Chapter 1
Introduction to Data Analytics

Section 1.1
Information for Decisions

David W. Gerbing

The School of Business
Portland State University

- Information for Decisions
  - Why Data Analytics?
  - Variables and Data
  - Organize Data for Analysis
  - Appendix: Levels of Measurement

1.1a
Why Data Analytics?

Managers as Decision Makers

Understanding of reality needed for effective decisions

- **Managers**: The decision makers who formulate the strategic goals of an organization and the tactics to realize those goals
- Effective decision making to accomplish these goals requires an accurate understanding of reality, knowledge of the . . .
  - Internal business environment and processes
  - External environment, such as customers, markets, suppliers
- To understand reality, acquire information and ultimately knowledge for topics such as
  - Marketing: Consumer preferences, market segmentation
  - Supply Chain: Ship time of raw materials, tolerances
  - Finance: Financial modeling, analysis of returns
  - Accounting: Account auditing, time to process invoices
  - HR: Job satisfaction, accumulated employee sick time

Statistics as a Tool to Understand Reality

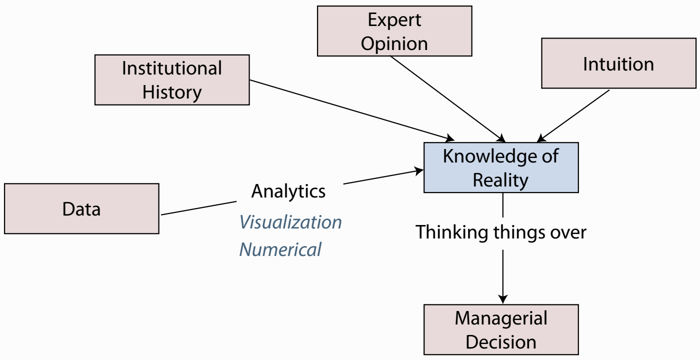
Many sources of information to guide decisions

- A central concern for the manager is gain the needed knowledge of reality for which to make appropriate decisions
- The manager can further understanding from many sources of information
- Data provides a primary source of information
- **Data**: Varying measurements of people, organizations, places, things and events
- Statistical methods provide the tools to process data
- **Data Analytics**: Apply statistical methods to transform data into usable information from which to make decisions
- Data analytics results in . . .
  - graphics, data visualization
  - numerical analysis in the form of text output

Information Sources for Decision Making

Integrate information from many sources

- The application of statistics to data analytics can provide useful information, but does not provide the decision rules to make a decision, which is based on many considerations



The Ubiquity of Data in the Modern World

Every manager benefits from access to data

- Given modern computing power and storage, massive amounts of data are now routinely collected, stored and analyzed
- Data are available from many sources
  - Virtually all business transactions are now computerized including every sale, every payment, every inventory entry, every web page interaction
  - Surveys and questionnaires from customers and potential customers, as well as employees, generate data
  - Many Internet sources are available, such as government websites regarding the census, the national and regional economies, and specific industries such as health care
- **Key Concept**: The resulting reams of data are useful only after their analysis with statistical methods

Interpretation of Data is Fundamental to Management

Data is an increasingly important source of information

- Some examples of using data to support decision making
  - To plan the budget for next quarter's payroll, the manager needs to know the typical total overtime hours
  - To plan for staffing in the emergency room, the manager needs to know the patient inter-arrival times
  - To decide which truck to purchase for the fleet, the manager needs to know the fuel costs for the different types and brands of trucks
- A key characteristic of the data is variation from one data value to the next
  - Total hours of employee overtime each month
  - Inter-arrival times at the emergency room
  - Cost of filling up a delivery truck's fuel tank

Variation is Everywhere

The presence of variation is the reason for statistical analysis

- The data values not only vary, they vary randomly
- **Random variability**: The value of the next outcome of some event is not known until it occurs
  - How many overtime hours next month will occur?
  - When does the next patient arrive at the emergency room?
  - How much will the diesel fuel cost for the next fill up?
- The presence of this random variability, the random fluctuation of data values, obscures knowledge of reality
- Virtually everything of interest that is sampled over time exhibits random variation
- **Key Concept**: The ability to understand and interpret the statistical analysis of data provides an enhanced understanding of random variability and its consequences for decision making

Machine Learning: The Forecast

Decisions are made in the context of random variation

- **Descriptive Analytics**: Analysis procedures to describe a known past
- A consequence of variability is that the outcome of any event, such as number of overtime hours, next patient arrival or fuel cost, is not known in advance
- Managers make decisions regarding future events and outcomes
- **Forecast**: A prediction of the outcome of a currently unknown future event, here based on the statistical analysis of data
- **Predictive Analytics**: Used machine learning to forecast the uncertain future
- **Key Concept**: Managers forecast the future in the context of the inherent random variability of the relevant outcomes
- What is the basis of this forecast in terms of data analytics?

Knowledge that Underlies Random Variation

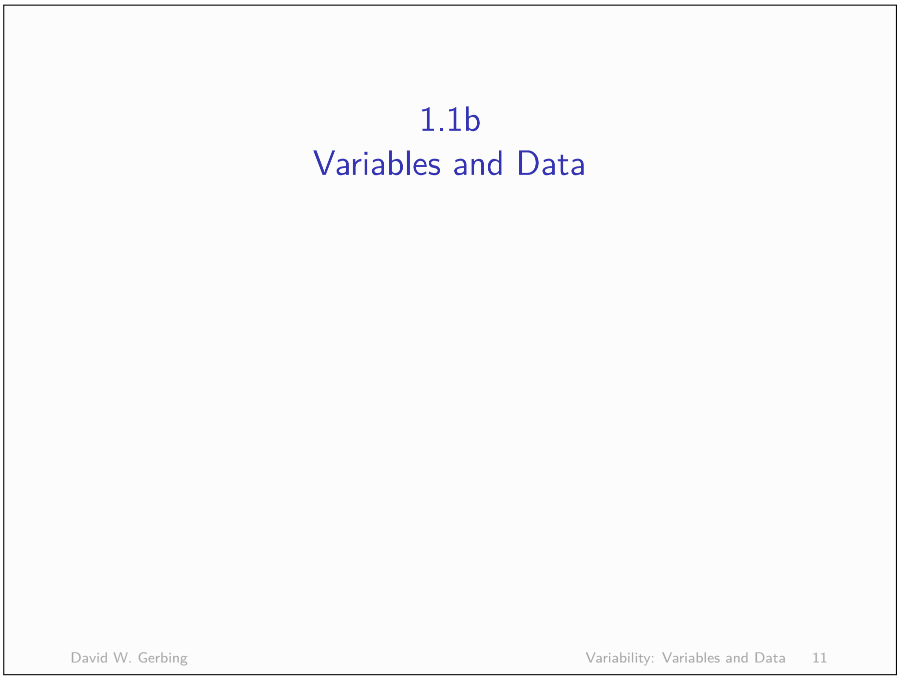
Decisions are made in the context of random variation

- **Statistical Thinking**: Focus on understanding the presence of random variation and attempts to understand its causes and consequences in terms of risk and opportunity
- More effective decisions generally result from analysis of the data, which provides an understanding of reality in the form of forecasts for the uncertain future outcomes
- Statistical analysis of data collected from *past* events provides the needed knowledge for the forecast of similar *future* events
- Ex: To reduce wait time in the emergency room, need to know, that is, forecast, the underlying pattern of when patients show up and the severity of their conditions
- **Key Concept**: Each data value results from two influences
  - Chance, random fluctuations that create variation
  - An underlying stable component common to all data values

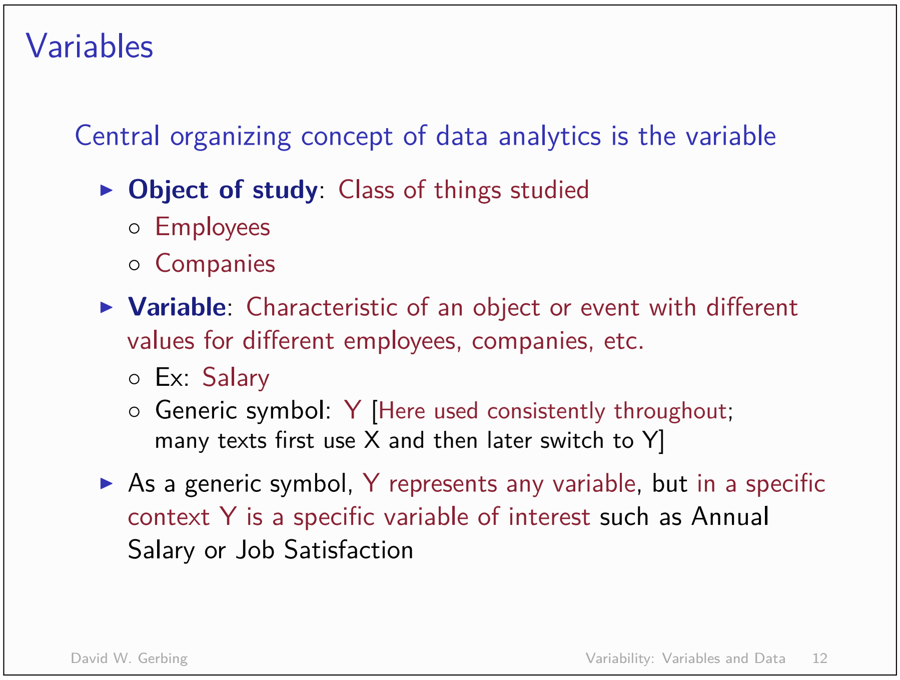
Analyze Data from the Past to Manage the Future

Describe the past, infer a stable pattern, project into the future

- The unpredictable random variation obscures the underlying structure, the stability that projects into the future
- **Key Concept**: data analytics is the quest for stability, true knowledge, in the presence of obscuring random variation
- **Model**: An expression of each data value in terms of the underlying structure and the random component
- **Key Concept**: data analytics follows four sequential steps
  1. Describe: Assess inherent variation in the data
  2. Infer: Build a model that expresses the knowledge of the underlying structure that underlies this variation
  3. Forecast: From the model project this stable component into the future as the estimate of future reality
  4. Evaluate: Wait for some time to pass and then compare the forecast to what actually happened



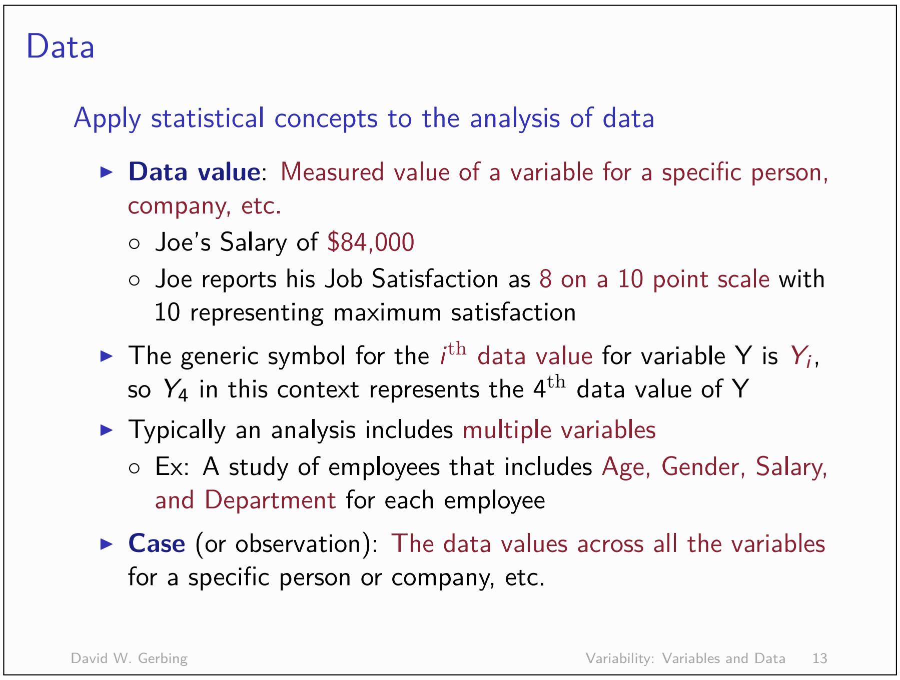
1.1b
Variables and Data



Variables

Central organizing concept of data analytics is the variable

- **Object of study**: Class of things studied
  - Employees
  - Companies
- **Variable**: Characteristic of an object or event with different values for different employees, companies, etc.
  - Ex: Salary
  - Generic symbol: Y [Here used consistently throughout; many texts first use X and then later switch to Y]
- As a generic symbol, Y represents any variable, but in a specific context Y is a specific variable of interest such as Annual Salary or Job Satisfaction



Data

Apply statistical concepts to the analysis of data

- **Data value**: Measured value of a variable for a specific person, company, etc.
  - Joe's Salary of $84,000
  - Joe reports his Job Satisfaction as 8 on a 10 point scale with 10 representing maximum satisfaction
- The generic symbol for the i^{th} data value for variable Y is Y_i, so Y_4 in this context represents the 4^{th} data value of Y
- Typically an analysis includes multiple variables
  - Ex: A study of employees that includes Age, Gender, Salary, and Department for each employee
- **Case** (or observation): The data values across all the variables for a specific person or company, etc.

More Examples of Variables and Data Values

Object of Study	Variables	Values
Employee	Annual Salary	$58,500
	Age	46
	Gender	Female
Company	Size of Facility	327,470 sq ft
	Employees	185
	Annual Sales	$5.1 million
Hospital	Annual Utility Bill	$136,530
	Number of Beds	249
	Location	Seattle
Country	Continent	Europe
	Number of Doctors	85,314
	Average Age of Death	69

Categorical Variables

Variables with values that define a small number of categories

- Some variables have values that consist of unordered groups of the people, companies or whatever is the unit of analysis
- **Categorical variable** or factor: The values classify an object or event into one of a relatively small set of distinct categories or groups of similar objects or events

Categorical Variable	Values
Gender	Male, Female
State of Residence	OR, WA
Payment Type	Cash, Check, Credit Card
Severity of Injury	Mild, Moderate, Severe
Jersey Number	44, 15, 38

- **Level**: A category or value of a categorical variable

Categorical Data Values

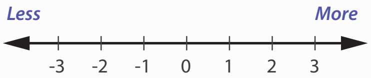
Classification instead of measurement

- Data values of a categorical variable consist of a relatively small number of unordered categories
  - Gender with values of Female and Male
  - Payment Type with values of Cash, Check or Credit Card
- More formally, for a categorical variable the data values are classified, not measured
- **Classification**: Assign a property of a person or other instance of the unit of analysis to a level of a categorical variable
- The categories, or levels, can be labeled with numbers, but such values are merely labels that do not have numeric properties
  - Payment Type with each of the three types labeled 1, 2 or 3
  - Jersey Number such as 44, 15, 38

Continuous Variables

Variables with numeric values that fall on a continuum

- **Continuous variable**: The values are ordered along the abstraction of the real number line in which an unlimited number of numeric values lie between any two values

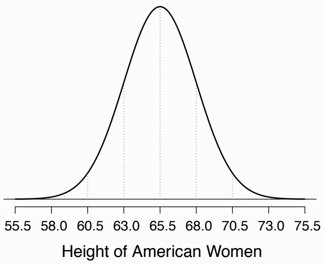


  - ○ Question: What number immediately follows the number 2?
  - ○ Answer: In the abstract world of mathematics, there is no such number, not 2.0001 or even 2.00000001
- Continuous variables include money, time, weight, length, volume, distance and the extent of agreement (such as regarding a statement of opinion)

Continuous Variables

Variables based on the real number line

- Much theoretical development of statistics applies to the theoretical abstraction of continuous variables
- Ex: The smooth normal (bell) curve shows how frequently different values occur, such as for Height



- Mathematical abstractions such as the normal curve are never directly observed, but are instead approximated with data
- **Key Concept**: Distinguish between the true value of a continuous variable and its corresponding approximation with a measurement, a data value at a specified level of precision

Numerical Data Values

Data are expressed in discrete units

- A measurement, a numeric data value, is *not* the true value
- **Key Concept**: A measurement only approximates the true value of a continuous variable with a category of similar values, usually one of a large number of similar categories
  - ○ Ex: . . . $2.00, $2.01, $2.02, $2.03 . . .
  - ○ There is a next data value, $2.01 follows $2.00
- The measurement of each true value should be placed in its closest category such as
  - ○ 2.15248. . . would be measured as 2.15
  - ○ 2.15957. . . would be measured as 2.16
- The extent of the approximation is based on the accuracy of the measurement procedure and its precision, such as 2.1 lb or the more precise 2.13 lb, neither of which is the true value

Time Orientation of Data: Cross-Sectional

Data collected at the same or different times?

- Data analytics is the study of variability, which can occur in two different forms
  - At about the same time, with different instances of the unit of analysis
  - At different times with the same unit of analysis
- **Cross-sectional data**: Data collected at about the same time with variation over different instances of the unit of analysis such as employees or companies
  - Mail a marketing survey to consumers within same week
  - Observe traffic flow at 3pm on Monday afternoon
  - Measure job satisfaction of all employees in a department next Thursday

Time Orientation of Data: Longitudinal

Variability can occur over time

- **Process**: Repeatable sequence of events that occur over time to accomplish a specific goal
  - Answering phone
  - Scheduling personnel
- **Longitudinal data**: Data collected over different times from the same process
  - Time each patient is admitted throughout a day
  - Quarterly revenues over 10 years
- Longitudinal data can be classified into two different types
  - **Event driven**: A measurement taken at each time an event occurs, such as the length of the next machined part
  - **Time series**: Measurements taken at regularly spaced time intervals, such as quarterly revenue

1.1c
Organize Data for Analysis

Variables and Variable Names

Organize the data for multiple variables

- A data analytics project generally involves several variables
- Some of the variables in the same analysis may have categorical data values and others may have numerical data values
- There needs to be a structure that defines how the data are organized and then stored in a computer file
- The basic structure of the data is a table, which begins with the name of each variable included in the data file
- **Variable name**: A concise name, usually less than 10 characters, that identifies a variable in a computer analysis
- The data values for a variable are listed in the same column, with the variable's name usually at the top of the column
- **Key Concept**: With computer software for data analytics, indicate the analysis of a specific variable by its name

Choose the Variables for the Analysis

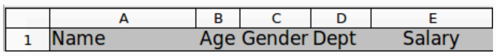
An example

- Following this format, the data are stored in a table, such as represented in a worksheet
- To illustrate, suppose an HR manager plans to analyze Salary as it relates to three other variables: an employee's Age, Gender and Department within which the employee works
- It is convenient to keep the variable names short, so in this example refer to Department with the abbreviation Dept
- For each employee, record his or her Name plus the value of each of the four variables in the analysis:
  Age, Gender, Dept and Salary
- To prepare the data values of these variables for analysis, enter the corresponding data values for each employee into a table stored within a worksheet such as MS Excel or the free LibreOffice Calc

Structure of the Data Table

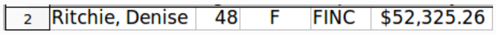
Enter the data values into a table

- To construct the worksheet table, first enter the variable names in the first row, beginning with the column that contains each person's unique identifier, his or her name



	A	B	C	D	E
1	Name	Age	Gender	Dept	Salary

- Now add the data for an employee
- For example, Denise Ritchie is 48 years old, is female, and works in the finance department with a salary of $52,325.26
- The second line of the worksheet is the first row of data values, here the data for Denise Ritchie



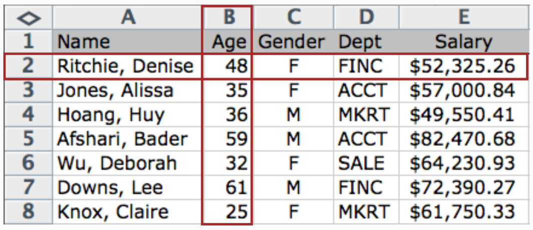
2	Ritchie, Denise	48	F	FINC	$52,325.26

- Then continue adding data to the data table row by row, such that each row contains the data values for one person

Data Analytics Proceeds from the Data Table

Data arranged in the form of a table, usually called `d` for data

- **Data table** or data frame or data matrix: A rectangular table of data with multiple observations across multiple variables
  - Each row contains all the data for a single case (observation)
  - Each column begins with a variable name and then its data



◇	A	B	C	D	E
1	Name	Age	Gender	Dept	Salary
2	Ritchie, Denise	48	F	FINC	$52,325.26
3	Jones, Alissa	35	F	ACCT	$57,000.84
4	Hoang, Huy	36	M	MKRT	$49,550.41
5	Afshari, Bader	59	M	ACCT	$82,470.68
6	Wu, Deborah	32	F	SALE	$64,230.93
7	Downs, Lee	61	M	FINC	$72,390.27
8	Knox, Claire	25	F	MKRT	$61,750.33

- Table of cross-sectional data with an ID field for employees in the 1st column, with 4 variables across 7 observations, $n = 7$

Properties of a Data Table

The data table structures the data by variable

- Data table identified on the computer system with a file name, the name of the computer file in which the data are stored
- Each data table also has a name in the application used for the analysis of the data
  - Excel: name of *worksheet*
  - R: name of *data frame*, usually `d`, or any valid name
- **Key Concept**: All data values *within* a column, for a single variable, are of the same type, categorical or numerical
- Regarding the previous example of a data table of employees
  - Age and Salary have numeric data values, each with a relatively large number of potential values
  - Gender and Dept have categorical data values or factors, each with a relatively small number of non-numeric values

Different Forms of the Data Table

Same data, different locations

- **Key Concept**: Distinguish between the following concepts as they relate to the data table
  - name of the data file on the computer system that contains the data table
  - name of the data table within the running statistical application such as R or Python
  - name of one or more specified variables within the data table
- Note the data table exists both within the data file stored on the user's computer system as well, when applicable, in a data analytics application, such as R or Python
- The data table can easily be moved to and from the computer file, Excel or other worksheet application, and the data analytics application, such as R or Python

Data Storage Formats

Excel

- Perhaps the best way to manually enter data into the computer and view data is with a worksheet application such as Excel
- Data analysis systems read data tables from many formats, including Excel, so can leave data in this format

csv text file

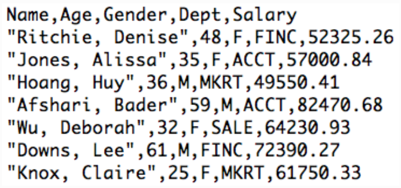
- **csv text file**: Plain text file with “comma separated values”
- Almost any application can read a csv file, and although the pure text contains no formatting, not needed for data files
- A related file format is a tab delimited text file, where invisible tab marks separate adjacent values

Structure of a .csv File

Can open a .csv text file in almost any application

- To illustrate, here is the exported .csv text file, example.csv,

  `http://lessRstats.com/data/example.csv`



```
Name,Age,Gender,Dept,Salary
"Ritchie, Denise",48,F,FINC,52325.26
"Jones, Alissa",35,F,ACCT,57000.84
"Hoang, Huy",36,M,MKRT,49550.41
"Afshari, Bader",59,M,ACCT,82470.68
"Wu, Deborah",32,F,SALE,64230.93
"Downs, Lee",61,M,FINC,72390.27
"Knox, Claire",25,F,MKRT,61750.33
```

File *example.csv*: Variable names in the first row, all remaining rows are data

- When exporting data from Excel, usually there is no need to directly view the resulting .csv text file
- Unless there is some problem reading the data, directly open the exported data file into the application for data analytics, such as R or Python

Appendix
Levels of Measurement

Measurement: Ratio Data

Different types of measurements provide different quality

- **Ratio data**: Differences of unit length between each pair of numbers represent the same distance, and a natural zero point
  - Distance between 3 and 4 is same as between 103 and 104
  - Natural, meaningful zero point implies that, for example, 6 is twice as much as 3
- Examples include measurements of
  - Dimensions such as weight, length
  - Monetary Units such as costs or revenues
- Generally preferred level of measurement as it provides the most information with data values that have properties generally expected of numbers
- **Key Concept**: Procedures such as computing the average, or the ratio between two values, can apply to ratio quality data

Measurement: Interval Data I

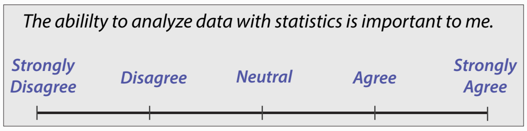
Other types of measurements provide less quality

- **Interval data**: Equal differences of magnitude for equal differences of the measurements, but no natural zero point
- For example, there are different scales for measuring temperature, of which two are Fahrenheit and Celsius, yet each scale has a different value of 0°
  - Fahrenheit and Celsius scales represent different temperatures with the same number, so 0°F $\neq$ 0°C
  - Yet the difference between 0°F and 1°F is the same temperature difference as between 100°F and 101°F
- Without a natural zero point, ratios are not meaningful
  - 100°F is *not* twice as hot as 50°F
  - 100°C is *not* twice as hot as 50°C

Measurement: Interval Data II

Other types of measurements provide less quality

- A common source of data typically assumed of interval quality are responses to items on a survey in which the respondent provides the level of agreement on a disagree/agree continuum


The abililty to analyze data with statistics is important to me.

Coding Scheme 1:	1	2	3	4	5
Coding Scheme 2:	-2	-1	0	1	2

- Here the continuum is divided into 5 values, of which the respondent chooses one by circling the desired response
- The analyst converts the response to a number, of which there are many possibilities as long as equal intervals are maintained

Measurement: Ordinal Data I

The least quality of measurement of a continuous variable

- **Ordinal data**: Data values are ordered but have *unequal* levels of magnitude between adjacent values
- The extent of the differences between adjacent values are not quantified, so they are not meaningful
  - Severity of Injury with levels of Mild, Moderate and Severe Injury where Severe represents more severity than Mild, but how *much more* is not specified
  - Difference between the 1st and 2nd ranked employees on assessment of Job Performance may be small or large
- The data may be coded with numbers, such as 1 for first and 2 for second, but the ordinal data values are not numerical
- **Key Concept**: Statistics such as the average cannot be meaningfully calculated with ordinal data

Measurement: Ordinal Data II

Categorical properties of ordinal data

- Usually ordinal data consists of a relatively small number of categories, such as Mild, Moderate and Severe Injury assessed for all incoming emergency room patients
- When measured on a scale of ordinal values, the resulting data values for the continuous variable, such as Severity of Injury, are non-numeric data values
- For clarity, the underlying distinction between the abstract underlying continuous variable and the variable defined by the resulting data values must be maintained
- **Key Concept**: For ordinal data, the underlying continuous variable is numeric, but the corresponding ordinal data are ordered categories

Abstract Variable vs. Data Values

The word “variable” can refer to two different concepts

- **Nominal data**: Data values that consist of a relatively small number of unordered categories
- Compare the underlying variable to the resulting data values
  - Measure a value of a continuous variable: Ratio, interval or ordinal data
  - Classify a value of a categorical variable: Nominal data
- Compare the properties of the data values
  - Ratio and interval data values are numeric
  - Ordinal and nominal data values are categorical, ordered and unordered, respectively
- **Key Concept**: In the context of data analytics, the term “variable” typically refers to the property of the resulting data values rather than the underlying variable being assessed

Index Subtract 2 from each listed value to get the Slide

- The End