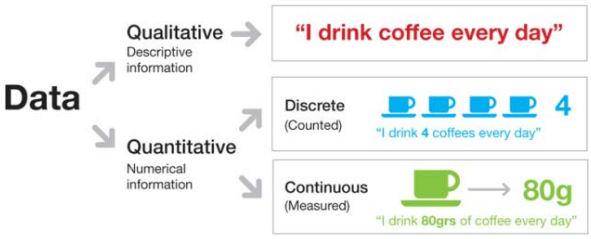
Data Analysis and Decision Making (DADM)



Lecture 1

Topic: Introduction Business Analytics

-Evidence-based management

-Sampling and data collection

-Big data

Masters Program- Learning Goals

Students will:

- have discipline specific knowledge and understanding;
- have the required cognitive, technical and research skills for lifelong learning;
- be effective communicators;
- be effective team members;
- be reflective learners; and
- demonstrate an awareness of ethical, cultural, social and sustainability issues.

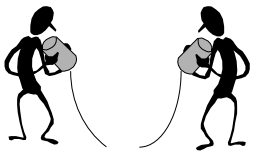
DADM Unit Learning Outcomes

- identify fit-for-purpose statistical approaches to address management concerns;
- explain data quality issues and collection methods;
- use a range of appropriate statistical tools and techniques;
- interpret statistical findings in a clear, concise and non-technical way;
- translate statistical findings into managerial actions; and
- communicate decision analysis results to stakeholders.

Lecture 1: Learning Objectives (LOs)

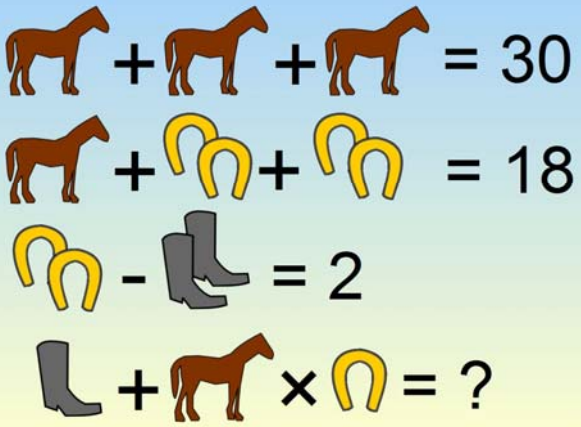
-Describe the importance of statistics.
-Differentiate between descriptive statistics and inferential statistics.
-Explain the need for sampling and discuss various data types.
-Describe variables and various types of measurement scales.
-Understand the research process.
-Discuss the difference between quantitative and qualitative research.
-Understand where business analytics fits with the ‘Big Data’ movement.

Getting to Know Each Other

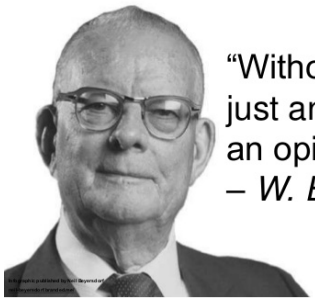


- How do you collect and use data? At work? In your daily life?
- On a scale of 1-10, where 1= very uncomfortable, 10 = very comfortable, rate your level of comfort with:
  - Using analytics?
  - Using MS Excel?
  - Working in teams?

Can You Solve This?



Article: Evidence-Based Management



"Without data you're just another person with an opinion."
– *W. Edwards Deming*

What Is Statistics?

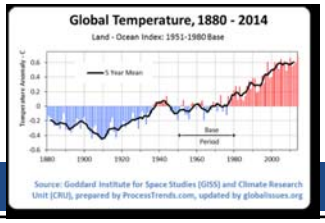
- Statistics is the methodology of extracting useful information from a data set.
- To do good statistical analysis, you must
  - Find the right data.
  - Use the appropriate statistical tools.
  - Clearly communicate the numerical information into written language.

The Relevance of Statistics

Example A

- Headline of newspaper states “What global warming?” after record amounts of snow in 2010.

✓ **Problem with Conclusion:** Incorrect to draw conclusion based on one data point. We must look at long-term trends.



The Relevance of Statistics

Example B

- A gambler predicts that he/she will roll a 7 on his/her next roll of the dice since he/she was unsuccessful in the last three rolls.

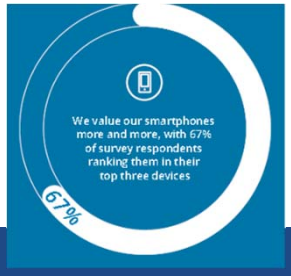
✓ **Problem with Conclusion.** The probability of rolling a 7 stays constant with each roll of the dice.

The Relevance of Statistics

Example C

- A magazine survey asks its readers for an opinion, then states that X% of Australians agree or disagree with the statement.

✓ **Problem with Conclusion.** The prediction was based on the opinions of readers from that magazine who decided to respond and does not necessarily reflect all Australians.



The Relevance of Statistics

Example D

- The CFO of Starbucks Corp. claims that business is picking up since sales at stores open at least a year climbed 4% in the quarter ended December 27, 2009.

✓ **Problem with Conclusion.** The CFO overstated the company's financial position by failing to mention that Starbucks closed more than 800 stores over the past few years.

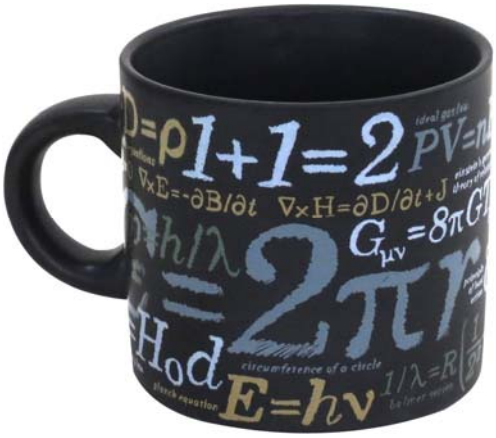
The Relevance of Statistics

Example E

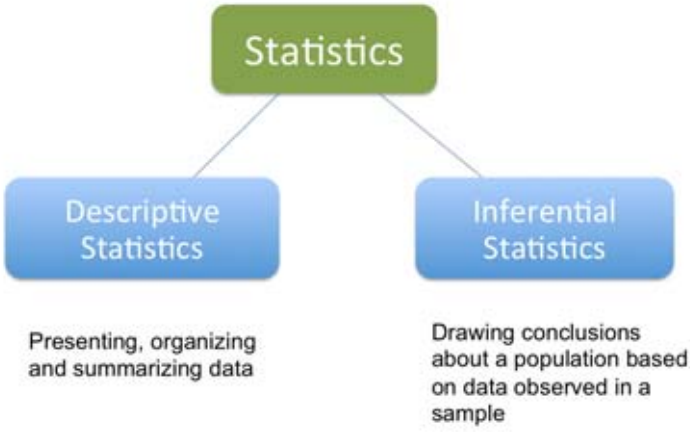
- Researchers showed that infants who sleep with a nightlight are much more likely to develop myopia.

✓ **Problem with Conclusion.** This is an example of the correlation-to-causation fallacy. Even if two variables are highly correlated, one does not necessarily cause the other.

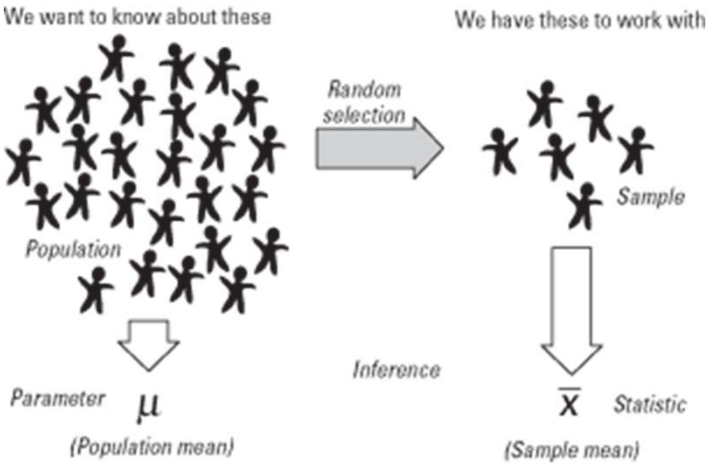
Coffee Break



Branches of Statistics

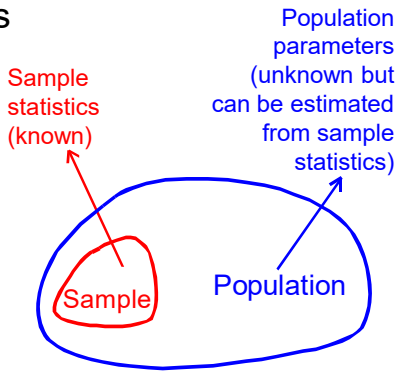


Inferential Procedures



Generalising from Sample Data

- Organisational research is mostly based on sample data
- A population consists of all people with which the study is intended to generalise.
- A sample is a subset of the population.
- The sample needs to be representative of the population for the sample results to be generalisable.
- The sample is representative when random sampling is used.



Class Discussion- So why sample? Why conduct a sample and not a census?

Extrapolating from Small Samples

- There are more extreme values in small samples. As the sample size increases, you are more likely to get a sample mean closer to the population mean. This is known as Central Limit Theorem.
- Often people wrongly assume that small samples may represent underlying characteristics of the population but fail to take into account the variability (ie chance) in the data. They readily use the sample's characteristics (such as the mean) to infer the population's characteristics.

Sampling

- **Population**—consists of all items of interest in a statistical problem.
  - **Population Parameter** is unknown.
- **Sample**—a subset of the population.
  - **Sample Statistic** is calculated from sample and used to make inferences about the population.
- **Bias**—the tendency of a sample statistic to systematically over- or underestimate a population parameter.

Sampling Methods

- Sampling Methods
  - ✓ **Simple random sampling**
  - ✓ **Cluster sampling**
  - ✓ **Stratified sampling**

Stats Joke

The Physicist, the Chemist, and the Statistician

Three professors (a physicist, a chemist, and a statistician) are called in to see their dean. Just as they arrive the dean is called out of his office, leaving the three professors there. The professors see that there is a fire in the wastebasket.

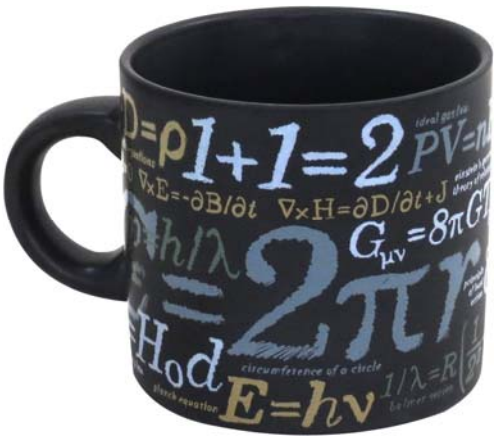
The physicist says, "I know what to do! We must cool down the materials until their temperature is lower than the ignition temperature and then the fire will go out."

The chemist says, "No! No! I know what to do! We must cut off the supply of oxygen so that the fire will go out due to lack of one of the reactants."

While the physicist and chemist debate what course to take, they both are alarmed to see the statistician running around the room starting other fires. They both scream, "What are you doing?"

To which the statistician replies, "Trying to get an adequate sample size."

Coffee Break

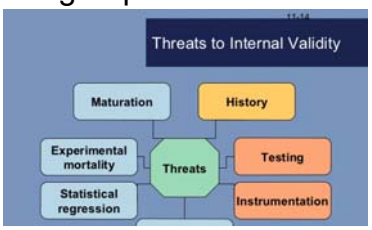


Data Collection Issues

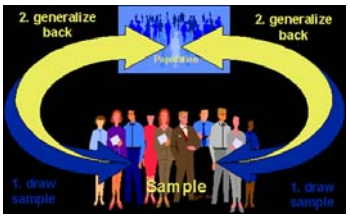
- Data Accuracy
- Interviewer Bias
- Nonresponsive Bias
- Selection Bias
- Observer Bias
- Measurement Error
- Internal Validity
- External Validity

Internal Validity

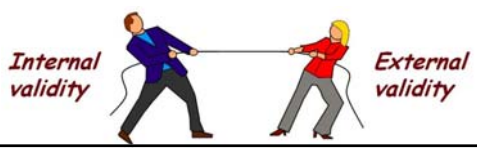
- Often relates to an experiment and how well extraneous variables were controlled
- Ways to address issues of internal validity
  - Randomly assign subjects to control and experimental groups, or
  - match subjects in the control and experimental groups in terms of age, gender, race, health etc
  - Randomly assign the treatment to groups
  - Use blinding where possible



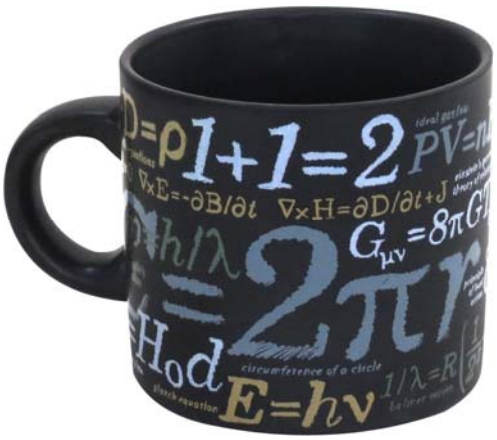
External Validity



- External validity relates to the generalisability of the study
- Can the results be replicated in another part of the company? With another sample? Time?
- If sampling is done well, and the sample is truly representative of the population, then the results should generalise.. that is, if all respondents respond and answer correctly….



Coffee Break



What is BIG Data?

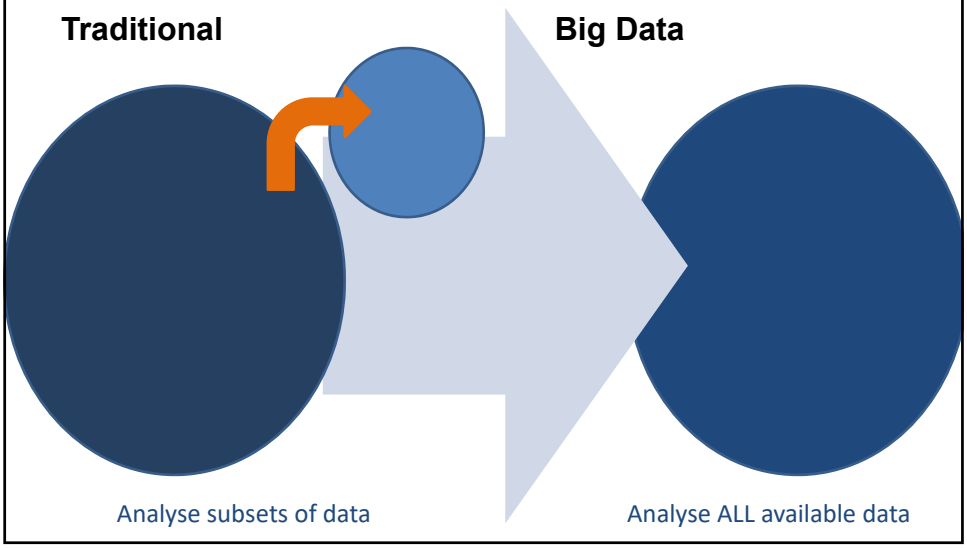
- No definitive definition
- Relates to large data sets but mostly big data is....

..data that is difficult to analyse using traditional data analysis techniques

Class Q- any concerns with the use of 'big' in terms of big data?

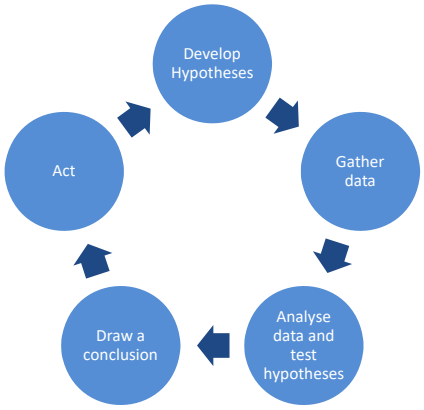
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Big Data Changes to Research Methods- Traditional vs Big Data Approach Sampling vs Scraping

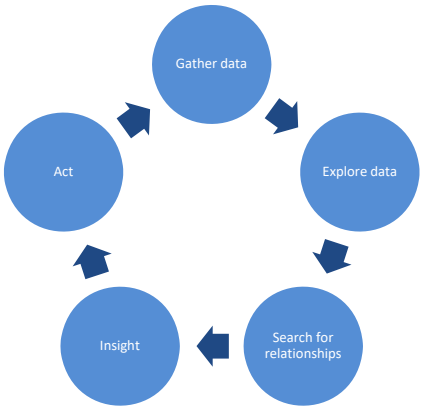


Big Data Changes Research Methods- Traditional vs Big Data Approach to Testing

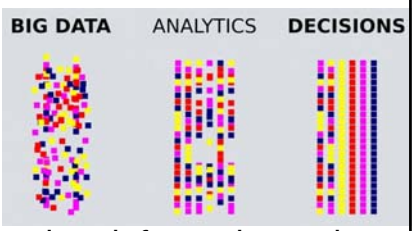
Traditional



Big Data



Why Are We Interested in BIG DATA?



- Data = Knowledge
- Big Data = Big Knowledge?
- It has value only if we convert data into information and use it for decision making
- Unstructured data constitute 95% of Big Data, therefore new approaches (mixed-methodologies) are needed
- It's everywhere
- "Social physics" (Pentland, 2014)

The use of large-scale data to predict human behaviour is gaining currency in business and government policy practice, as well as where the physical and social sciences converge.

Challenges

Technological (different needs and collaboration)

- The key components of machine learning - expertise, computing power, data, and algorithms – are not concentrated in any one domain, and industry, academia, and government, all play significant roles

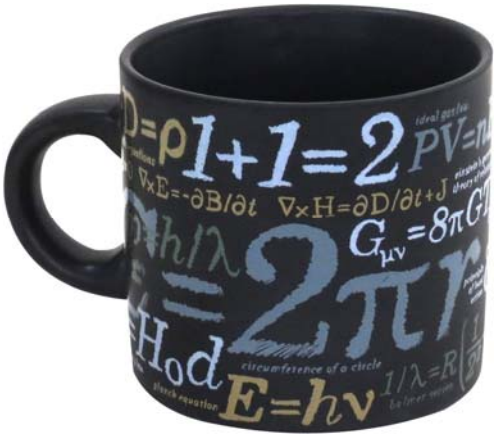
Educational

- Data literacy (**EVERYONE SHOULD CARE ABOUT QUANT SCIENCES!**)
- Ethical issues

Societal (Fairness, Privacy, Consent, and Cybersecurity)

- Trust, Transparency, and Interpretability
- Engagement of stakeholders
- Open distribution of tools and knowledge
- Living alongside machine learning

Coffee Break

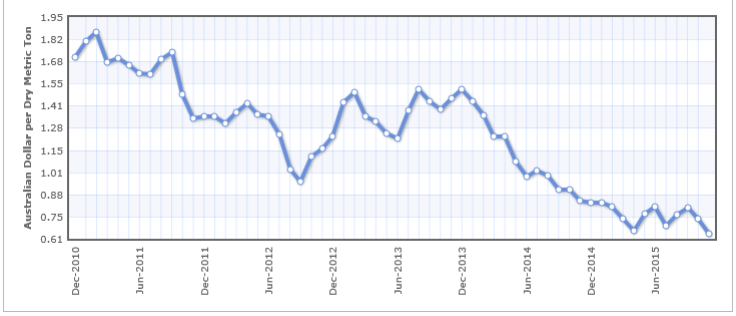


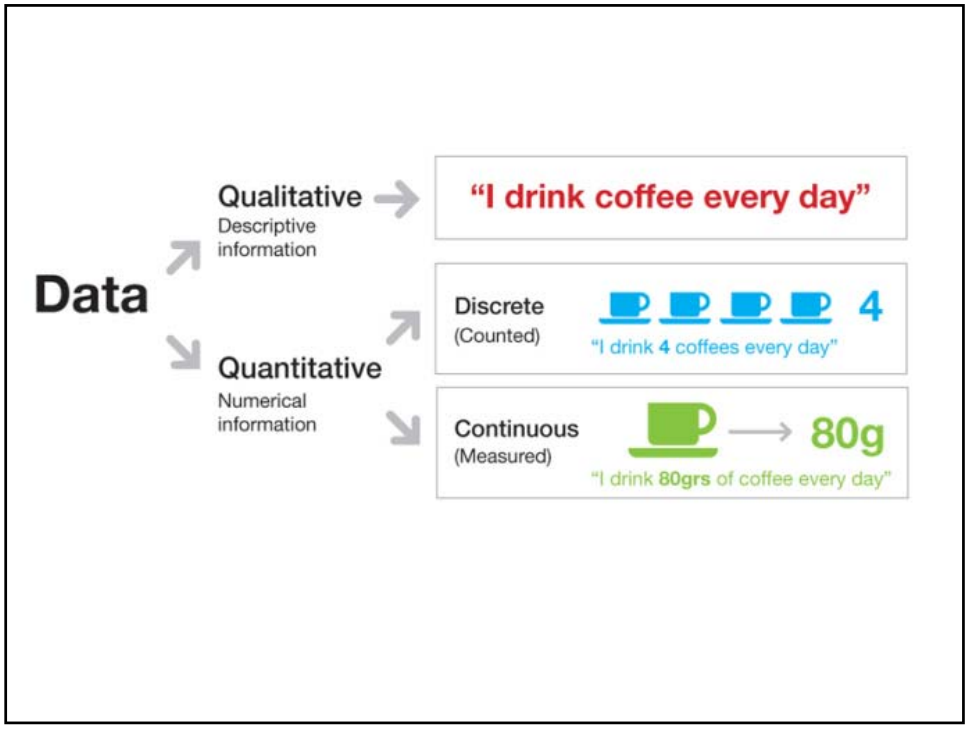
Types of Data

- Cross-sectional data
  - Data collected by recording a characteristic of many subjects at the same point in time, or without regard to differences in time.
  - Subjects might include individuals, households, firms, industries, regions, and countries.

Types of Data

- Time series data
  - Data collected by recording a characteristic of a subject over several time periods.
  - Data can include daily, weekly, monthly, quarterly, or annual observations.
  - This graph plots the Iron Ore price from Dec 2010 to Nov 2015 - it is an example of time series data.





Variables and Scales of Measurement

- Scales of Measure

- Nominal
- Ordinal

→ Qualitative Variables

- Interval
- Ratio

→ Quantitative Variables

Variables and Scales of Measurement

- The Nominal Scale
  - The least sophisticated level of measurement.
  - Data are simply categories for grouping the data.
- The Ordinal Scale
  - Ordinal data may be categorized *and* ranked with respect to some characteristic or trait.
    - For example, instructors are often evaluated on an ordinal scale (excellent, good, fair, poor).
  - Differences between categories are meaningless because the actual numbers used may be arbitrary.
    - There is no objective way to interpret the difference between instructor quality.

Variables and Scales of Measurement

- The Interval Scale
  - Data may be categorized *and* ranked with respect to some characteristic or trait.
  - Differences between interval values are equal and meaningful. Thus the arithmetic operations of addition and subtraction are meaningful.
  - No “absolute 0” or starting point defined. Meaningful ratios may not be obtained.

Variables and Scales of Measurement

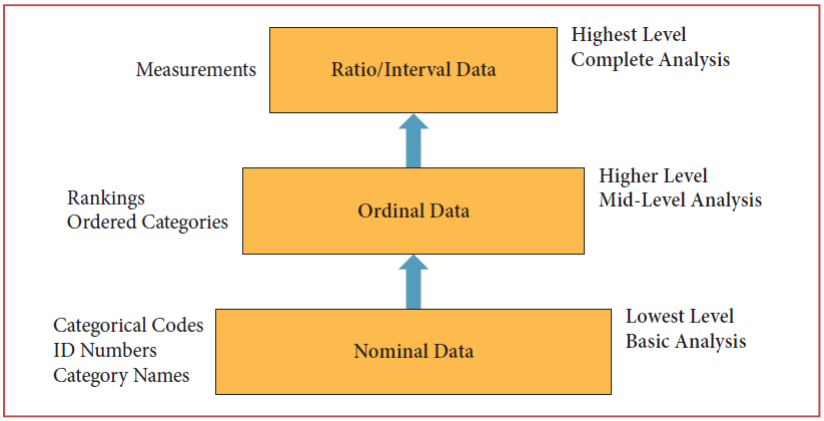
- The Interval Scale
  - For example, consider the Fahrenheit scale of temperature.
  - This scale is interval because the data are ranked and differences (+ or -) may be obtained.
  - But there is no "absolute 0" (What does $0°\text{F}$ mean?)

What does $\frac{80°\text{F}}{40°\text{F}}$ mean?

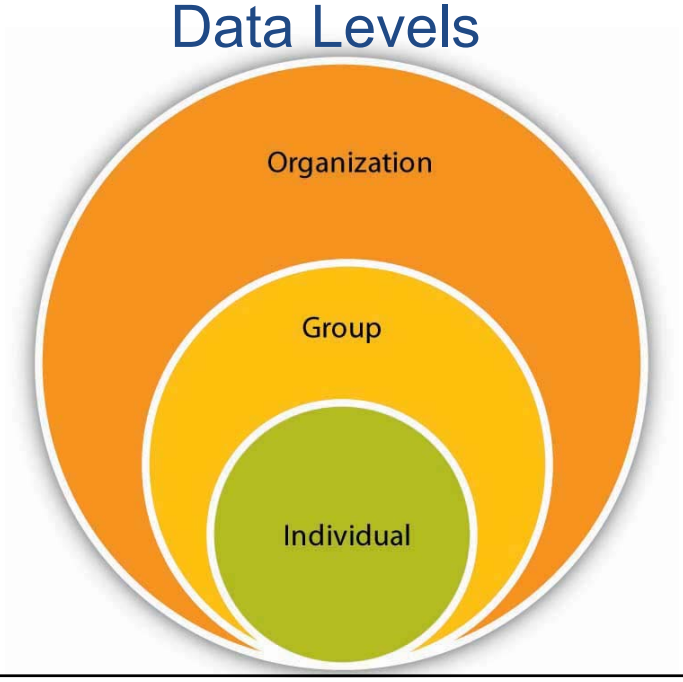
Variables and Scales of Measurement

- The Ratio Scale
  - The strongest level of measurement.
  - Ratio data may be categorized *and* ranked with respect to some characteristic or trait.
  - Differences between interval values are equal and meaningful.
  - There *is* an "absolute 0" or defined starting point. "0" *does* mean "the absence of …" Thus, meaningful ratios may be obtained.
  - The following variables are measured on a ratio scale:
    - General Examples: Weight, Time, and Distance
    - Business Examples: Sales, Profits, and Inventory Levels

Data Measurement Levels



Data Levels



Summary

-Describe the importance of statistics.
-Differentiate between descriptive statistics and inferential statistics.
-Explain the need for sampling and discuss various data types.
-Describe variables and various types of measurement scales.
-Understand the research process.
-Discuss the difference between quantitative and qualitative research.
-Understand where business analytics fits with the 'Big Data' movement.

Next Lecture….

- Brief revision of Lecture 1
- Topic:
  - Presenting and Reporting Data