
Data Analysis Boot Camp

Classroom Manual

Learn practical, hands-on data analysis skills and tools to maximize and leverage your data.

29400PG_7.0_2017

www.aspetraining.com

www.techtowntraining.com

A Techtown Training program from ASPE

877-800-5221

Data Analysis Boot Camp

29400M_7.0_2017



NOTES:

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**



This class discusses data as it is in the real world. When analyzing any data, the first step is to get data. Very often these come from external sources and will require extensive methods to evaluate, clean, store and manage. Once that data is “in house” a variety of techniques can be applied to extract information from the data you have managed to collect. All of this starts with a basic understanding of probability and statistics. Analysis techniques applied to data in the real world are very often statistical in nature for the simple reason that the world is a complex place. We don’t know many “textbook” equations that we can use to determine a business cycle, an economic cycle, or social phenomenon. We have to gather data and start to model it. Then we can work from the models to create predictions, but these are inherently hazy. They follow probability and statistical rules. Starting from these, you can gain insight into what complex software algorithms are telling you. After you gain confidence in your predictions, you can begin to tell the story of your conclusions.

Data and Information

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

Section 1



NOTES:

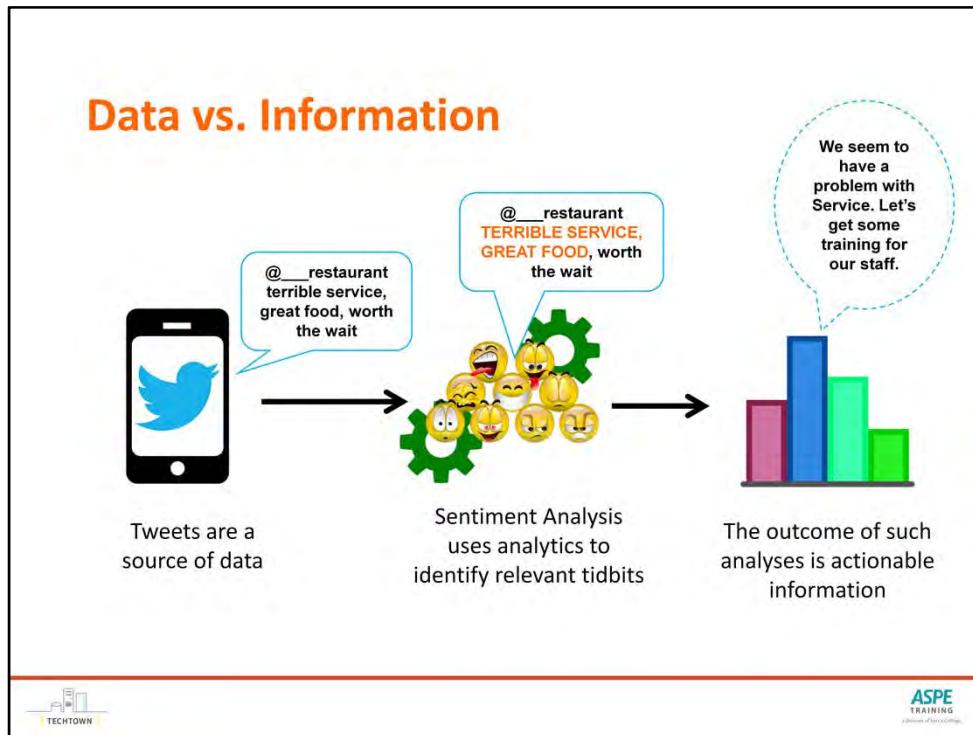
Data in the Real World

- **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 is all of the little bits and pieces of digital “things” we generate every minute of every hour. It is not of itself information that can be immediately used for productive conclusions. Data must be gathered, probed and prodded, and understood before it can be brought into the organization for useful analysis. This is called profiling. After the data is profiled, it must be cleansed of all sorts of inconsistencies that is inherent in the way all of these bits are captured. Not all data is numerical in its nature, so some sort of scheme must be applied to quantify anything we want to analyze. The scheme must be numerical in nature because we are dealing, in the end, with numbers. The numbers are how we model, how we predict, how we determine data that is not where we might expect it to be (outliers).

Data in the real world is growing rapidly, and so are ways to gather it and process it. How we do so is really the scope of many classes of programming and data processing, but they are growing almost as fast as the data itself.



This is the path of data into the enterprise from random bits to actionable information from which conclusions can be drawn. In general, this is done by identifying the type of data you want to collect and perhaps the sources from which the data should come. Then, programmatically, the sources are polled for their data and the software brings the data into the enterprise. For example, the day's exchange rates for all currencies might be pulled from a web service at your favorite bank.

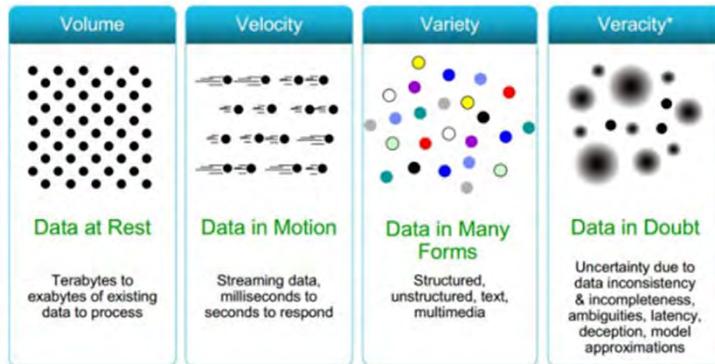
When the data is gathered, it must be stored and managed. There are many ways to do this, the most common in the enterprise being the relational database (Oracle, MS Sql Server, MySql). At this point, database professionals work with the analysts to clean the data to make it usable.... but more on all this later.

Consider Twitter and its effect on customer service...

Twitter is often where people will go to express their opinion of a company's customer service. The person's tweet is issued from the Twitter software service and captured by any client software listening for it. How do we gauge the customer's opinion? The tweet is natural language with all possibilities of interpretation as one would expect. Moving from natural language to quantifiable aspects of the mood those tweets are trying to convey is known as Sentiment Analysis and involves complex algorithms to determine the mood the tweets are trying to convey. Once the mood is read and quantified, analysis can begin and actionable information developed.

Now in the real world, we cannot develop a model of customer service progress from one person's tweet(s). We need to identify all customers who are on Twitter talking about the company. When we get enough, the sentiment analysis of that scope of data can turn into real information.

The Many “Vs” of Data



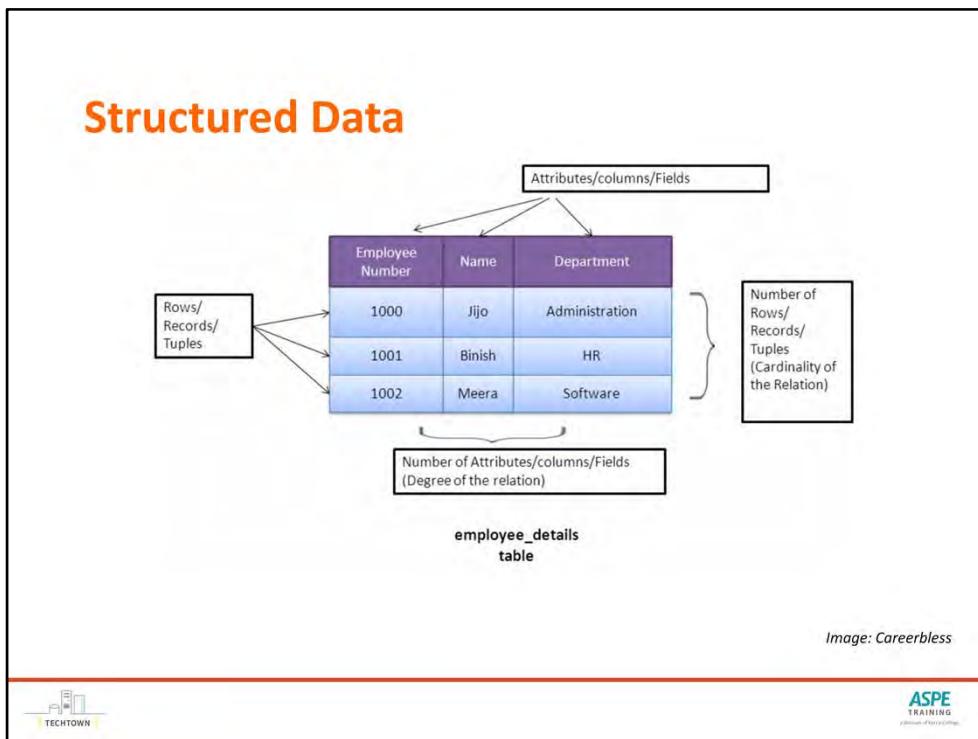
The process of bringing data into the enterprise requires the consideration of the Vs of data

Volume – what is required, what is available, and what can be stored internally all force one to think about the amount of data (Mega, Tera, Peta) being considered for a given project

Velocity – issues of availability of data due to the physical limitations of moving the data through the networks have to be considered when planning an analysis project

Variety – bringing data from the many different external sources (tweets, Yelp, Facebook, Government centers, etc) to be used internally requires transformation from many different forms

Veracity – when using external data, what is the quality of that data? Is it good enough to lead to actionable information with a high degree of confidence?

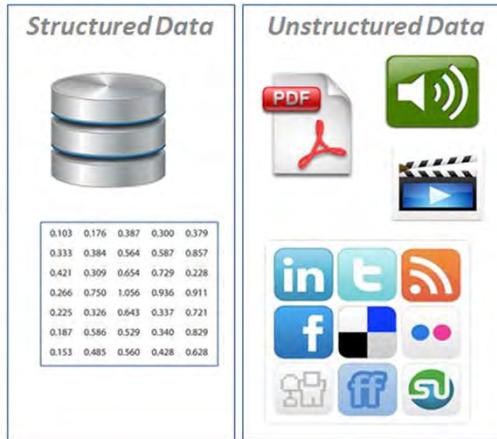


The internal enterprise usually makes use of highly structured data. Relational databases are the norm inside of most major organizations. These implement rules around highly structured collections of information enforcing:

- Data types
- Relationships between tables
- Constraints on what data can be entered (beyond just data type)

When data is managed by a RDBMS, a lot of the dirtier aspects of data are automatically handled. This is because the data types and relationships will “kick out” anything that violates these when the data comes into the enterprise. Of course, this can also mean that some valuable data is lost. Why is it valuable if it is dirty? Depends on the situation. For example, some survey responses may be rejected because the respondent did not enter their city in the form. It all depends on what you decide is necessary when you frame the problem. We will see this many times throughout this course. In the end, the analyst is indispensable. The automated processes and analytical tools are not foolproof means to an end.

Unstructured Data



...everything else.

Image: Infosys

Even though different applications like PDF and MPEG have a structure, the data placed in them can be unstructured. Consider the title of a document in a PDF. PDF rules may be enforced to center and bold the title, but the words in the title itself may be anything from simple text to text with numbers and symbols, of a variety of lengths and case. This way, PDF readers know how to typeset the document, but the data they are typesetting is not in any structure in the RDBMS sense.

Types of Data

Internal/External

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



Internal data is data that is within the enterprise. The data is created by the enterprise, stored and managed by the enterprise. It spends its entire lifetime under your control. External data comes from somewhere else. It is gathered by you (or someone with your team) but its initial integrity is under the control of an outside group. Generally, if you want high quality data, that data comes with a fee. However, it may be that the fee is reasonable compared what it would take for the internal team to develop it.

Consider demographic data for an advertisement you want to run. There are companies that specialize in maintaining demographic databases that you can purchase. They may be expensive, but to reproduce this data on your own requires a considerable effort and expense.

Data Quality should be a priority



In the end, bad data is worthless and potentially very damaging to the organization. The damage can be in many forms, but, ruling out actual malicious code, the damage is in the form of erroneous conclusions for the enterprise. A great deal of money can be lost if conclusions are faulty and the first place to start to prevent that is to ensure that the data is clean. What does it mean to be clean? We will cover these in the next slides.

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**



One aspect of dirty data is data duplication. Perhaps you are tracking a group of people for a medical study over a period of one year. Imagine if you do not account for spellings of a person's name when you take the second or third survey – Jane Doe submits the second survey as J. Doe and becomes two different people in the database. Now the data is not tracking Jane's first and second surveys, but you are tracking two different people. One simple way to prevent this is the Single Source of Truth where every participant is tracked by a unique ID and the name is keyed into the DB once and only once. This prevents multiple copies of a name in multiple places in the data that are seemingly different but actually the same.

Starting on the right foot

The risks of skipping or neglecting data cleansing:

- Inaccurate analysis
- Wasted time/money
- Inconsistent results



There are many ways that dirty data leads to these results. Some more ways,

- Duplicate records can lead to improper sample sizes when performing simple statistics
- Data sources which provide continuous data updates (say from stock information) can generate inconsistent trending information when the data is not treated properly on import.

Common Issues/Causes

Key Issues

- **Duplicates**
- **Misspellings**
- **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



These are a list of some common causes of dirty data. Many of these are caused simply by working with manual entry. For instance, hand typing a department in a text field instead of offering a dropdown list of available departments the way most state dropdowns are available. The Y2K problem was an example of poor future planning as only two digit years were required for a very long time.

To look at just a few...

Inconsistent formatting is often found in dates.

Misspellings in names of people and places (from registration forms or surveys)

Inconsistent data population is often found when downloading an Excel file from an external data source. This is because data was not rigorously collected.

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**



We already spoke about SSOT. This is very important. Consider the state dropdown mentioned a few slides ago. The SSOT for this is usually a database table with all of the state codes in it. Perhaps each one has an integer ID in the table as well. When there is any record stored for a user filling out this form (whatever form it may be), the state code is referenced by the ID in the table. The presentation on the screen comes from the SSOT table so that the spelling is always the same. When you need to generate a report, or view the state data for a given person in any way, you resolve the keys linking to the SSOT table, so that you always get consistent spelling.

If you have older projects, you can implement projects to review the older data and determine how it may be fixed. Even with new projects, review policies are important. Having a policy to kick data out when you need to preserve data integrity is usually referring to the minimum “atomic unit” of data. The analytics team has to decide the minimum effective amount of data in any record. If the data integrity dips below the minimum, then it should just be rejected and noted in order to maintain quality of any conclusions drawn from the data.

Data Analysis Defined

We define some essential terms

Section 2



NOTES:

Data Analysis Defined

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



Data is everywhere, but of itself is not useful. Information comes from analyzing data, but what does it mean to analyze data. There are, of course, specific analyses that everyone is familiar with – averaging the semester grades, percentage of people who shopped at Wal-Mart in the last 10 days, etc. To analyze data is to perform research in a field, organize data that is relevant in that field, and shape it in ways that tell us useful things. Changing data here does not mean that we change the data in any way we like to draw our conclusions, it means to refactor or reshape the data so that different dimensions are compared together, or to rescale so that two different variables are measured on the same graduation...things like that.

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**



We analyze data because we are looking for some useful information. The information can be to make a prediction, or to see where we have been and understand what happened in the past. Of course, the past is the past, but analysis of data about a past event can help us determine what is to come. Economics is a good example of analyzing past data. A recession is generally defined as two consecutive quarters of negative GDP growth. Basically, by definition, you cannot tell if the economy is in a recession until past data has been analyzed.

In order to make a prediction, we generally have to have a model. In order to make a model, we gather data and analyze it. Later in the course, we will work with moving data by analysis to a model. From the model, you can push outward and try to make predictions about the future. The accuracy of the predictions depends on many things, no the least of which is the quality of the data and analysis.

Data can also be used for monitoring and reporting problems. Why does it take data to determine if something is a problem? A simple error in a computer network can just be reported as is when it occurs, but what about reporting credit card fraud? Credit card fraud, or at least potential credit card fraud, is found from mining millions of records and developing a model for what may be considered fraud. Then all new transactions are monitored and compared against data for what is considered normal. If the new transaction fails the “normal” test, then the monitoring system can notify the company about the potential fraud.

And then there is SWOT – strengths, weaknesses, opportunities and threats. You can read more about SWOT at Wikipedia (https://en.wikipedia.org/wiki/SWOT_analysis) but basically it is the structured approach to business planning. You can answer questions like “who may be moving into our market?” and “what markets may we be missing with our product?” by data analysis.

Data Analysis Mindset

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



The mindset of an analyst is one of curiosity and tenacity. Finding the story embedded in numerical data requires a lot of thought and persistence. One must also be comfortable with the lack of clear results. Very often the results of analysis are suggestive, but not definitive. Simply put, there is not an answer, only solutions which are better or worse than other solutions.

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?



In the end, computers calculate. This is all that they do. It takes a person to discern the quality of the result. That quality relies on the proper framing of the problem, the best data sources, the proper analytical techniques for the field, and good communication of the results.

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**



This set of steps is a guideline for the data analyst. Define your problem and what you want to accomplish, determine what metrics to measure and what constitutes success or failure (when applicable), gather the data and cleanse it, analyze and build your model and go over the cycle again.

When building a model, remember to test, test, test. Use the original data and make sure you get out of your model the actual result that the data suggests. Find other data and test against it. Make a short term prediction and test against the actual result. The process is a cycle because each new piece of relevant data improves the model.

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



The difference between analysis and analytics is one of degree. Analysis is usually reserved for using mathematics to look at data and turn it into information. Analytics often refers to a prebuilt or purchased package of analysis techniques commonly applied to a particular business or science problem.

For example, finding averages and percentiles is analysis. An analytics package will run the average and percentiles automatically. It obviously is not that simple, but that is the idea. Analytics is packaging a set of analysis techniques which we know to be relevant to a particular field.

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



These are some common terms in analysis. Notice that a sample is a subset of a population, but the sample can be the entire population. This is a distinction which leads to two formulas for something called the standard deviation which will be seen later.

The average is never just the average. There are at least three commonly used averages which we will cover in this class. The average measures the central tendency of a set, which means that it describes where the numbers tend to be. It stands to reason that three averages probably measure slightly different things about that tendency. There will be more discussion in coming slides.

The distribution is an important concept and critical in developing models. It describes how outcomes are distributed. Of course, this uses almost the exact word it is meant to define so lets dig a little deeper. When you have outcomes of an event – coin tosses, games, catastrophic economic collapse on a global scale – the outcomes can be noted and, when performing the same test over and over again, the number of each outcome can be counted. When the outcomes are all noted and counted, they form a distribution.

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



NOTES:

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.



NOTES:

Descriptive Statistics

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

- How much of a starting salary to expect
- Salary averages
- Variation in salaries

- *Descriptive statistics emphasizes measures of central tendency.*



Here is that phrase, central tendency, again. When we have a set of numbers, in this case starting salaries as reported by various companies, the central tendency describes how those numbers cluster together. That is the point of descriptive statistics – describing how things fit together.

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.***



This is much more complex than simple descriptive statistics. The bulk of this course will ignore it, but it is important to be aware of the distinction. Inferential statistics has a broad array of tools and techniques used to come to conclusions about small samples of much larger populations. This is realm of medical studies for pharmaceuticals etc.

Types of Variables

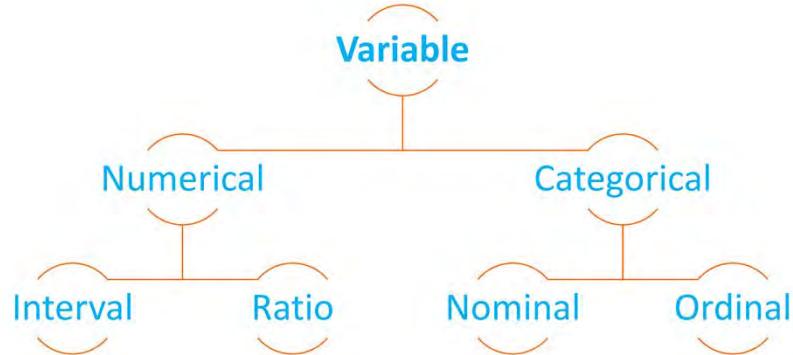
We look at what kinds of variables are involved in data analysis

Section 3



NOTES:

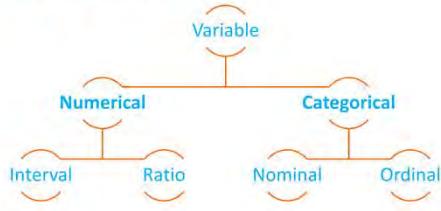
Definitions: Variables



Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

NOTES:

Definitions: Variables



Categorical: categorizes or describes an aspect of a population

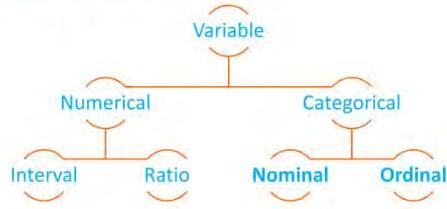
Numerical: quantifies an element of a population



Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

NOTES:

Definitions: Variables



Nominal: categorizes, describes, or names an element of a population.

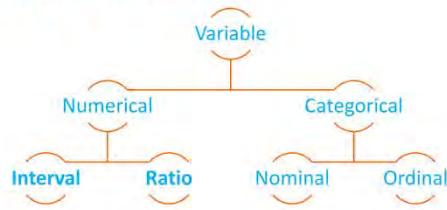
Ordinal: determines an ordered position or ranking



Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

NOTES:

Definitions: Variables



Interval: the distance between two values is meaningful (Celsius on the thermometer)

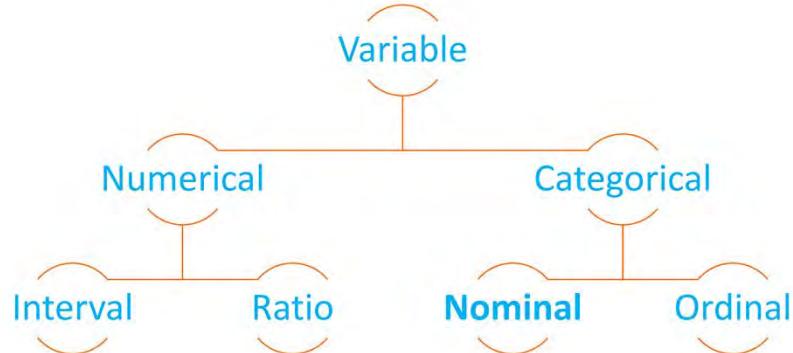
Ratio: possesses a unique, meaningful and non-arbitrary zero value (Kelvin on the thermometer)



Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

NOTES:

Definitions: Variables



Let's go through these in some details with examples – starting with Nominal...

NOTES:

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.**

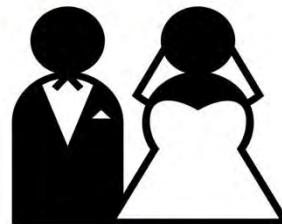


Eye color: blue, brown, green, etc.; Biological sex (male or female)

NOTES:

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!!!)



All of these are categories. Yes, the second one of these has numbers, but they are really just short hand for the labels. They are not really ordered numbers.

Notice that there is nothing here that is ranked in order of importance or quality or anything else. They are just labels.

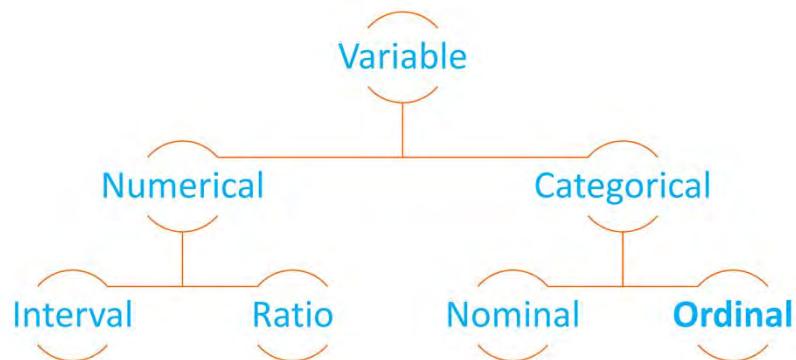
Nominal Variable: What Statistics/Operations Can I Compute?

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



This slide describes what calculations we can perform on nominal variables. Notice that it is not much. We can only describe the frequency distribution. If you do not already know what that is, don't worry, we will cover it later. For now, just notice that nominal variables are limited in what they can be used for.

Definitions: Variables



NOTES:

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



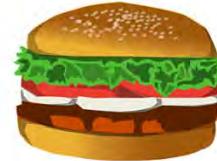
As discussed, order matters here. Consider data that is collected from Jan 2016 through May 2017 every month. When displaying in a chart, January 2016 comes before January 2017, so any outcomes in 1/2016 should be displayed before 1/2017. We should not necessarily group all January data together regardless of year. Notice that January is still just a label. The labels do have a time ordering to them which ranks the outcomes in the labels.

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.**

Also notice that the difference between first and second is not necessarily equivalent to the difference between second and third, or between third and fourth. You may love hamburgers slightly more than chocolate, but both way more than sushi and papaya.

Ordinal Variable: What Statistics/Operations Can I Compute?

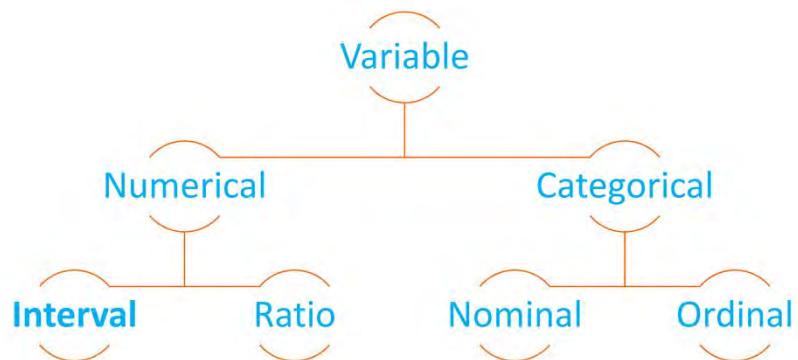
Statistic/Operation	Compute?
Frequency Distribution.	Yes
Median And Percentiles.	Yes
Add Or Subtract.	No
Mean, Standard Deviation, Standard Error Of The Mean.	No
Ratio, Or Coefficient Of Variation.	No



What can we compute? Well, since we have not covered the techniques yet, just notice that we have added one more group from our list by rank ordering the set.

NOTES:

Definitions: Variables



NOTES:

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.**



Now we are in the variables that work with numbers. You can imagine that this will open a new set of calculations we can perform on our set. These are some general statements about interval variables. Note that with numbers, we can define explicit intervals and make them equal in size (1-10, 11-20, 21-30, etc)

Interval Variables: Examples

Some Temperature Scales

- **Fahrenheit**
- **Celsius**



ASPE
TRAINING
a division of Foothills College

It is meaningful to say that 25 degrees Celsius is 3 degrees hotter than 22 degrees Celsius, and that 17 degrees Celsius is the same amount hotter (3 degrees) than 14 degrees Celsius. Notice, however, that 0 degrees Celsius does not have a natural meaning. That is, 0 degrees Celsius does not mean the absence of heat!

Interval Variables: Examples

Likert scale:

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!



A Likert scale is common in surveys. This is an example of a survey question with a numerical values mapped to statements. They are often balanced around the middle point. Here you see 3 is the midpoint and there are two options higher and two options lower.

Interval Variables: Which Statistic/ Operation Can I Compute?

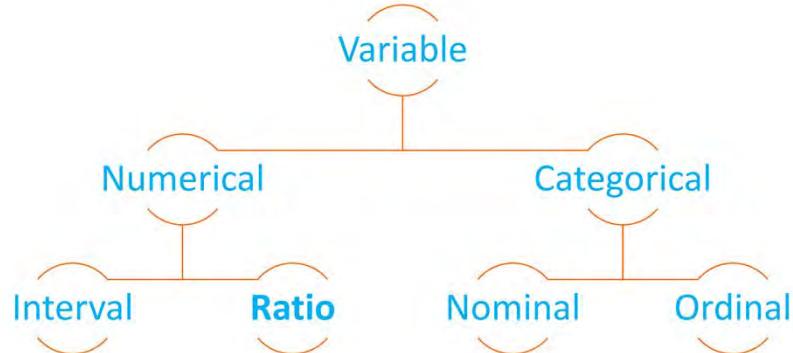
Statistic/Operation	Compute?
Frequency Distribution.	Yes
Median And Percentiles.	Yes
Add Or Subtract.	Yes
Mean, Standard Deviation, Correlation, Regression, Analysis Of Variance	Yes
Ratio, Or Coefficient Of Variation.	No



Now we have greatly expanded what we can do for analysis. Numbers give us a lot of power.

NOTES:

Definitions: Variables



Lastly...

NOTES:

Ratio Variables

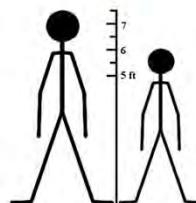
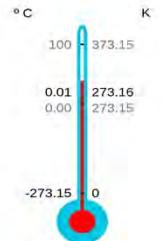
- This is the highest level of measurement and has the properties of an interval scale; coupled with fixed origin or zero point.
- It clearly defines the magnitude or value of difference between two individual items or intervals in same group.



NOTES:

Ratio Variables

- Temperature in Kelvin (zero is the absence of heat. Can't get colder).
- 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.



We already saw Kelvin, now we add height (no one is negative inches tall), and heart beats (no one has negative heart beat).

NOTES:

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



And with ratio variables, we now have the spectrum.

NOTES:

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



The summary of everything covered to this section about the utility of variable types.

Central Tendency of Data

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

Section 4



NOTES:

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**



When confronted with most sets of data, we want to know how it tends to cluster together. Where are the most points? Now we are speaking of this independent what the data actually measures, so it is an abstract concept, which is what makes it powerful. By powerful, we mean that applying the same math to vastly different sets allows you to draw conclusions about those sets. Remember, for the information to be useful, you have to know something about the field the sets are from. Knowing the average of the number of particles emitted from a radioactive isotope and the average of the population of North American cities doesn't really lead you to any conclusions without knowing the problem.

(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**



This is what most people mean when they say average. The mathematical definition is next.

(Arithmetic) Mean

Given a data set of

1,2,3,3,4,11

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



We just add the numbers in the set together and divide by the number of elements in the set. The result is a single number that we set clusters around. Note that the numerical result may not actually actually be in the set (the average of 2 and 3 is 2.5).

Notice what happens if I add another number, say 215, to the set. What happens to the average?

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

$$\text{Median}(1,2,3,3,4,11) = \frac{3 + 3}{2} = 3$$



This is another way of clustering around a single number in a set, but here, the number is in the set. This average is insensitive to outliers because adding another number that is really large or really small compared to the existing set doesn't significantly change the actual count of the data by much, so the "half above and half below" value is not affected much.

Mode

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

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



This clustering description also generates a number that is in the set. It is also insensitive to outliers because outliers don't usually occur multiple times.

Basic Probability

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

Section 5



TECHTOWN



NOTES:

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.**
- **Generally expressed as a number between 0 and 1, where 0 means no chance of occurring and 1 means it will certainly occur.**



Probability is just a game of counting. It describes such things as

Heads vs Tails

Number of times you will draw an ace out of a deck

Likelihood you will roll a 7 on two dice

How likely are you to get a red 32 in roulette

Are you sensing a theme here? There is a reason. Gambling was a primary motivator behind the study of probability.

You often hear things like a 10% chance of rain, but it is important to remember that probability is really measured between 0 and 1. We just tend to multiply by 100 when we present results to non-analysts.

Probability Uses In Business

- **Quality Control (Manufacturing Defects, etc.)**
- **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)**



All of these are commonplace applications of probability (and statistics).

Probability: Several Ways We Can Calculate It

- **Simple probability calculated from frequencies**
- **Conditional probability from contingency tables**
- **Conditional probability from tree diagrams**

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%

Geographic Location	Years of Experience			
	0-5	6-9	10 or more	Managers
Midwest	4	3	2	9
East Coast	7	5	3	15
West Coast	12	10	7	29
Total	23	18	12	53



ASPE
TRAINING
a division of North College

The charts pictured here show 3 common ways of calculating probability

- Calculated from Frequencies – the chart shows how often something of a particular color is observed (i.e. a marble in a bag of marbles) and the total count of marbles
- Contingency Table – the chart tracks two factors in a matrix. The probabilities are read from the totals
- Tree Diagram – tracking the different options through the branches of the tree

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



The event is the thing that we are performing (flipping a coin) and the outcome is the result of the event when measured (got heads). We assign the probability of getting the outcome of heads when we have the event of flipping a coin (in this case, it is $P=0.5$).

The sample space describes all of the possible outcomes that an event can have.

“Independent” will be an important term for us. The notion that one coin flip does not affect the outcome of another coin flip, or that a person’s height does not depend on another randomly selected person’s height, is very powerful and leads to insightful mathematics.

Calculating Probability

- Assuming a sample space of equally likely events

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

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

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

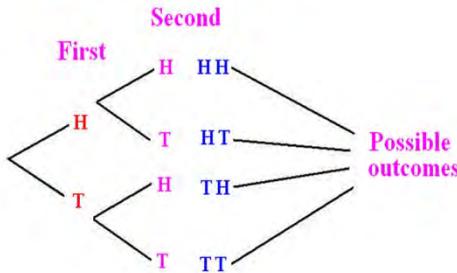
- If I flip 2 coins, what is the sample space and what is the probability I get at least one head?



The math described here speaks for itself, but it is one of the first mathematical descriptions, so spend some time understanding the notation. $P(A)$ is the probability of an outcome that we label as A. The probability is calculated by dividing the number of ways that A occurs by the total number of possible outcomes.

When flipping a coin, heads can only occur one way and there are two possible outcomes for the coin toss. Therefore, $P(H) = 1/2$

Calculating Probability



$$\bullet \quad P(\text{At least one head}) = \frac{\text{Outcomes of HH or HT or TH}}{\text{Total outcomes}} = \frac{3}{4}$$



This one is harder. Here we are flipping 2 coins and we want to know the probability of getting at least one head. Using this tree diagram makes it easy to visualize the space of outcomes.

The way we read this is:

Follow the line to H for the first toss, then we can follow the line to H for the second toss and get a possible outcome of HH

Follow the line to H for the first toss, then follow the line to T for the second toss to get a possible outcome of HT

Repeat for T as the first toss.... TH and TT

There are four possible outcomes and 3 of them have an H in them.

Calculating Probability from a Contingency Table

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

Mode of travel	Amount spent on drink			
	None	\$1 and under	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

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



The contingency table allows the analyst to read the count of the outcomes directly from either the row or column. In this example, to determine the probability that a customer spends less than \$20 for refreshments is found by reading the last row (Total) for all columns where the money spent is less than \$20. Adding up the counts in the two columns is the outcome count for this outcome. Then divide by the total count of outcomes in the last column.

Can you repeat this to answer the question “What is the likelihood that a person will spend less than \$20 given that they took the bus to the store?”

Conditional Probability

Mode of travel	Amount spent on drink			
	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?**

$$P = \frac{10}{70} = 0.143$$



A little more practice but this involves the challenge in the last slide. The part about the "given that they took the bus". Here, there are 70 people total who came on foot and 10 of those spent at least \$20. This is the conditional part... the "given that". Now try the challenge question from the previous slide.

Conditional Probability

Mode of travel	Amount spent on drink			
	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

More Formally ($A = (>\$20)$, $B=(\text{On foot})$)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{10}{70}$$



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

- $E(x) = \sum_n P(n)x(n) = 0.27 \times 105 + 0.30 \times 100 + 0.23 \times 156 + 0.20 \times 130 = 120.23$



The "Expected Value" of this set of numbers is defined by the probability with which they occur. Given this set of numbers and the probability that each number occurs in the table, the value one would expect on any observation is defined by the sum of the products of $x(n)$ and the $P(n)$. In this set of observations, the expected value is calculated to be 120.

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



This is the definition of the frequency distribution. It is the workhorse of descriptive statistics. First we define categories, then we observe something. We assign each observation to a category and then count how many are there. That's all there is to it!

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



Be careful about how you interpret though, the type of variable limits the analysis that can be done.

Frequency Distribution

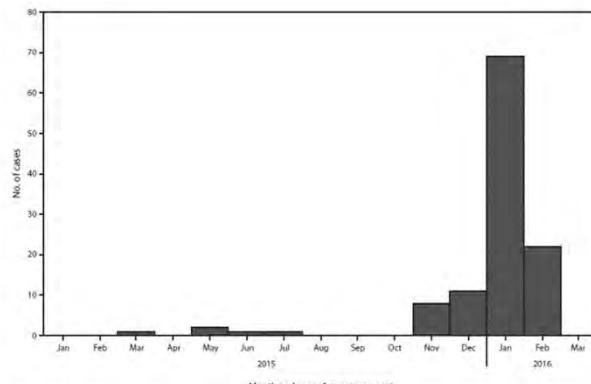


Figure. Month of illness onset for 115 patients with laboratory evidence of Zika virus infection among residents of U.S. states and the District of Columbia — January 1, 2015–February 26, 2016*



The distribution you see here has two important points to note:

- The buckets (bins) are time ordered, or ordinal, values
- The sum of all bins is 115 patients

Additionally, see how the x axis is laid out over two years. This is an example of where it is important that we do not combine both Januaries together because the time ordering tells us something important.

Distributions

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

Section 6



NOTES:

Distributions

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

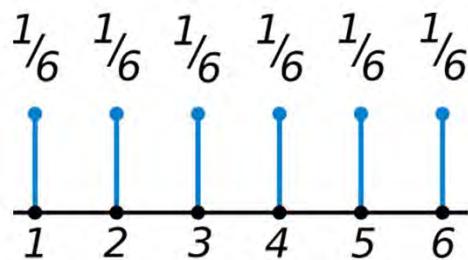


When frequency plots categorize all of the population, they lead us to distributions. Distributions show how all of the population is distributed among the space of outcomes. They can be discrete, where the outcomes are small and countable, or continuous, like the real number line. When looking at a distribution, the frequencies of each outcome added together should add to 1.

Discrete Distributions

- **Probability Mass Function - probability of a particular X value occurring.**

A fair die

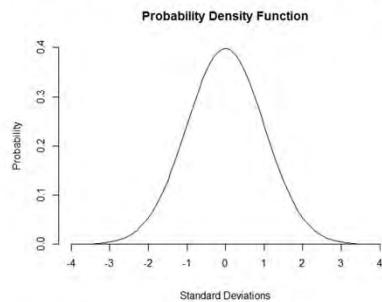


You can determine the probability that a discrete variable will be exactly equal to a particular value.

In this case, we are looking at a single six sided die. The probability of rolling any number on that die is shown in the PMF pictured. When we add all of the probabilities, we get $6*(1/6)$ to get 1.

Continuous Distributions

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

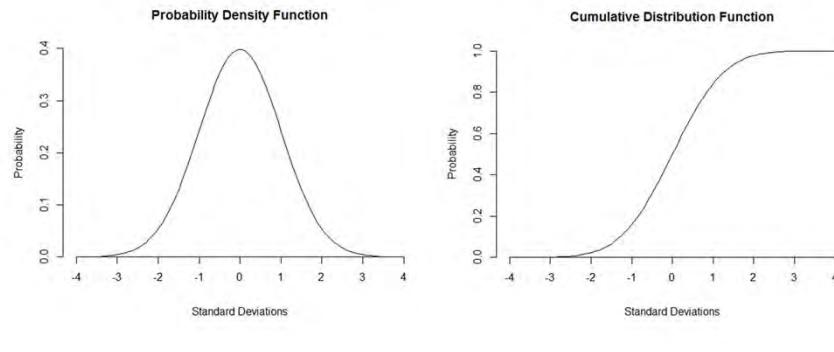


You can determine the probability that a continuous variable will be equal to a particular value by summing (integrating) over a range. In this case, we are looking at a what's called a normal distribution.

We will not be integrating in this class, we will let the tools do it for us, but it is important to understand what it means. In this distribution, we must integrate from -4 to 4 to obtain 1. Notice that there are an infinite number of points between any two numbers on the scale.

Continuous Distributions

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



The continuous distribution on the left is the probability density function that we saw in the last slide. The function on the right is what happens when we integrate the one on the left at each stage. It is cumulative, or it keeps adding, as we move from -4 to 4. When you use Excel functions to sample distribution functions, the cumulative function is often used.

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.**



We are building up the basic tools for analysis. So far, we have learned the basics of probability and how they lead to distributions (of which, we will cover a number of them later). Now we begin to describe those distributions.

Since distributions are the plots of outcomes, they are made from data. Real data always has variability because the world is a messy place. Statistics helps us characterize the variability. By doing this we can determine such things as quality in a manufacturing process, or the likelihood that some even will happen in the future.

What follows are some definitions of concepts that define the variability of data.

Range

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

$$Range(1,2,3,3,4,11) = 11 - 1 = 10$$



The Range includes all outliers. The definition of an outlier will come later, but it is important to remember when characterizing a data set that the range, when reported, includes the entire set.

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



NOTES:

Variance

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



The Variance, defined in this slide, is the Expectation Value of the squared deviation from the mean. Compare the definition above with the definition of the Expectation Value and note how the values of $P(n) = 1/N$ for each observation in the data set. Then $x(n)$ is the squared difference between the mean and the value of each observation.

Variance

Variance(1,2,3,3,4,11)?

- **Take the Mean ($24/6 = 4$)**
- **Sum the squared differences with the mean**
$$(1 - 4)^2 + (2 - 4)^2 + (3 - 4)^2 + \dots = 64$$
- **Divide by the count (6)**
$$64/6 = 10.667$$



Standard Deviation

- The square root of the variance
- It is scaled to the observations and is also positive
- Standard Deviation of (1,2,3,3,4,11)?

$$\begin{aligned}\sqrt{\text{Variance}(1,2,3,3,4,11)} \\ = \sqrt{10.667} = 3.266\end{aligned}$$



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.



Remember earlier in the course when we talked about population and sample? We consider analysis when we have knowledge of the full population for most of this course. Often, the real world considers samples of a population. This leads to a subtle difference functions for the standard deviation and the variance. The difference is in the use of Bessel's Correction.

The explanation of the N-1 (Bessel's Correction) is complex, but is related to the fact that you are estimating the mean of the population from the sample mean. This introduces bias into the calculation that is corrected by the N-1.

Application of the Standard Deviation

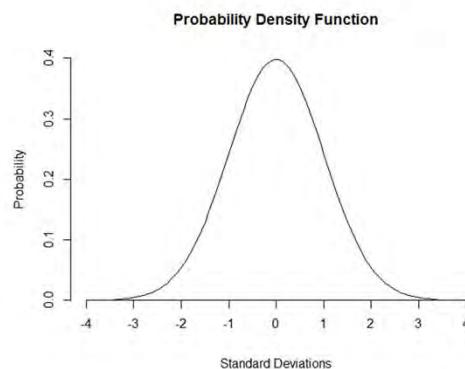
- **Measures deviation of any distribution**
- **Low value indicates values are close together**
- **High value indicates values are spread over wider range**
- **But... when applied to the Normal Distribution, it means specific things**



We talked about the variance and the standard deviation. They apply everywhere, but when applied to a specific distribution, the std dev tells us many things. So let's talk about the normal distribution.

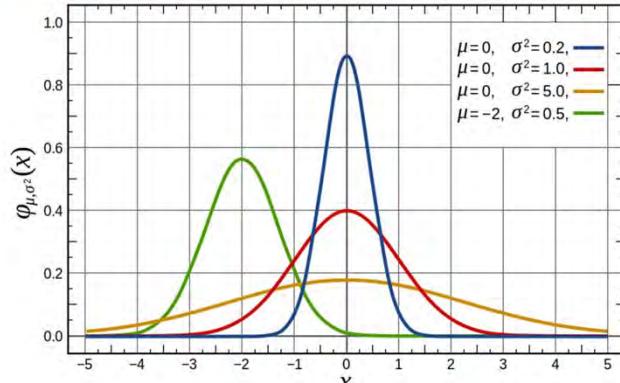
Normal Distribution

- Also called Gaussian
- Appears when events are independent
- Useful because of the Central Limit Theorem



Central Limit Theorem say, basically, that the outcome of a large number of measurements of independent random variables will be distributed along the normal curve seen here. The details are unimportant to the scope of the class, but because of this theorem, the normal distribution show up in a lot of places in nature, technology, business, etc. Most popular appearance is in the Bell Curve of IQ in the last couple of decades.

The Normal Distribution



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

Image is from Wikipedia https://en.wikipedia.org/wiki/Normal_distribution#Definition

This plot shows the normal curve and how it changes with the value of the mean (moves left or right) and the standard deviation (tall and thin, or short and fat).

The mathematical equation is not important for this class, but sometimes, people like to see what is happening behind the scenes. Notice that the funny little u character, called mu, is the arithmetic mean, and the o with loop on it, sigma, is the std dev. When we square the std dev, we get the variance. So this formula has the average, the std dev and the variance all in the same equation.

Standard Deviation and the Normal Distribution

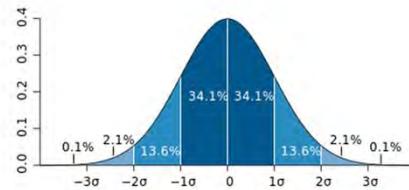


Image: https://en.wikipedia.org/wiki/Standard_deviation



As promised, the application of the standard deviation to the normal distribution shows some interesting information. What this shows is the percentage of the population of a series of observations that lie between the + and – values of the standard deviation. Those values are on the next page.

Sigma (σ) Values (Standard Deviations)

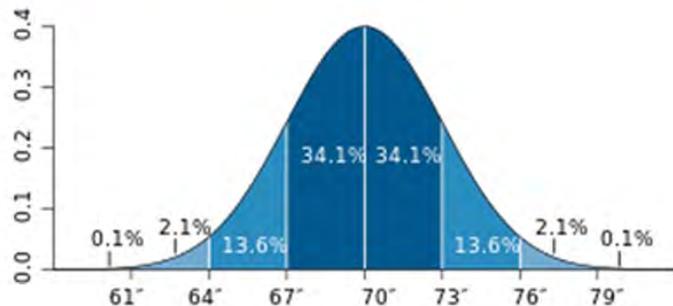
For Normally Distributed Data

<u>Sigma</u>	<u>Percent</u>
+/- 1 sigma	68.27%
+/- 2 sigma	95.45%
+/- 3 sigma	99.73%
+/- 6 sigma	99.999998%



That last number is why the organization that developed the process improvement tools for business calls the suite Six Sigma.

Estimated Heights of Men



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

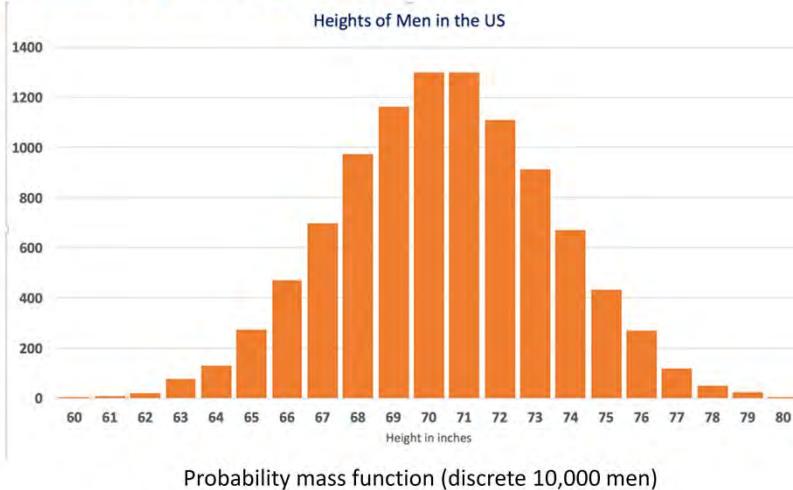
Image: en.wikipedia.org/wiki/File:Visualisation_mode_median_mean.svg



When we measure the heights of randomly selected men in the US, we find the normal distribution.

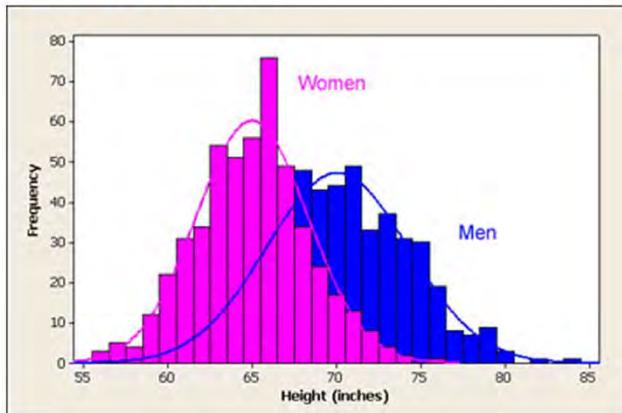
The standard deviation of this data set is 3 inches. Just over 68% of the men measured lie between $70 - 3$ inches and $70 + 3$ inches

Estimated Heights of Men



Here is a plot of the measurement of 10,000 men. It is discrete (countable to 10,000) but notice the shape it makes. When you measure another 10,000, you get more accurate distribution and if you take the buckets and subdivide them into decimal intervals, you will get a more continuous looking curve.

Men and Women

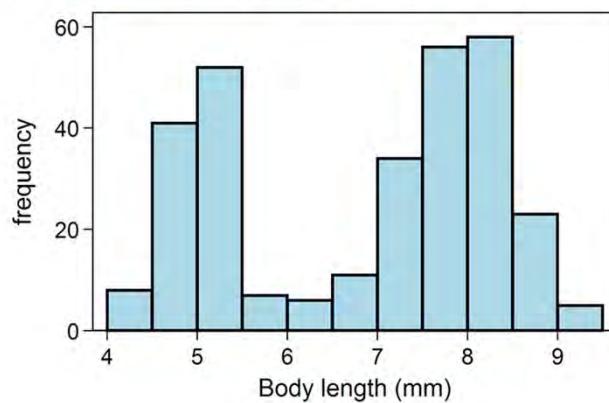


Men and women create two normal distros, or a bimodal distribution



This image shows measurements made of men and women. The bars in the histogram show the actual counts of the distribution (the discrete data set) and the lines show the fit to the distribution. This is very often the case. The discrete counts match more closely the fits when the number of people measured increase. It is unknown how many were counted in this sample.

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>

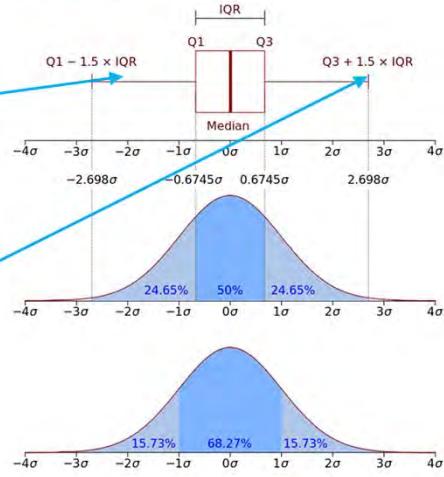


This is a bimodal distribution. They are built from two different normal distributions with two different means laid on top of each other.

Quartiles and the Normal

IQR from Quartiles slide

1.5 is variable value
based on field of study



The box plot is a typical way of describing the entire distribution of a data set graphically. It shows the relationship between the quartiles, the IQR, whiskers (the outer vertical lines) and any points that may lay outside these whiskers – commonly defined as outliers.

Normal but Slightly Skew

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

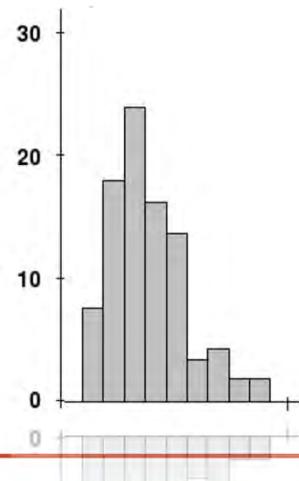


Image: en.wikipedia.org/wiki/File:SkewedDistribution.png

When a normal distribution is weighted more heavily on one side, we say it has skew.

Positive vs. negative skew

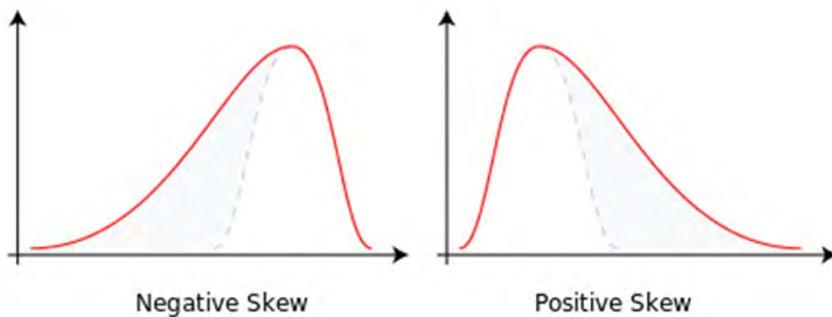


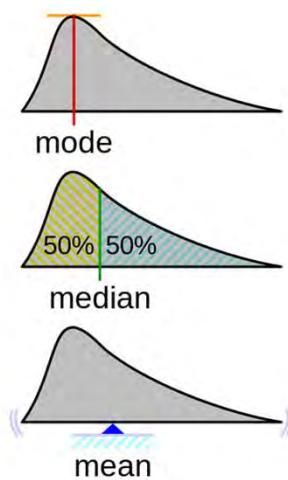
Image: <https://en.wikipedia.org/wiki/File:SkewedDistribution.png>



Be careful what we call positive and negative skew, it follows the tail.

NOTES:

So?



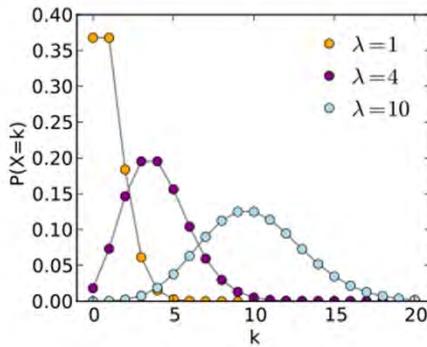
<https://en.wikipedia.org/wiki/Skewness>



When we have a skewed distribution, the mean, median and mode tell us something about the skew. Here we see that the mode is at the peak of the curve. The median tells us the 50/50 point and the mean is the “balance” point. The more radically skewed the distribution becomes, the more these three numbers will deviate from each other.

Poisson Distribution

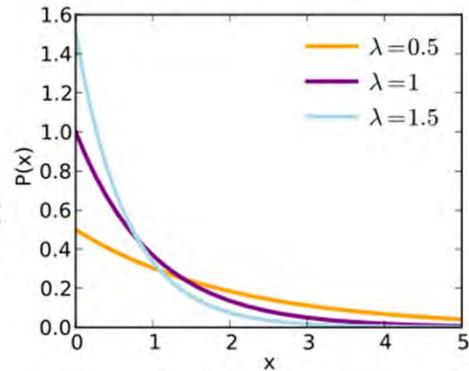
- Independent Events with a known average rate
- Calls between 11AM and Noon at a call center



Another type of distribution is formed from two constraints on the normal distribution. First, have positive only events (0 or above). Second, have few events. With these we see something called the Poisson distribution. The example here is a call center where we count the number of calls between 11am and noon. The average rate is low and the events are independent and we cannot have negative number of calls.

Exponential Distribution

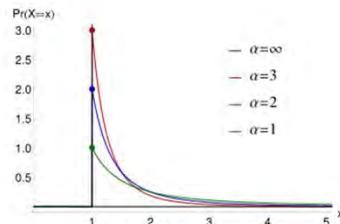
- Shows decaying probability
- Time until the next call at a call center



This is a different type of distribution altogether. It shows that fall off, or decay, as the x variable gets larger.

NOTES:

Pareto Distribution (“80/20”)

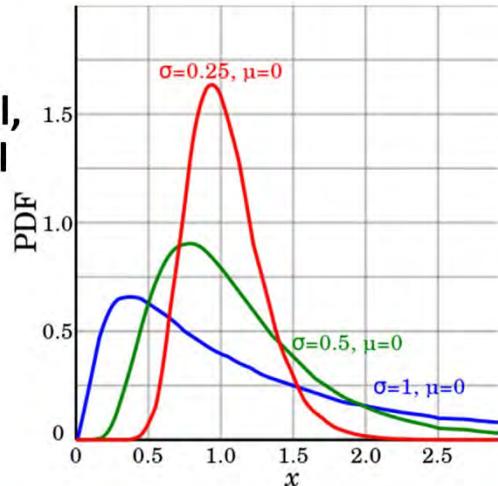


- Special type of Power Law distribution
- Wealth distribution
- Size of meteorites

The Pareto comes from a trend seen in manufacturing but applies to many things. We know it as “80% of the work takes 20% of the time and the remaining 20% of the work takes 80% of the time”.

Log Normal Distribution

- If X is Log Normal, $Y=\ln(X)$ is Normal
- Dwell time on online articles
- Repair times



NOTES:

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**



Now that we have seen a number of distributions, let's look at how we can use Excel to extract information out of them. What follows is a series of examples of how to use Excel formulas for the distributions.

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)



This Excel function allows us to pass the number of interest (x) and define a normal distribution with the mean and std_dev of our choosing. The last argument tells us whether we are using less than and equal to (TRUE) or just the probability density at that point (FALSE). FALSE requires more advanced understanding of the PDF, we use TRUE throughout.

Distributions in Excel

What's the probability of measuring a random man at less than or equal to 67 inches tall ?

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



NOTES:

Distributions in Excel

What's the probability of measuring a random man at 67 inches tall ?

NORM.DIST(68, 70, 3, TRUE)

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



We find the value at 68 inches and subtract the value at 67 inches because we are only interested in whole number inch values (not 67.2 or 67.6, etc).

NOTES:

Distributions in Excel

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

- BINOM
- POISSON
- LOGNORM
- T
- CHI



The "1 – " is because

- All distributions sum to a probability of 1
- The formula with cumulative set to true gives us "less than or equal to". The "greater than" part is 1 – "less than or equal to"

Fitting Data

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

Section 7



From here we are moving into basic modelling. We will look to fit math equations to data and notice what these fits can tell us. Once we know the math equations, we can make predictions about what is to come. This power is limited for a number of reasons which we will discuss along the way.

Bivariate Data

- **Scatter plots**
- **Covariance and Correlation**
- **Regression**
- **R²**



The first thing we will look at is data of two variables. Our best visual aid here is often the scatter plot where we put one variable on the x axis and another on the y axis. We can look at the covariance and correlation of the data set. We use regression to calculate the mathematical fits to the data, then we look at R² to see which is the best one. Be cautious, the literal best fit may disobey rules of the field you are working in. It is, as always important to remember that the math is not the only thing at work here. What is the problem? What is field of research?

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 – $Cov/(\sigma(X)\sigma(Y))$

- Normalized by standard deviations
- Unitless
- Between -1 and 1



Covariance is, simply put, how the two variables vary together in the set. The formula is shown in the slide. We need it because we really want the correlation, which is the covariance divided by the std dev of both variables. Don't worry, our tools will do this for us.

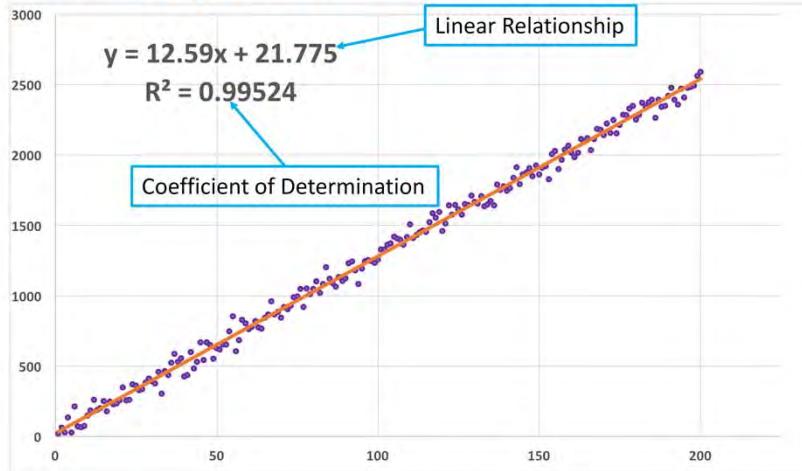
Simple Linear Regression

- **Finding a linear relationship which can predict the (average) value of Y for a given value of X.**
- **Coefficient of Determination (R^2)**



The “regression line” is a straight-line descriptor. If you have slope and y-intercept of the regression line, you can plug in a value for X and predict the average value for Y.

Linear Regression



This fit is calculated by minimizing the sum of the squared differences between each point and the line. This amounts to minimizing the Variance of the data set (see slide on Variance). This is called the **Method of Least Squares**.

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).



These are some warnings about curve fitting. One note about the moderate values of r ... in the real world, $r = 0.5$ is not enough for a good correlation.

NOTES:

Other Fitting Functions

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



Linear regression is a technique and we can fit a lot of different functions with it. Some of the function types are listed above.

NOTES:

Linear Fit

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

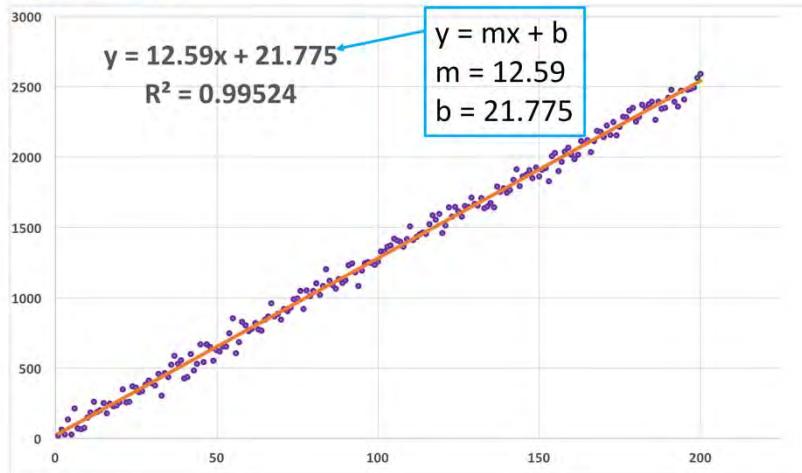
$$y = mx + b$$



The simplest is the straight line. The equation of the line, when you show it in Excel looks like the equation here, except m and b will be numbers determined by the fit.

NOTES:

Linear Fit



This is what Excel will display.

NOTES:

Polynomial Fit

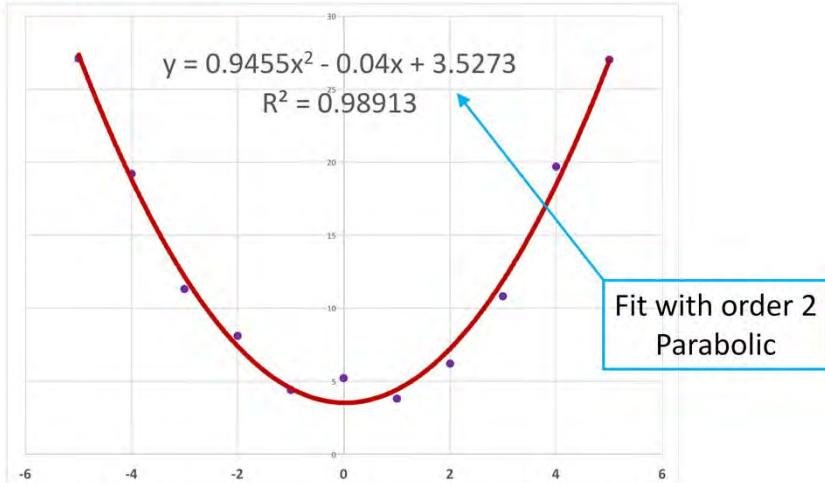
- **Can fit anything with enough parameters**
- **Be careful with extrapolation**

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



A polynomial is an equation with higher powers of the x variable (the independent one). Be very careful with this... you can almost always fit any data set perfectly by going to enough powers of x, but it does not mean that you are correct. Certain fields of study obey known laws or trends and these determine the type of equation you should use. A bad fit of the data to this equation could mean you have bad data, not require higher order fits. You, as the analyst, should know as much as you can about the field to determine these things.

Polynomial Fit



This is an example of a parabola. This is really common in studying new materials in physics and engineering. Notice the equation has a squared x term in it.

NOTES:

Power-Law Fit

- **Income counts histogram**
- **Frequency of words in Text**

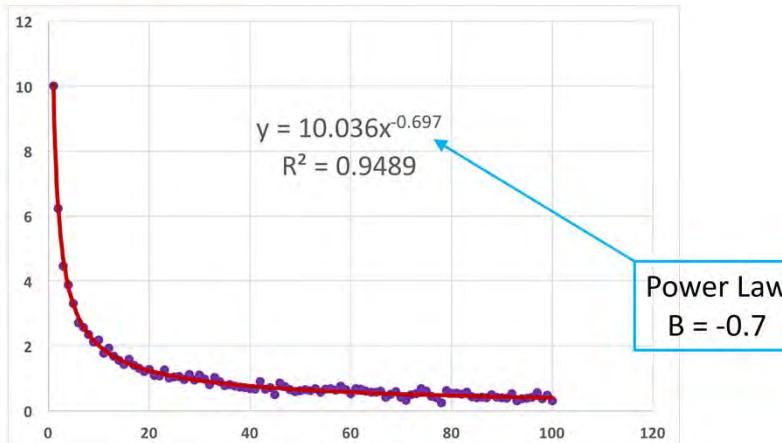
$$y = Ax^B$$



The power law is a fit where B is some number (2,3,4,5,6,etc).

NOTES:

Polynomial Fit



Here is one where B is negative and a fraction.

NOTES:

Exponential Fit

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

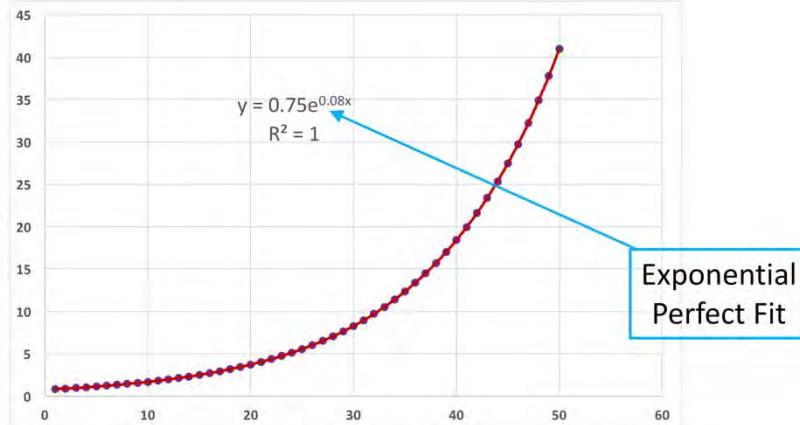
$$y = Ae^{Bx}$$



The exponential fit is a common fit to data such as the examples stated.

NOTES:

Exponential Fit



Continuously compounded interest? This is how much money I owe?!

NOTES:

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.**



Remember, choosing the fitting function is more than just trying the functions until one gets a high R² value. It is a powerful tool, but you are fitting things that have already happened. Be careful about predictions you make from it.

Cluster Analysis

We look at another way of drawing conclusions from visual data

Section 8



Let's look at a different way to visualize data. It still involves scatter plots, but the analysis is different than simply looking at a data fit.

Clustering

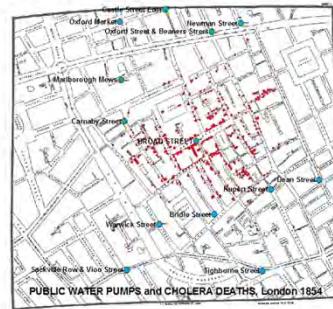
- **Finding common groupings in a multi-dimensional dataset (segmentation)**
- **Can be used to find patterns that are difficult to identify by eye**
- **Major factor in machine learning**



Clustering is where we apply a number of algorithms to find patterns in multidimensional data sets. These patterns are instrumental in machine learning. Remember that we mentioned credit card fraud some time ago in the course, well, that type of analysis is a major clustering and pattern recognition algorithm in which analysts look for patterns and anomalies from those patterns.

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>



In one of the most famous cases of pattern recognition from clustering analysis, a doctor in 1850's London stopped a deadly disease and at the same time proved the germ theory of disease.

John Snow investigated the cause of a cholera outbreak by documenting where the cases were located. Discovered that people who lived near, or travelled by, the Broad St. water pump were much more likely to get cholera. A nearby cesspool had leaked in to the water supply.

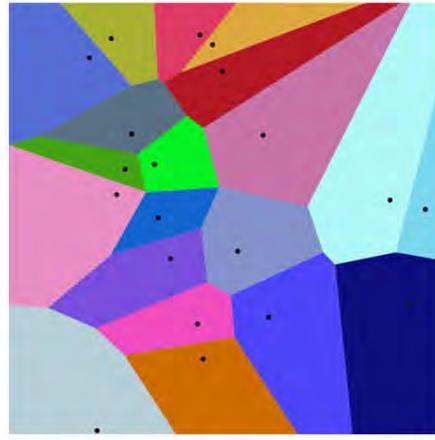
Left – Snows original map of cholera cases in London

Right – Cases shown in red, Water pumps shown in blue.

From: <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 an algorithm (e.g. K-means)**



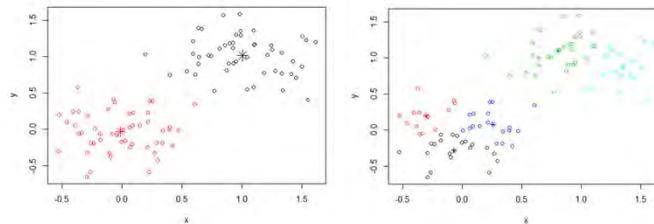
https://en.wikipedia.org/wiki/Voronoi_diagram



A Voronoi diagram provides a visual map of the distance of any point in space to a specified center point in a region. That center point is usually the mean of a cluster, though it could also be something called a medioid. The regions in the map above show areas in a color in which all points in that color area are closer to the black dot in that same color area than they are to any other black dot. This has been applied to John Snow's map and the computers arrive at the same conclusion he did about the source of the epidemic. Today, these algorithms can tell ambulances the location of the closest hospital in real time, and many other things.

K-Means

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



These charts are 2-D examples from R with 2 or 5 centers. The K-Means algorithm works in iterations. It calculates distances from a chosen “center”, then moves the “center” then calculates distances again, over and over, until the next iteration does not improve the accuracy any. At that point, in K-Means, you have the mean values of clustered data sets.

Monte Carlo

We discuss a few things that take us into advanced prediction

Section 9



Up to this point in the class, we have been looking at modelling data for explanatory power, but with limited predictive potential. There are many methods by which we try to add prediction to descriptive data and we have chosen to look at a common one with a simple fundamental premise – Monte Carlo.

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**



The idea behind Monte Carlo is that if you play games of chance long enough, probability defines how often you will win or lose. The only unknown, really, is the model of the game you are playing. In the case of black jack, you are looking to add to 21 given the cards in a deck and the players in the game. Of course, you can never know for sure how one game will end up, but you can calculate the likelihood that you will win any game, and play enough games, you will win that many times.

In Monte Carlo, we develop a model of our system and we use computers to "play a game" with that model over and over again. We generate random inputs and send them into the model. Then we keep track of what these results are and start adding and subtracting, multiplying and dividing to "ask the questions" of the results. Don't worry, we will keep our models simple, but in the real world, the power comes from developing complex models and letting probability and powerful computers find our answers.

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**



Here is a brief description of the steps involved. Note that here, PDF is the probability density function from the distribution of results, not the document format.

The next series of slides shows how to do it in Excel.

Monte Carlo Method

- Data set is 300 previous jobs from DB records
- Analysis shows
 - Average = 20 days
 - Standard Dev. = 3 days
 - Distribution = Normal

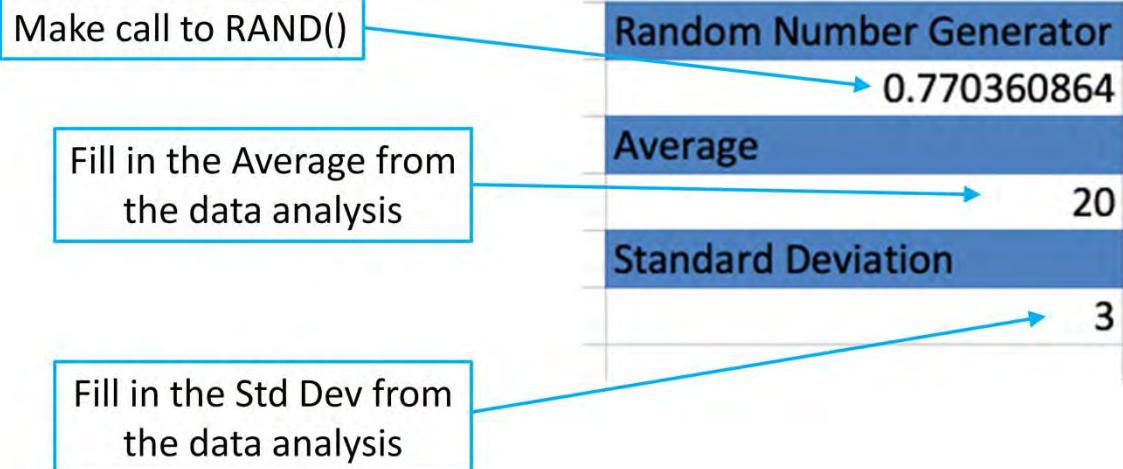
Job Number	Days To Deliver
1	18
2	20
3	22
4	17
5	20
6	19
7	20
8	19
9	15
10	23
11	21
12	20



The data set is 300 records, only some of which are pictured here. The first step is to analyze the data with the basics: Average, Std. Dev. and plotting the distribution.

NOTES:

Monte Carlo



NOTES:

Monte Carlo

Use these values in the model

Random Number Generator	Model
0.770360864	=NORM.INV(B3,B5,B7)
Average	22
20	
Standard Deviation	
3	



NOTES:

Monte Carlo

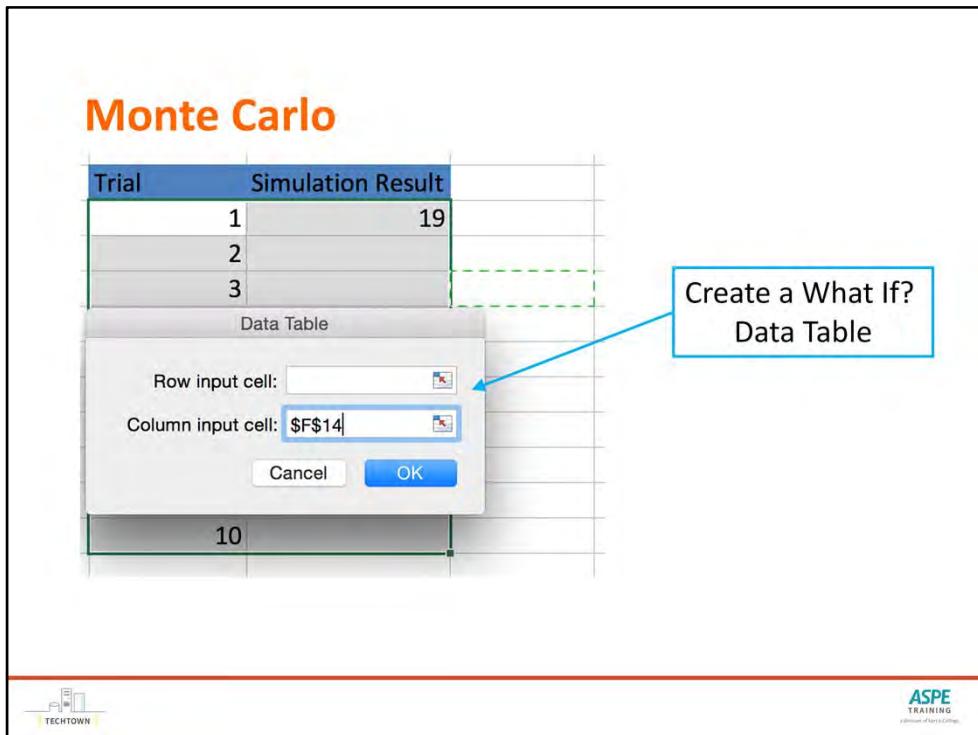
Model	
Raw Number	22.2201074
Rounded	22

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

Build a simulation table and reference the model



NOTES:



To create a What If? table, you need to click on an "off, empty cell" for the column input cell

NOTES:

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



In reality, this only works when we have many more trials (thousands, if not millions).

NOTES:

Monte Carlo

Query the results by finding relative frequencies

Trial	Simulation Result	Questions
1	19	$P(\text{Days} > 23)$ =COUNTIF(Simulation_Result, ">23")/10
2	18	$P(\text{Days} > 21)$ 3
3	17	$P(\text{Days} < 18)$ 3
4	22	
5	23	
6	12	
7	15	
8	25	
9	18	
10	19	

Using COUNTIF() and a Named Range

TECHTOWN

ASPE TRAINING

It is important to remember (cannot be stressed enough) that these numbers are not valid with 10 trials. The relative frequencies will converge as the number of trials grows very large.

R Programming

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

Section 10



Programming... this might scare you a little, but that is ok, we will not be doing much actual programming. This is more a section to motivate you to go beyond Excel. A real R programming course will take a substantial amount of time to cover.

What is R?

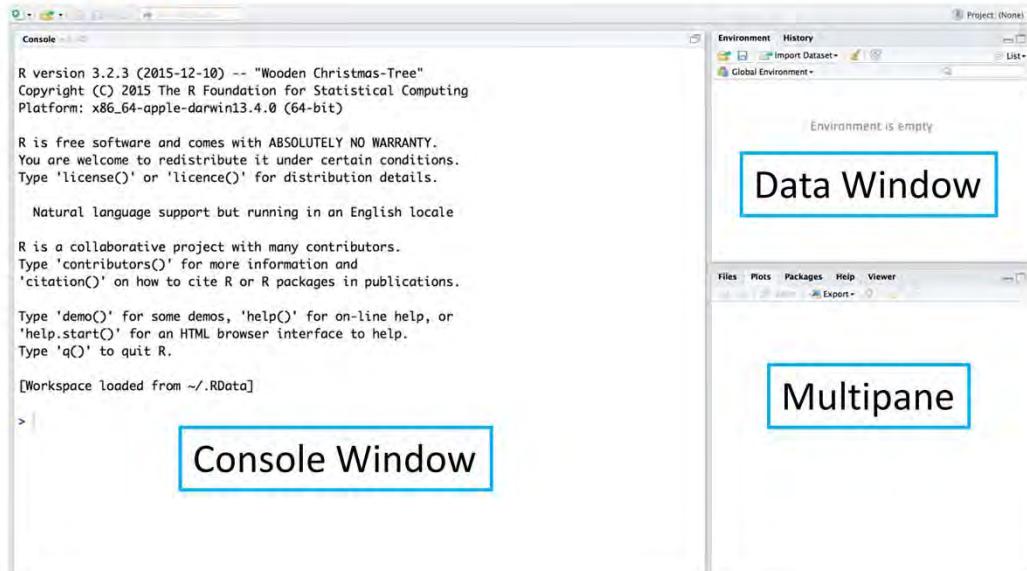
R is a programming language for

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



R is open source and freely downloadable from <https://www.r-project.org/>
We recommend R Studio to develop your code in, it is just easier to manage. It can be found
at <https://www.rstudio.com/>

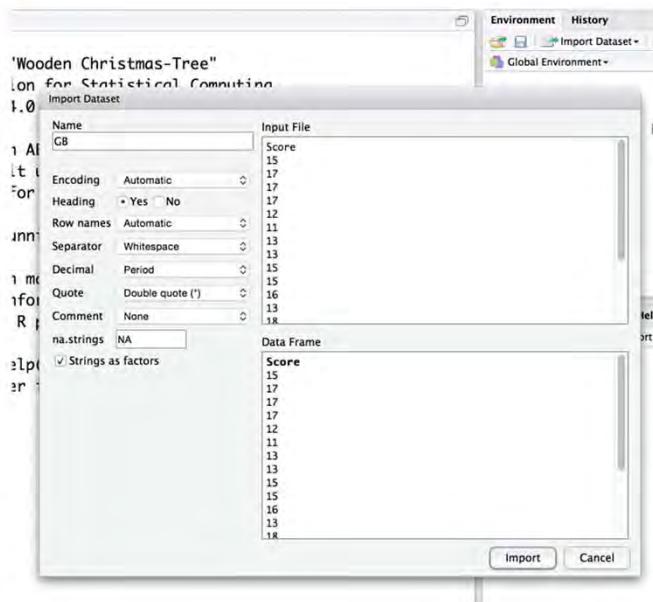
R Studio



This is a screenshot of R Studio. It is a development environment for code where commands are executed in the console window, data is available in the data window, and the multipane window shows help, charts, and a number of other information displays.

Basic Analysis in R

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



In R Studio, you can import data from a variety of sources – Excel files, CSV files and Tab delimited text files to name a few.

Here we are pulling in grade data from one of the CSV files in the labs.

Basic Analysis in R

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

```
Console ~/
> outOf100 <- GB$Score/20*100
> outOf100
[1] 75 85 85 85 60 55 65 65 75 75 80 65 90 70 75 75 65 7
0 75 75 80 75 90 95 80
[26] 75 90 75 85 75 70 70 75 75 80 75 75 60 90 75 55 85 8
0
>
```



This takes every row of data in the Score directory, divides each value by 20 and multiplies by 100, turning the value into a percentage. The arrow says to store the result of this calculation in a variable called outOf100.

Basic Analysis in R

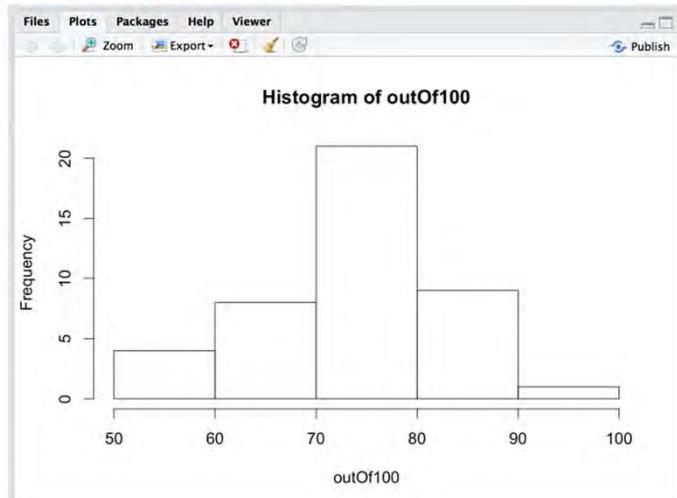
- Calculate central tendency
- Ask for help
- Plot histogram

```
Console ~/ ⌂
> mean(outOf100)
[1] 75.5814
> median(outOf100)
[1] 75
> sd(outOf100)
[1] 9.335627
> ?hist
> hist(outOf100, breaks=4)
```



NOTES:

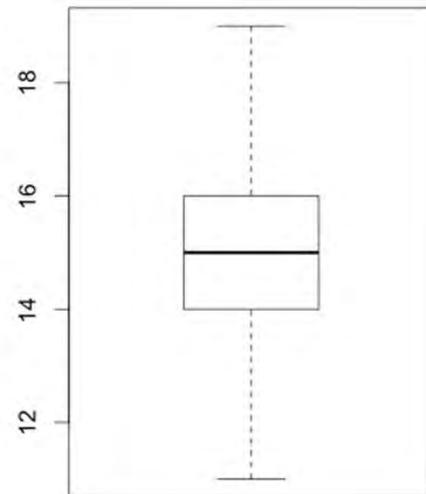
Basic Analysis in R



This is the result of the `hist()` command. It is basic, but R does have a lot of options for setting parameters like labels and ticks. Consult the documentation for more information.

Basic Analysis in R

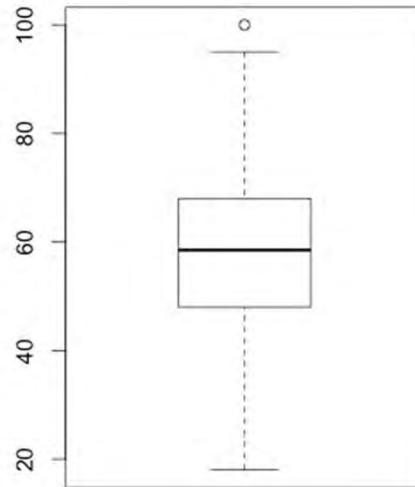
- Create a boxplot with
boxplot(GB\$Scores)



NOTES:

Basic Analysis in R

- A different set has an outlier



NOTES:

Monte Carlo in R

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

The screenshot shows the RStudio interface. The code editor window contains the following R script:

```
11 · concreteSlabTest <- function(iterations=1e4, myMean=20, mySD=3) {  
12   # this line generates random numbers (rounded) against the mean you pass in  
13   # and the standard deviation you pass in  
14   game <- round(rnorm(iterations, mean=myMean, sd=mySD))  
15  
16   # this line sums all of the parts of the game for days greater than 23 and divides  
17   # by the number of iterations (default is 10,000)  
18   numMoreThan23 <- sum(game>23)/iterations  
19  
20   # this line prints the percentage  
21   cat("Percent greater than 23 ", numMoreThan23*100)  
22 }  
23  
24 > source('~/Dropbox/Data Analysis Master/Exercises/7_Monte Carlo/Concrete Slab Co.R')  
> concreteSlabTest()  
Percent greater than 23 11.72  
> concreteSlabTest(iterations = 1e8)  
Percent greater than 23 12.16887  
>
```

The console window below shows the execution of the script and the output of the function calls.



This is a shot of the console window to create a function in R. Functions are ways to create reusable pieces of code.

NOTES:

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



NOTES:

Data Visualization

We talk about how to tell the story

Section 11



Hopefully, you have seen a number of ways to create data visualization by now, but to end the course lets talk about the final aspect of analysis, telling people about it.

Time to Tell the Story



After all the hard work , you have to communicate the results. Results are not just numbers, but also conclusions. You have to tell your story from problem definition to conclusion at multiple levels depending on who is listening.

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



A story has a beginning, middle, end, a plot, and hopefully, a reader. Telling your story is no different.

What problem were you attacking?
Why were you taking it on?
What did you do to solve it?
What assumptions, what data was used, rejected?
What was the conclusion?
And as for the reader, what level of depth do they require to "get it"?

THE GOAL OF VISUALIZATION

What you're really trying to accomplish

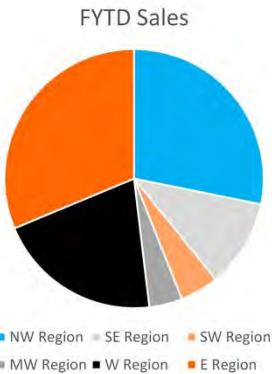
- **Efficiently communicate the right information to the right person**
- **Tell the story**
- **Critical Characteristics**
 - Has a Beginning, Middle, End
 - Has a punchline
 - Doesn't lose the reader (stakeholder)
- **Enable better decision making**



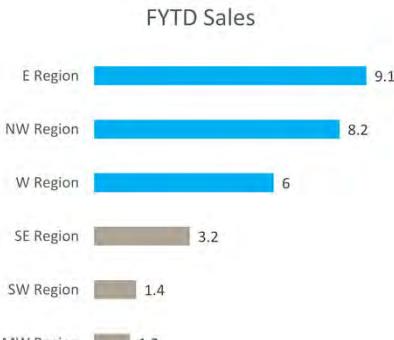
People see patterns very well. We create them where there is nothing a lot of the time (shapes of clouds for instance). Presenting data is leading people to the pattern and telling them the information necessary – not all the information. Charting is great for this. Tables of numbers don't work so well for most people. This is why we call it visualization.

Two Tales of The Same Information

Common go-to Object



"Bolder," more efficient choice



Here are two objects to display information. Which of the two looks “better”? Conveys information more effectively? Which one requires less eye movement and cognitive engagement to extract the information?

Hint: Pie charts are usually frowned upon in any scenario.

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



Knowing the user is critical. The project leader needs details, the CEO, not so much, they need broad strokes.

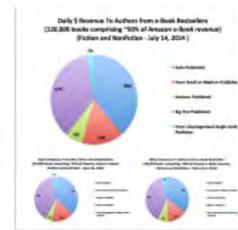
Stakeholder Cheat Sheets*

*Exceptions will apply

THE VERIFIER

- Likes Detail
- Tends to Question
- Methodical
- The Planner
- Resists Change

BEST BETS



These are quick reference slides to stakeholders and what they may like to see. Think about any organization you have been in and see if you agree and can match the personality type with the layout design.

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



ASPE
TRAINING
Executive Facilitation

NOTES:

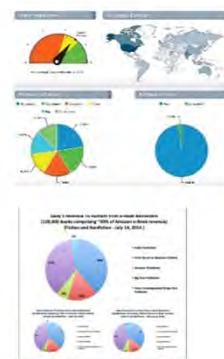
Stakeholder Cheat Sheets*

*Exceptions will apply

THE INTERPRETER

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

BEST BETS



NOTES:

Common Presentation Missteps

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



Avoids these when performing the analysis and preparing the presentation. Don't discard data that are outliers. They are important to indicate the outliers. Don't show graphs just to show graphs, concentrate on displaying the insights. All of the presentation should serve the 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**



NOTES:

Reports

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



NOTES:

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



NOTES:

Excel

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



NOTES:

KEEP IT SIMPLE

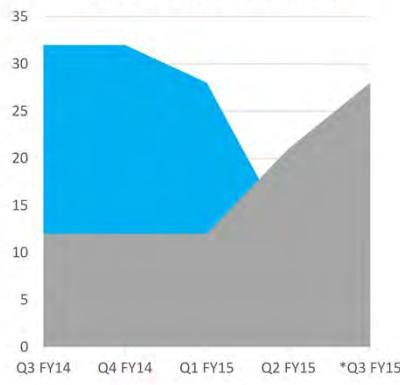
- Ask yourself the question: “Will my user ‘get it’ in less than 5 seconds?”
- Classic chart types are classics for a reason
- Don’t go crazy with color
- Take a minimalist attitude
- Remember the goal



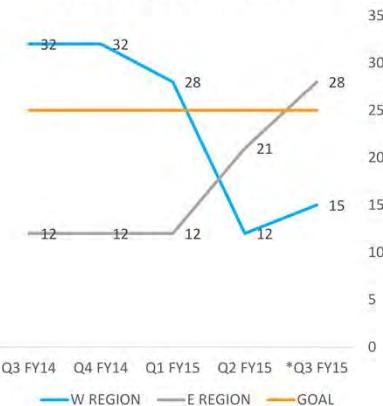
NOTES:

Which Graph Conveys Information Better?

PROFIT (\$M) BY QUARTER



PROFIT (\$M) BY QUARTER



There is no right answer, there is a good answer for the purposes. In this case, the level of detail desired is the key.

USE SPACE JUDICIOUSLY

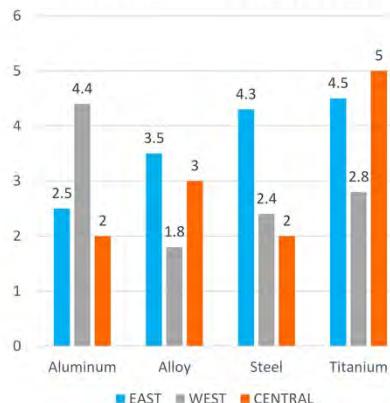
- **Reduce noise**
- **Continually ask, “How can I make it cleaner?”**
- **If it doesn’t help tell the story, it’s out**



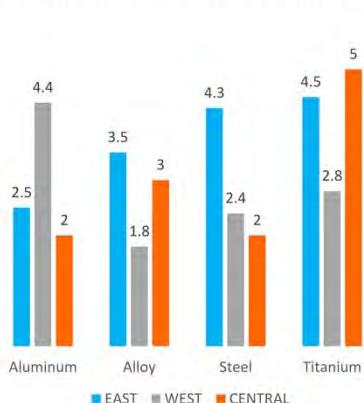
NOTES:

Zebras Don't Belong in Graphs

CORE METAL SALES BY REGION



CORE METAL SALES BY REGION



The message here is a little subtle, but horizontal lines to tell you the numbers when the numbers are on the vertical bars. Clean it up and the picture is more pleasing with less distraction.

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%



Tables are hard to use to deliver data effectively. However, it is clear, we hope, that the multi colored cells to serve as indicators is a visual onslaught. The bottom table shows how to clean up a table for better, more pleasing display.

FOCUS ON DESIGN

- **Color palette matters**
- **Go for a clean look**
- **Don't clutter with icons and FAQ**
- **Drill-down versus "Kitchen Sink" approach**
- **Not all screen locations are created equal**
- **Reduce or eliminate non-essential features**
- **Be consistent**



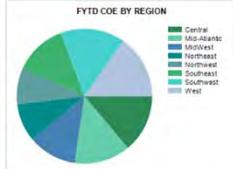
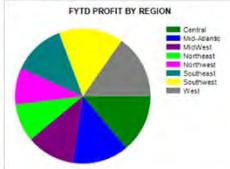
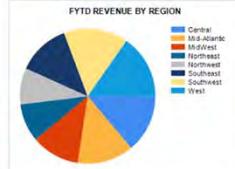
NOTES:

An Example of Poor Design

Acme Corporation

Today's Date: November 15, 2014

Welcome: John Henson



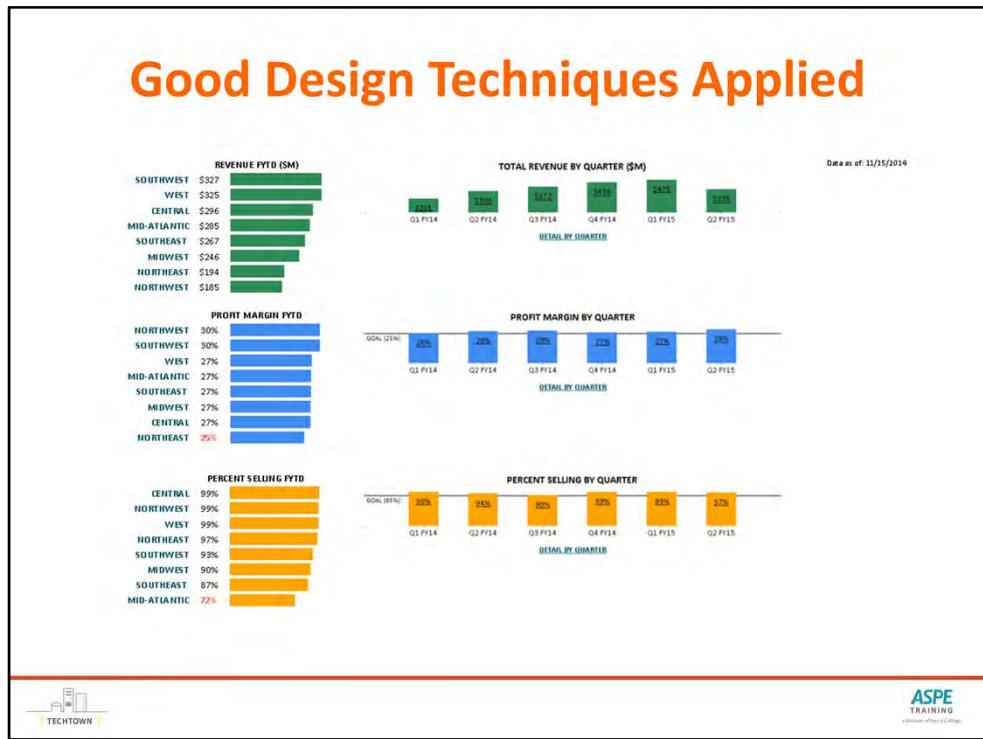
REGION	REVENUE	Q1 FY14			Q1 FY15			Q2 FY14			Q2 FY15		
		PROFIT %	% SELLING		REVENUE	PROFIT %	% SELLING	REVENUE	PROFIT %	% SELLING	REVENUE	PROFIT %	% SELLING
Central	\$39.00	27.9%	25.7%	\$30.00	20.0%	19.0%	\$38.00	20.0%	19.0%	\$12.00	35.8%	36.0%	
Mid-Atlantic	\$54.00	20.0%	19.4%	\$25.00	25.8%	22.4%	\$32.00	20.0%	19.0%	\$39.00	36.0%	36.0%	
MidWest	\$3.00	18.8%	40.4%	\$76.00	26.0%	29.0%	\$1.00	20.0%	(19.7%)	\$96.00	26.0%	49.3%	
Northeast	\$19.00	25.0%	36.0%	\$47.00	20.0%	18.0%	\$45.00	17.8%	12.8%	\$3.00	36.0%	37.1%	
Northwest	\$9.00	20.0%	36.0%	\$50.00	22.0%	33.2%	\$28.00	14.5%	9.7%	\$18.00	26.0%	5.4%	
Southwest	\$2.00	45.0%	38.8%	\$87.00	20.0%	19.0%	\$71.00	20.0%	16.4%	\$44.00	33.0%	36.3%	
West	\$64.00	21.8%	36.0%	\$80.00	26.0%	17.7%	\$49.00	26.0%	12.7%	\$44.00	43.0%	36.0%	



ASPE
TRAINING

Everything here is meant to offend the eyes, it seems.

NOTES:



This is complicated but clean.

NOTES:

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**



NOTES:

The End of the Show

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



NOTES:
