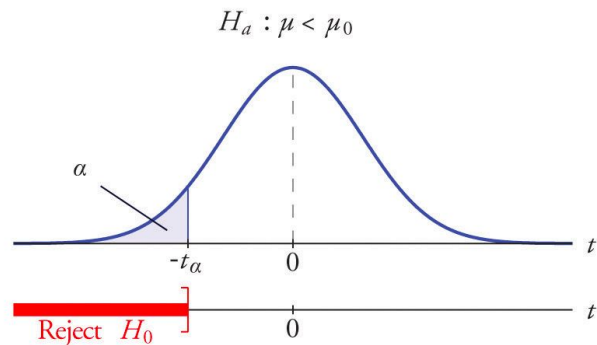
Normalizing Data → Fitting a distribution to data



Normal curve centered at mean of data set with standard deviation equal to the deviation of the sample data.

How good is this model?

Significance Testing and Confidence Intervals



Statistical test provides a mechanism for making quantitative decisions about a process or processes.

The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process.

The conjecture is called the **null hypothesis**

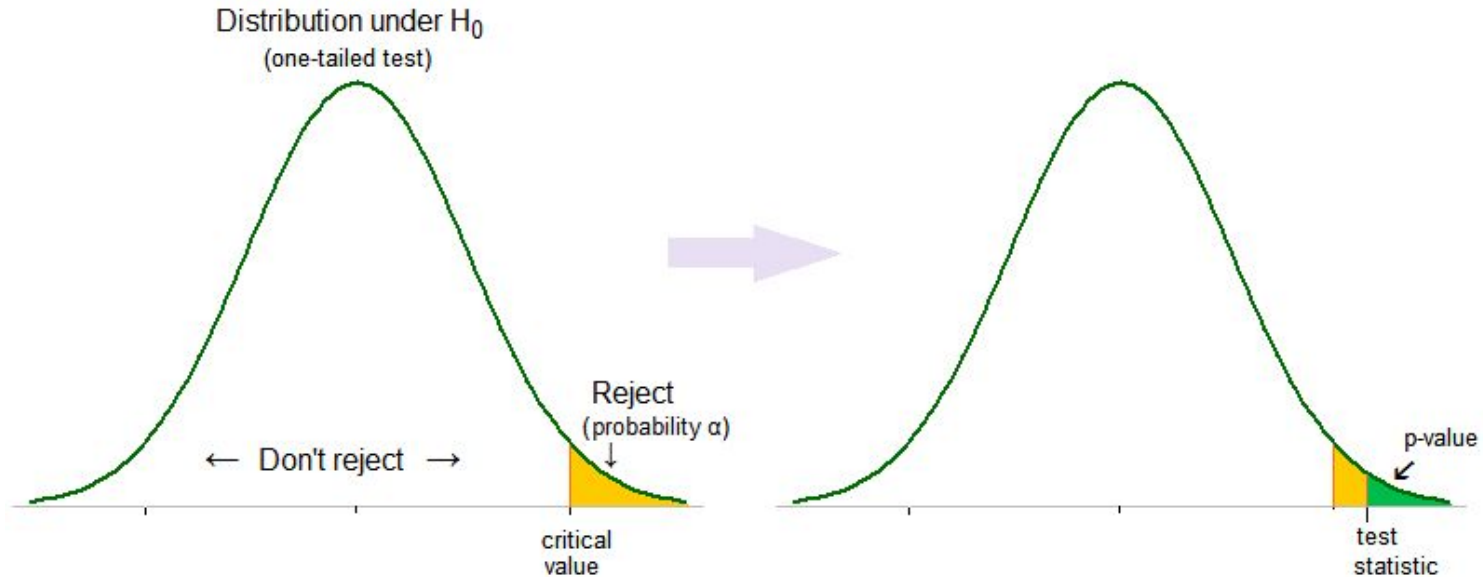
Hypothesis Testing

Steps:

1. **State the hypotheses.** This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
2. **Formulate an analysis plan.** The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
3. **Analyze sample data.** Find the value of the test statistic (mean score, proportion, t-score, z-score, etc.) described in the analysis plan.
4. **Interpret results.** Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.



Significance Levels (α) & P-Values



Alpha sets the standard for how extreme the data must be before we can reject the null hypothesis.

The P-value indicates how extreme the data are.

- If the p-value is less than or equal to the alpha ($p < \alpha$), then we reject the null hypothesis, and we say the result is statistically significant.
- If the p-value is greater than alpha ($p > \alpha$), then we fail to reject the null hypothesis, and we say that the result is statistically nonsignificant (n.s.)



Statistically Significant Results

- A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- A test result is statistically significant when the sample statistic is unusual enough relative to the null hypothesis that we can reject the null hypothesis for the entire population.
- The common alpha values of 0.05 and 0.01 are simply based on tradition.

Introduction to R

R is a language and environment for statistical computing and graphics. It was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.

You will need to install **RStudio**. RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

R Prep:



NYU

TANDON SCHOOL
OF ENGINEERING

Course GitHub

GitHub is a web-based Git or version control repository and Internet hosting service. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features.

Use to access:

- Lectures
- RCode
- Data



<https://github.com/jcbonilla/BusinessAnalytics>