# Business Statistics

## Lecture 1: Course Introduction & Descriptive Statistics

# Goals for this Lecture

- Introduce professor & course
- Define some basic statistics terminology
  - Populations vs. samples
  - Descriptive vs. inferential statistics
- Numerical descriptive statistics
  - Measures of location
  - Measures of dispersion
- Short introduction to JMP

# Contact Information

- Professor Ron Fricker
  - Phone: 831-869-8414
  - E-mail: rdfricker@nps.edu
  - Located in Monterey

Call or e-mail anytime!

# A Little Bit About Me...

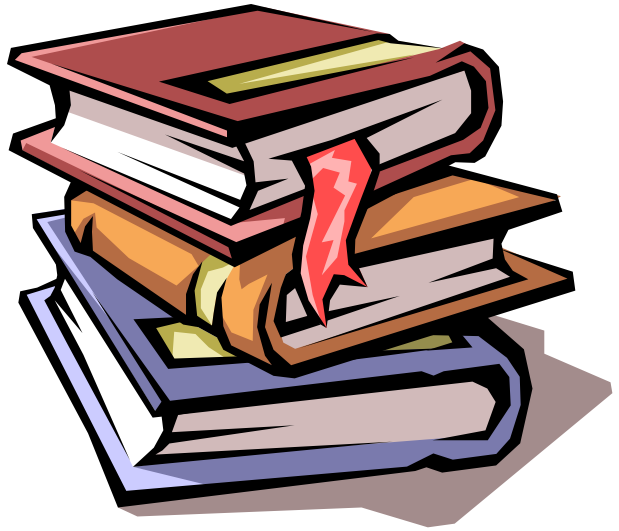
- Academic credentials
  - Ph.D. and M.A. in Statistics, Yale University
  - M.S. in Ops Research, The George Washington University
  - B.S. in Mathematics from the United States Naval Academy
- Teaching credentials
  - Started teaching post-graduate courses in mid-80s
  - Have taught at NPS, RAND Graduate School, and USC
- “Real world” credentials
  - Former active duty naval officer
  - Commercial managerial experience
    - Two defense-related organizations
    - One non-profit
- Can find out more at http://faculty.nps.edu/rdfricke/

# Course Goals

- Be able to:
  - *Apply* basic statistical methods to business problems
  - *Understand* more advanced statistical techniques and how they are properly applied
  - *Judge* good statistics and statistical practice from bad
  - *Know* when to call in statistical experts

# Course Outline

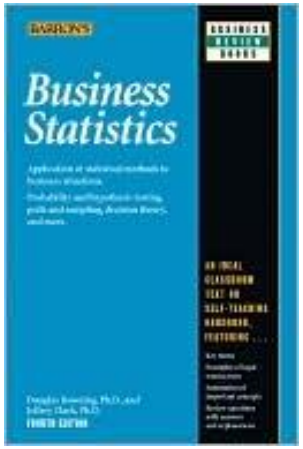
- Eleven lectures over nine class meetings:
  - Descriptive statistics
  - Basic probability
  - Confidence intervals
  - Hypothesis testing

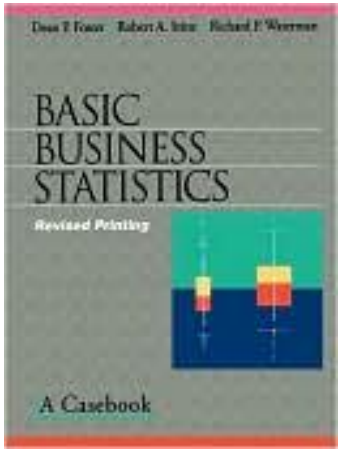


- See the course syllabus for class policies
- Course website: http://faculty.nps.edu/rdfricke/Business_Stats.htm

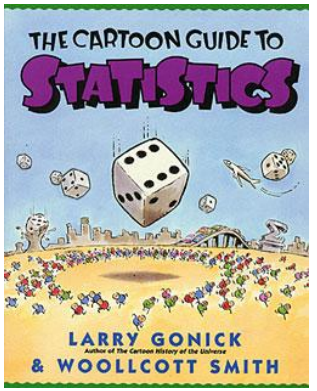
# Course Texts & Resources

- Course texts:




  - *Business Statistics* by Downing and Clark
  - *Basic Business Statistics: A Casebook* by Foster, Stine and Waterman
- If supplemental reading is required, recommend *Cartoon Guide to Statistics* by Gonick and Smith
  - It's a rigorous treatment of the material, but done in a very accessible style
- Course software: Excel & JMP

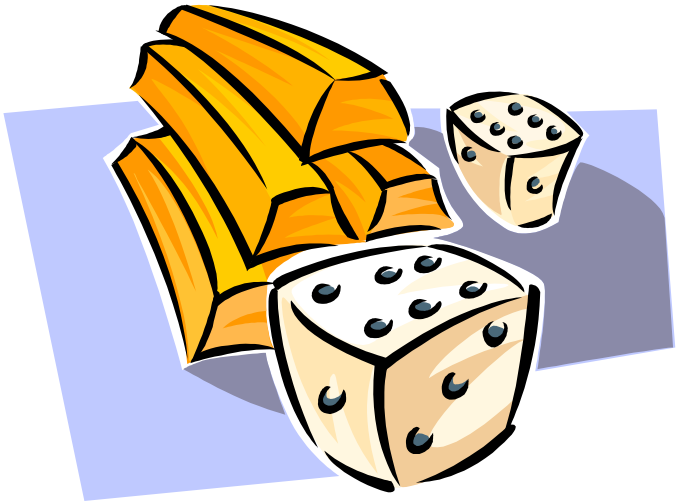


# Descriptive Statistics

- Numerical
  - Mean, median, mode
  - Variance standard deviation, range
- Graphical
  - Histograms
  - Boxplots
  - Scatter plots

# Probability

- Basic concepts
- Discrete distributions
- Continuous distributions
- Conditional probability

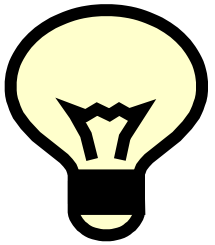


# Inferential Statistics

- Point Estimation
- Interval Estimation
  - E.g., confidence intervals
- Hypotheses testing
  - Testing sample means and variances

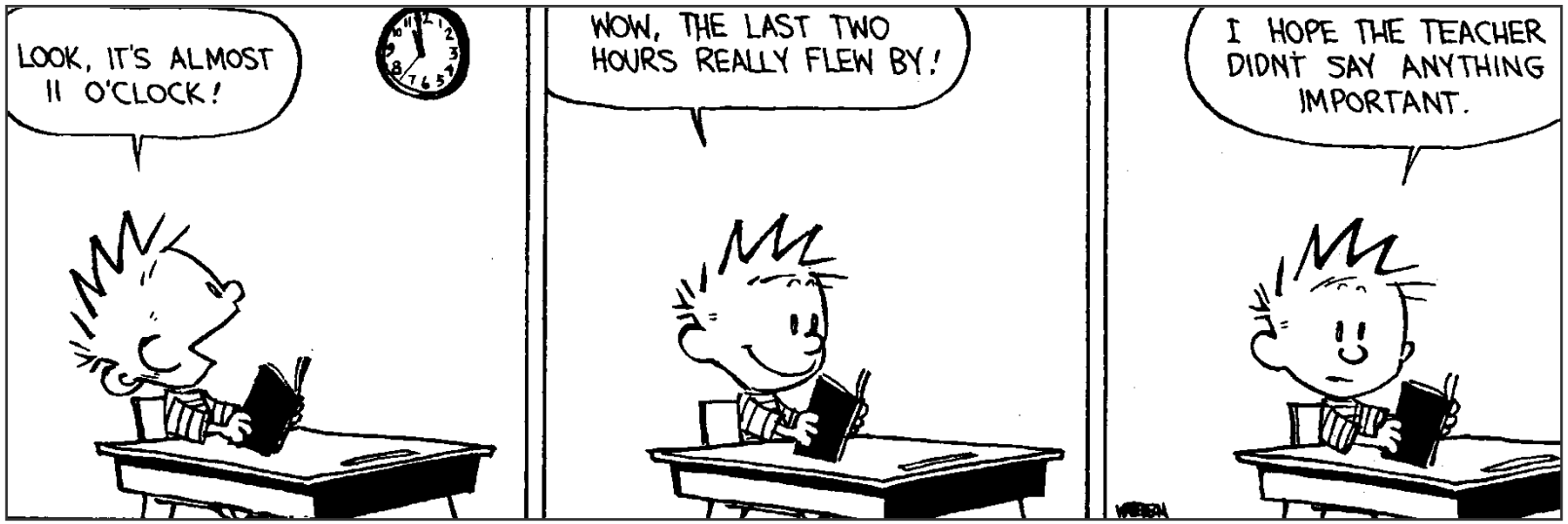
# How to Study Statistics

- Do the reading in multiple passes
  - First skim for major ideas before the lecture
  - After the lecture, go back for details
  - Re-read as necessary to solidify concepts
- Do practice problems (homework)!
  - Only after first completing reading assignment
  - If necessary, make up simple data to see what equations are doing
- *Don't just depend on your colleagues to explain the concepts to you…*



# How *Not* to Study for this Course

**Calvin & Hobbes** by Bill Watterson



# “Statistics”

- “Statistics” has two uses in English:
  - Can mean “a collection of numerical data”
  - Also refers to a branch of mathematics that deals with the analysis of statistical data
- This class is all about the latter
  - Though we must use “collections of numerical data” to do our analyses

# Why Study Statistics?

- The world is an uncertain place
  - Your company is recruiting a new CEO. What compensation should you offer?
  - What GMAT score do you need to get in to an MBA program?
- *Statistics gives you the tools to make informed decisions in uncertain conditions*

# Statistics Uses Data

- Statistics attacks uncertainty with data
  - CEO: Salaries of other CEO's
  - GMAT: Other students' scores
- *Statistics turns raw data into information that speaks to your question*

# Variability

- Statistics is *more* than tabulating numbers
- Data exhibit variability
  - CEO's have different backgrounds, work in different industries, etc.
  - Students vary in ability and luck
- Standard statistics question: *"Given the data I have seen, what is the truth likely to be?"*

Understanding and describing variability is one of the main jobs of statistics

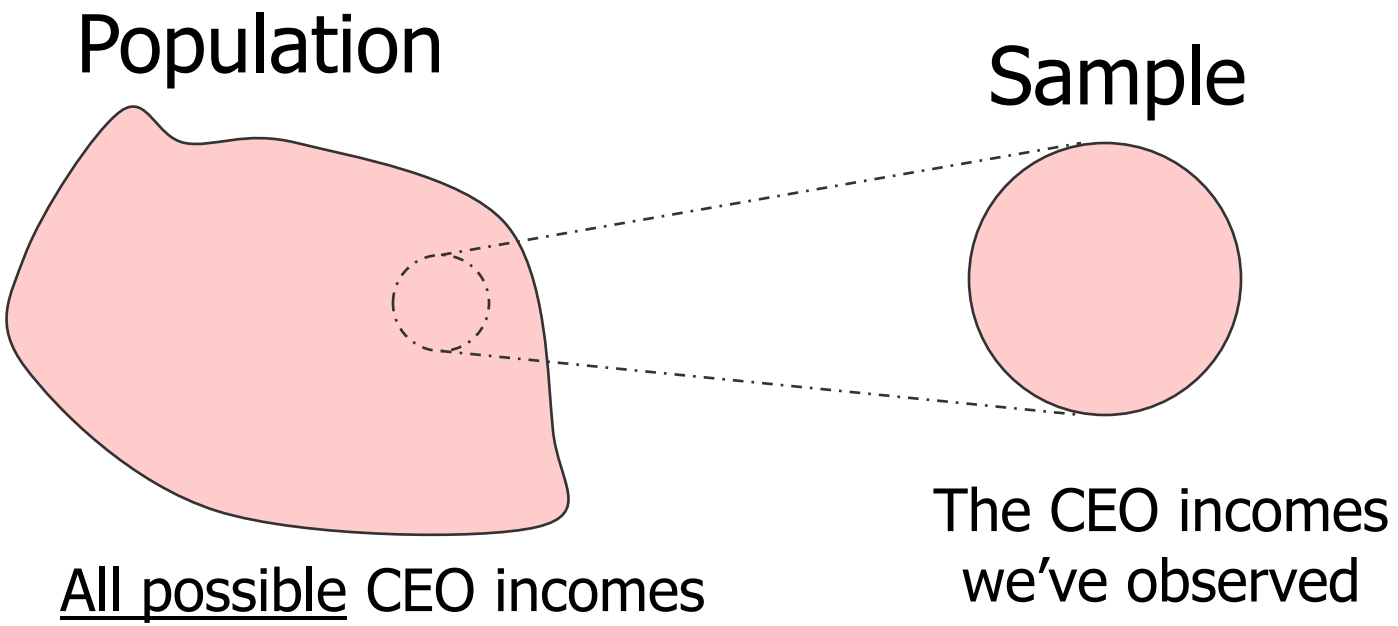
# Some Types of Variation

- Cross sectional
  - Data are a snapshot in time
  - Use one variable to explain another
- Time series (also called longitudinal)
  - Trend (long run changes)
  - Seasonality (retail sales up in December)
- Random
  - Not explained by anything
  - That's why we call it random!

# Samples versus Populations

- A population consists of all possible observations
  - Example: All students enrolled in an MBA program
- A sample is a subset of the population
  - Example: Global MBA students are a sample of all MBA students
- A *random* sample is a subset not drawn in any systematic way from population

# Samples versus Populations



# Why Sample?

If we could see these:

- The TV viewing preferences for every individual in the US
- The diameter of every shaft ever produced by a manufacturing process
- The proportion of potential customers who know of your product

We wouldn't need these:

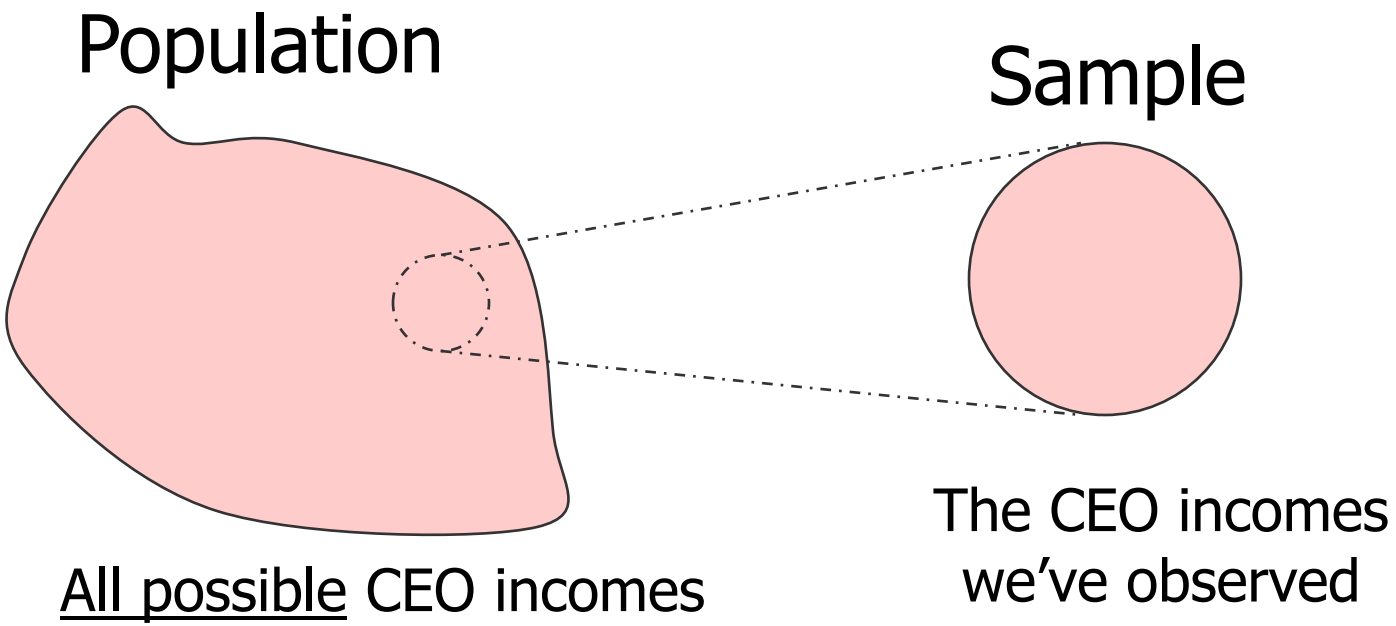
- Nielson survey of a sample of US television viewers
- The diameters of 100 shafts produced by the same process
- The proportion of individuals in a survey claiming knowledge of your product

✓ Collecting data for whole populations can be *expensive* and/or *impossible*

# Two Roles of Statistics

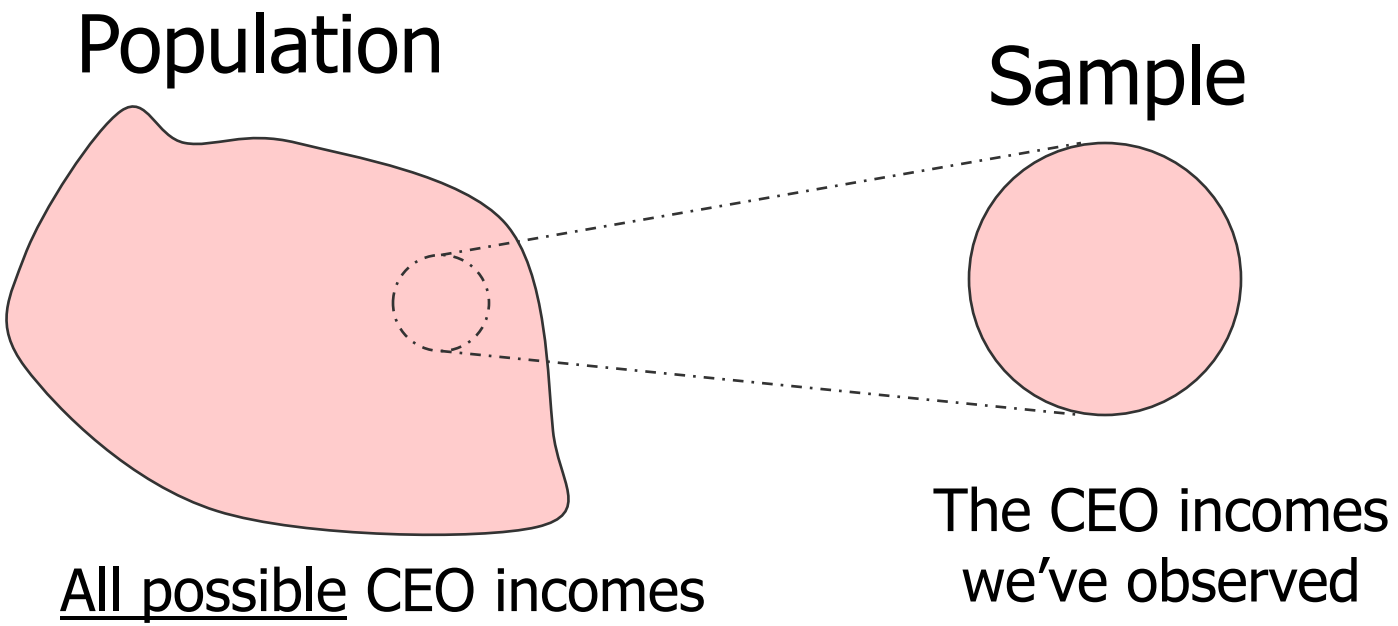
- Descriptive: Describing a sample or population
  - Numbers: (mean, variance, mode)
  - Pictures: (histogram, boxplot)
- Inferential: Using a sample to *infer* facts about a population
  - Making guesses (average income of MBA's)
  - Testing theories (does an MBA increase your income?)

# A Descriptive Question:




What is the average CEO income in our sample?

# An Inferential Question:

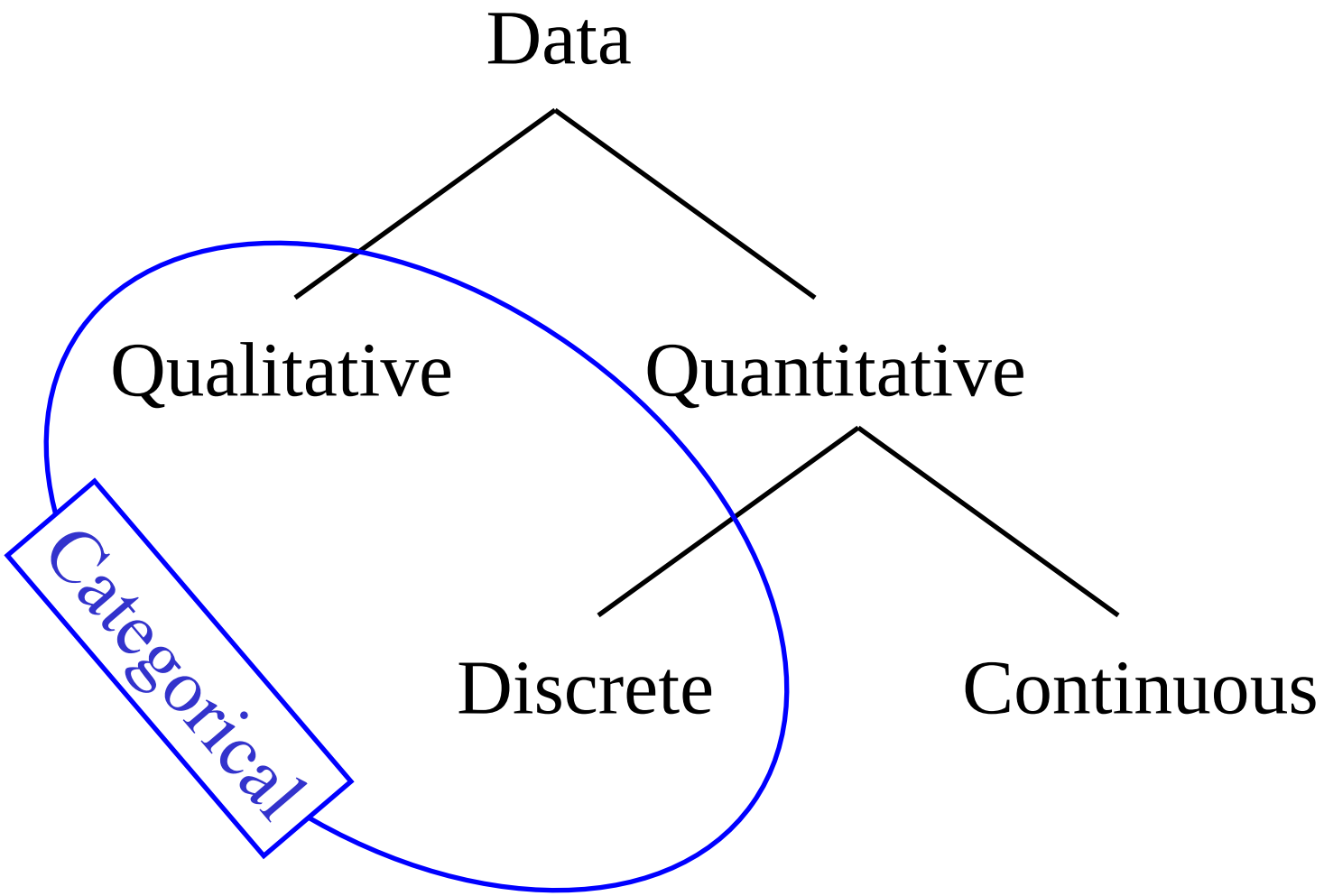



Given what we have observed, what can we say about the average CEO income for the population?

# Types of Data

- Continuous: Can divide by any number and result still makes sense
  - Examples: Salary, height, weight, age, etc.
- Categorical:
  - Nominal: unordered categories
    - Example: Country of origin, product color
  - Ordinal: ordered categories
    - Example: Small, medium, large
- Different types described in different ways

# Types of Data



# Notation

- Capital roman letters usually represent an unknown quantity
  - Example: What the outcome of a dice roll?
  - Label this outcome "$X$"
  - $X$ can be 1, 2, 3, 4, 5, or 6
- A small $i$ subscripted on a letter represents a series of observations
  - Example: The dice is rolled many times
  - $X_i$ is the outcome from the $i^{\text{th}}$ roll

# Notation

- A greek letter capital sigma ( $\sum$ ) means to sum up
  - Subscripts tell what to sum
- Example:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3$$

# Continuous Data

- Numerical Summaries
  - Location:

    Mean, median
  - Spread or variability:

    Variance, standard deviation, range, percentiles, quartiles, interquartile range
- Graphical Descriptions
  - Histogram
  - Boxplot
  - Scatterplot

Next class

# Sample Mean ( $\overline{x}$ )

- Sample average or sample mean
  - Often denoted by $\overline{X}$ (spoken “x-bar”)
- From previous example:

$$\overline{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{x_1 + x_2 + x_3}{3}$$

- In general: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

**Excel tip.** Use the built-in function:
= AVERAGE ( *cell reference* )

# Population Mean ($\mu$)

- Population mean
  - Often denoted by $\mu$ (Greek letter “mu”)
  - In general: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

**Excel tip.** Built-in AVERAGE function works for both samples and populations

# The Median

- The median is the “typical” value
- Steps to calculate the median:
  - Order your data from smallest to largest
  - If the number of data is odd, the middle observation is the median

    1 3 5 6 12 12 99

  - If the number is even, then the average of the two middle observations is the median

    1 3 5 6 12 12

    5.5

**Excel tip.** Built-in function:
= MEDIAN ( *cell reference* )

# Mean vs. Median

- Both are measures of location or “central tendency”
  - But, median less affected by outliers
- Example:
  - Imagine a sample of data: 0, 0, 0, 1, 1, 1, 2, 2, 2
    - Median=mean=1
  - Another sample of data: 0, 0, 0, 1, 1, 1, 2, 2, 83
    - Median still equals 1, but mean=10!
- Which to use? Depends on whether you are:
  - characterizing a “typical” observation (the median)
  - or describing the average value (the mean)

# Sample Variance ($s^2$)

- Sample variance measures data variability
- For $n$ observations, the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

**Excel tip.** Built-in function for *sample variance* = VAR ( *cell reference* )

# Population Variance ($\sigma^2$)

- Population variance measures data variability too
- For $N$ observations, the population variance is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

**Excel tip.** Built-in function for *population variance*
= VARP ( *cell reference* )

# Standard Deviation ($s$ or $\sigma$)

- The standard deviation is the square root of the variance

$$s = \sqrt{s^2}$$

- Also a measure of the variability
  - It's in the same units as the sample mean
  - For populations, the standard deviation is denoted

$$\sigma = \sqrt{\sigma^2}$$

**Excel tip.** Built-in functions for the *sample standard deviation* = STDEV ( *cell reference* ) or = STDEVP ( *cell reference* )

# Calculating Variance and SD

- Variance:
  - Sample numbers: 1 3 7 9 ← $X_i$
  - Mean: (1+3+7+9)/4 = 5 ← $\overline{X}$
  - Deviations from Mean: -4 -2 2 4 ← $X_i - \overline{X}$
  - Squared: 16 4 4 16 ← $\left(X_i - \overline{X}\right)^2$
  - Summed: 16+4+4+16 = 40 ← $\sum_{i=1}^{n} (X_i - \overline{X})^2$
  - Divide by n-1: 40/3 = 13.3333 ← $\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- Standard deviation:
  - SD = $\sqrt{13.333} = 3.65$

# The Range

- Range is another measure of variability
  - Denoted by R
- In words, it is the largest observation in the sample minus the smallest observation
  - Example: Imagine we collect the ages of students in the class
    - Data: 21, 23, 23, 25, 25, 26, 27, 31, 33, 33, 35
    - Range = 35 - 21 = 14

# Other Measures of Variation

- Percentiles
  - $p$th percentile: value of $x$ such that $p$% of the data is less than or equal to $x$
  - Special Percentiles:
    - Max: 100th percentile
    - Min: 0th percentile
    - Median: 50th percentile
    - Quartiles: 25th and 75th percentiles
- Interquartile Range (IQR):

  IQR = 75th percentile - 25th percentile

# Categorical Data

- Numerical Measures:
  - Mode: most commonly occurring value
  - Frequency table: how often each value occurs
- Graphics:
  - Bar chart of frequencies (histogram)
  - Mosaic chart (stacked bar chart)
  - Pareto chart

Next class

# Mode

- Mode is the most frequently occurring value in the sample or population
  - It is the “typical” or “common” value
- For example, in the following data

  1, 1, 1, 1, 2, 2, 2, 3, 4, 5, 5, 6, 7

  the mode is “1”
  - “1” occurs 4 times
  - All other observations occur less than 4 times

# Frequency Tables

- Tables of counts by two or more categorical variables
- Example: Executive compensation (Forbes94.jmp)

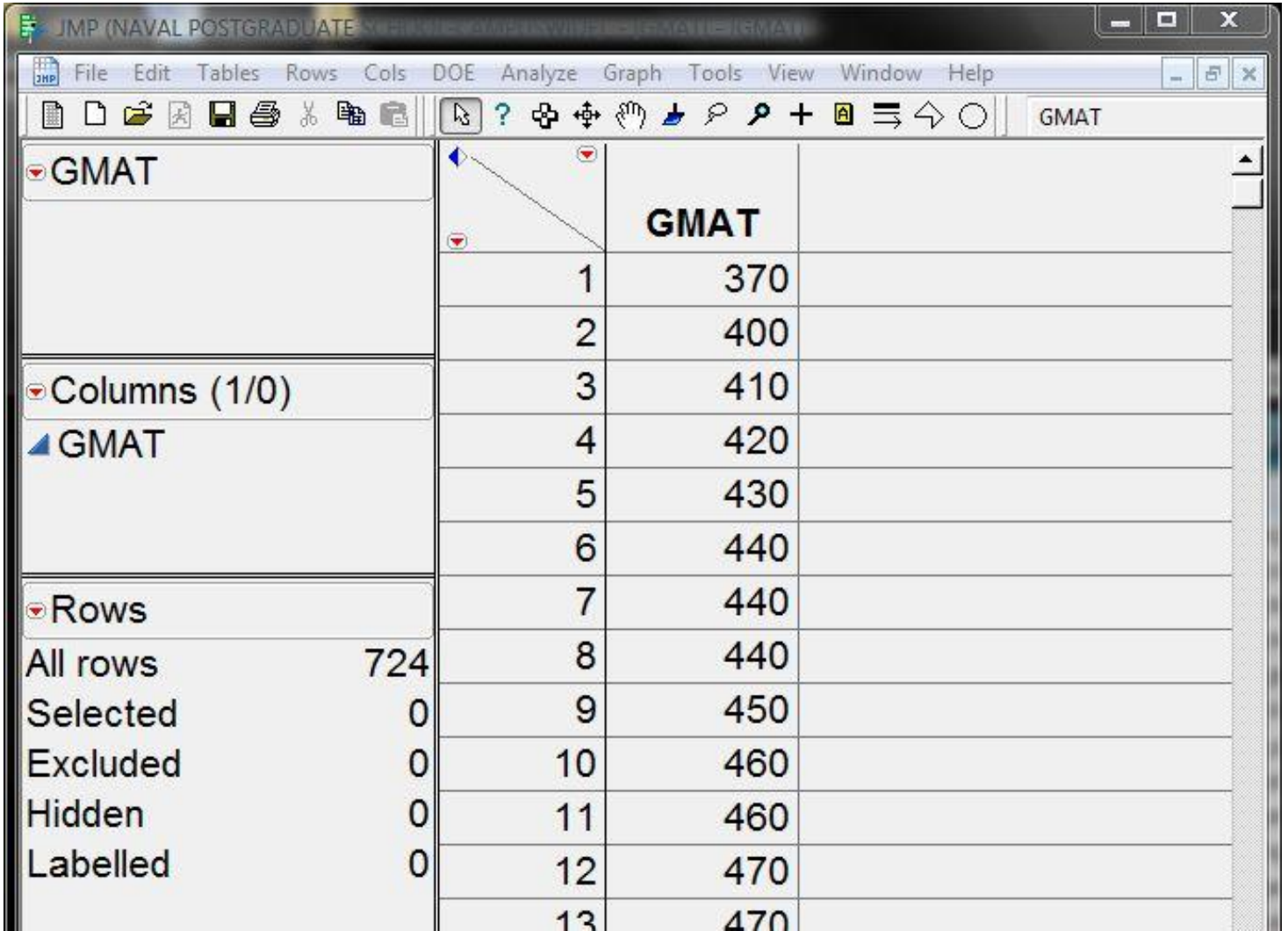
| WideIndustry | MBA | | |
|---|---|---|---|
| Count | No | Yes | |
| Aerospacedefense | 15 | 4 | 19 |
| Business | 22 | 5 | 27 |
| Capital goods | 13 | 6 | 19 |
| Chemicals | 17 | 8 | 25 |
| ComputersComm | 46 | 21 | 67 |
| Construction | 7 | 4 | 11 |
| Consumer | 36 | 18 | 54 |
| Energy | 32 | 10 | 42 |
| Entertainment | 23 | 4 | 27 |
| Financial | 112 | 56 | 168 |
| Food | 45 | 17 | 62 |
| Forest | 12 | 7 | 19 |
| Health | 37 | 12 | 49 |
| Insurance | 43 | 11 | 54 |
| Metals | 14 | 4 | 18 |
| Retailing | 43 | 3 | 46 |
| Transport | 11 | 4 | 15 |
| Travel | 13 | 2 | 15 |
| Utility | 48 | 15 | 63 |
| | 589 | 211 | 800 |

# Introduction to JMP

- Statistical analysis software
  - More powerful than Excel for statistical analyses
  - Designed to facilitate analyses and to do advanced statistics
  - Particularly good at interactive analyses
    - Interactive graphics
    - Delete points and repeat analysis
    - Conduct multiple analyses

# Introduction to JMP

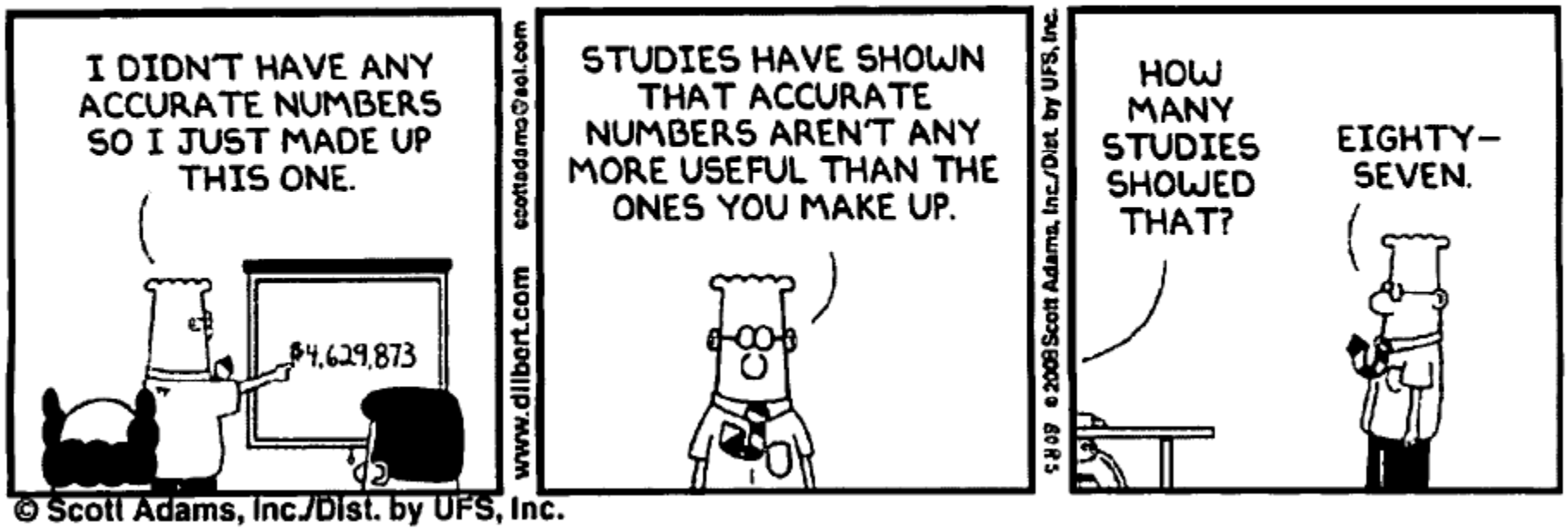
- Demonstration using GMAT case study (GMAT.jmp)



# Remember the Notation

- Summation
  - $\Sigma$ notation and subscripts
- Size
  - $n$ denotes size of sample
  - $N$ denotes size of population
- Knowns vs. unknowns
  - Small letters (i.e., "$x$"): quantity is known
  - Capital letters (i.e., "$X$"): quantity unknown
    - Later we will call these random variables

# People Will Believe Any Statistic…



# What We've Covered

- Introduced professor & course
- Defined some basic statistics terminology
  - Populations vs. samples
  - Descriptive vs. inferential statistics
- Learned about some numerical descriptive statistics
  - Measures of location
  - Measures of dispersion
- Introduced JMP