ECO 6416: Review

Joshua L Eubanks

University of Central Florida

What is Statistics?

Statistics is a collection of tools and methods applied to **data** which helps managers to:

- Make Decisions
- Analyze Data
- Identify Patterns, Trends, and Relationships
- Identify and understand variation
- Use samples to draw conclusions about the population

Statistical Decision Making

Statistics is the field of study dealing with the collection, analysis, presentation and interpretation of data, and how to interpre data as evidence to make decisions when faced with uncertainty. Statistical Decision making is done in 4 major steps:

- Identify the relevant population and variables
- Locate the data
- Analyze the data
- Generate a statistical report

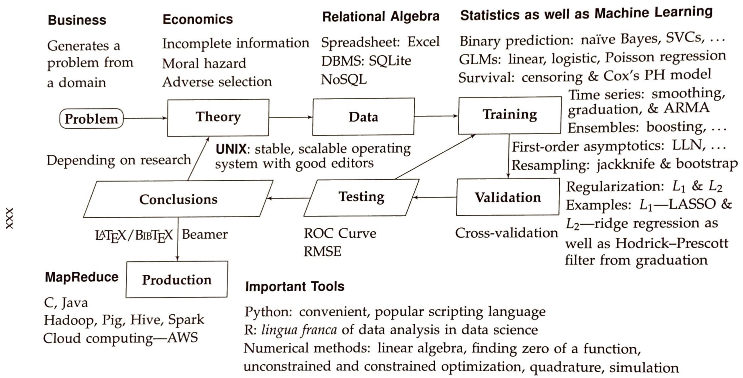

Business
Generates a problem from a domain
Economics
Incomplete information
Moral hazard
Adverse selection
Relational Algebra
Spreadsheet: Excel
DBMS: SQLite
NoSQL
Statistics as well as Machine Learning
Binary prediction: naïve Bayes, SVCs, …
GLMs: linear, logistic, Poisson regression
Survival: censoring & Cox's PH model
Time series: smoothing, graduation, & ARMA
Ensembles: boosting, …
Problem
Theory
Data
Training
Depending on research
UNIX: stable, scalable operating system with good editors
First-order asymptotics: LLN, …
Resampling: jackknife & bootstrap
Conclusions
Testing
Validation
Regularization: L_1 & L_2
Examples: L_1—LASSO & L_2—ridge regression as well as Hodrick–Prescott filter from graduation
xxx
LaTeX/BibTeX
Beamer
ROC Curve
RMSE
Cross-validation
MapReduce
Production
C, Java
Hadoop, Pig, Hive, Spark
Cloud computing—AWS
Important Tools
Python: convenient, popular scripting language
R: *lingua franca* of data analysis in data science
Numerical methods: linear algebra, finding zero of a function, unconstrained and constrained optimization, quadrature, simulation

Figure 5 Methods and Tools of Business Analytics in Workflow of Problem Solution

Figure 1: Foundations of Empirical Intuition - Harry J. Paarsch

Data

Data is information, especially facts or numbers, collected to be examined and considered and used to help decision-making - Cambridge Dictionary

For information to be data, it must have context:

- **Who** are the individuals being measured or counted?
- **What** has been measured or counted?
- **When** or for which time period were the data measured?
- **hoW** were the data measured and recorded?
- **Why** were the data measured and recorded?

Data Types

- **Categorical:**
  - Often recorded as names, but can be recorded as numerals
    - Ex: Grade level, county, gender
  - Usually dealt with by counting how many cases fall within each particular category
- **Quantitative:**
  - Measurements or counts of something
    - Ex: Dollars, cases, temperature
  - **ALWAYS** has units of measure attached

Data Types

- **Cross Section:**
  - data collected during a single time period
  - snapshot of similar things: stores, workers, cities
- **Time Series:**
  - data collected at regular time intervals
    - hour, day, week, month, quarter, year
  - distinct business feature: relies on time series
  - businesses and government are ongoing institutions

How Data is Summarized and Displayed

- **Univariate Analysis: Single Variable (Center/Shape/Spread)**
  - *Descriptive Statistics*
    - Center (mean, median, mode)
    - Shape (unimodal, symmetric)
    - Spread (standard deviation, high/low, range)
  - *Histograms* of frequency distribution
- **Bivariate Analysis: Relate two variables (Direction/Strength/Form)**
  - Correlation, Regression Equations
  - Scatterplots, Trend Lines
- **Multivariate Analysis: Relate three or more variables**
  - Regression Equations and ANOVA Models
  - Often times, data in 3 or more dimensions is hard to visualize

Univariate analysis: Center

- Population Mean: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ where N is the size of the population, x_i is the value for each member of the population .

```r
grades <- c(78,79,80,81,82)
mean(grades)
```

```
## [1] 80
```

- Median: The middle value in a sorted list. If the dataset is an even amount, you take the average of the two closest values to the center

```r
median(grades)
```

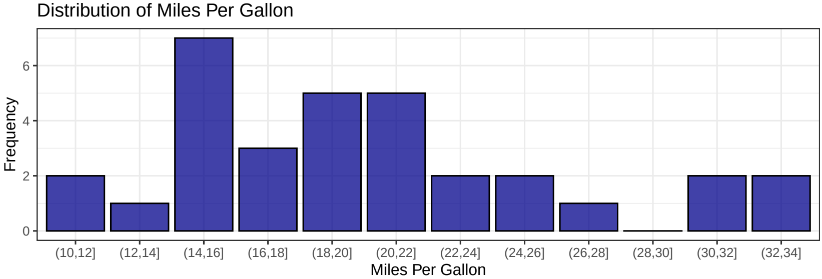
```
## [1] 80
```

- Mode: Which value occurred the most, also tells you if the distribution is unimodal

Univariate analysis: Shape

The mode can help tell you the distribution, however the easiest way is with a histogram. Also another way to see if the distribution is symmetric is if the mean is approximately equal to the median.

If the mean is greater than the median, then it is skewed-right, if it is less than the median, then it is skewed left



Data is from 1974 Motor Trend US Magazine

Univariate Analysis: Spread

- Variance: $\sigma^2 = \frac{\sum_i^N (x_i - \mu)^2}{N-1}$ = Sum of squared deviations of x_i from μ divided by degrees of freedom

```
var(grades)
```

```
## [1] 2.5
```

- Standard Deviation $= \sigma = \sqrt{\sigma^2} =$ Square root of variance

```
sd(grades)
```

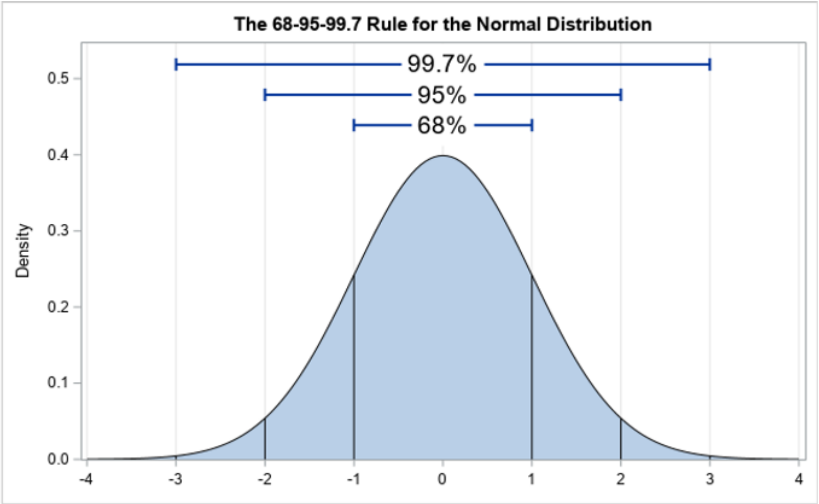
```
## [1] 1.581139
```

- Interquartile Range (IQR): Difference between the 75th and 25th percentile

```
IQR(grades)
```

```
## [1] 2
```

Normal Distribution and the Z-Score



The Z-Score tells you how many standard deviations (σ) an observation x_i is from the mean (μ) where $Z = \frac{x_i - \mu}{\sigma}$

Predictions of a Univariate Model with a Normal Distribution

If the population is unimodal and symmetric then:

$$Prediction = \mu \pm cv * \sigma$$

where μ = population mean, cv = critical value = Z-value, σ = standard deviation.

If you want 68%, 95%, or 99% accuracy, the critical values are 1,2,3 respectively. Some businesses like using 1.3 as the critical value for 80% confidence.

Univariate Data that is not Normally Distributed

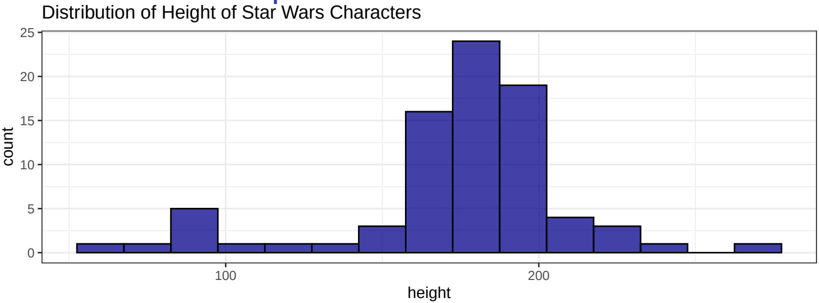
In many business applications, you will find that varaibles do not always follow a normal distribution. There are many reasons such as:

- Bi or Multimodal
  - There might be an underlying reason as to why there are multiple modes, you may want to break them down by different categories
- Outliers

Univariate Outliers

- Clearly larger or smaller than the rest of the data
- Can find using histograms
- **Mean** and **standard deviation** are very sensitive to outliers
  - If you have both high and low outliers, the mean will not be affected much as the outliers would cancel eachother out, but outliers always increase the standard deviation
- The median is not affected by outliers
  - median is a resistant measure because it is ordinal
  - You will see large differences between the mean and median as a result
- You can also detect outliers using Tukey's Fences:
  - Lower fence: $Q1 - 1.5 * IQR$
  - Upper fence: $Q3 + 1.5 * IQR$

Starwars Mass Example



```
mean(starwars$height,na.rm = T)
```

```
## [1] 174.358
```

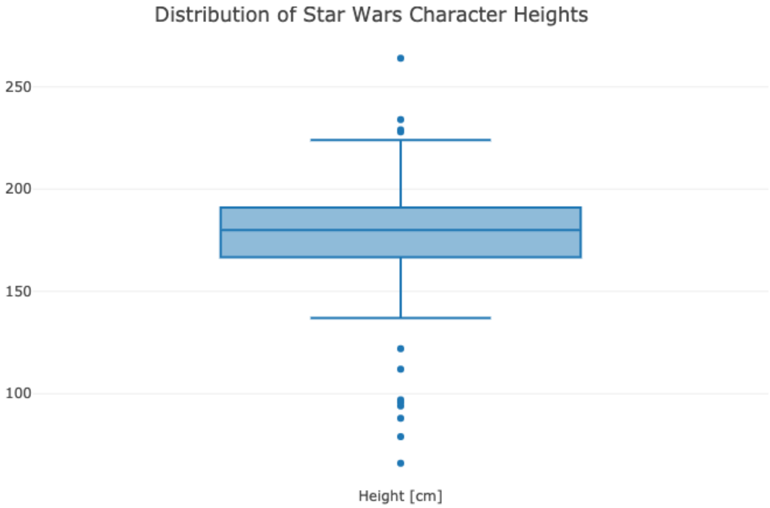
```
median(starwars$height,na.rm = T)
```

```
## [1] 180
```

```
sd(starwars$height,na.rm = T)
```

```
## [1] 34.77043
```

Box Plot


Distribution of Star Wars Character Heights
250
200
150
100
Height [cm]

Bivariate Statistics: Correlation

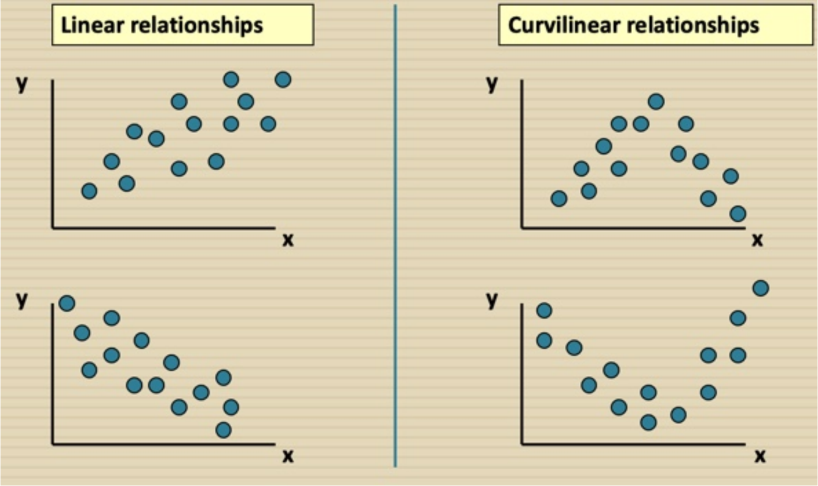
This measure helps determine if there is a strong linear relationship between two quantitative variables.

- Unit free
- Range is [-1, 1]
- Closer to -1, the stronger the negative linear relationship
- Closer to 1, the stronger the positive linear relationship
- Closer to 0, the weaker the linear relationship
- Correlation is not causation. The correlation between x and y is the same as y and x, so it tells you nothing about causation. In fact, there might be a lurking variable that is impacting both

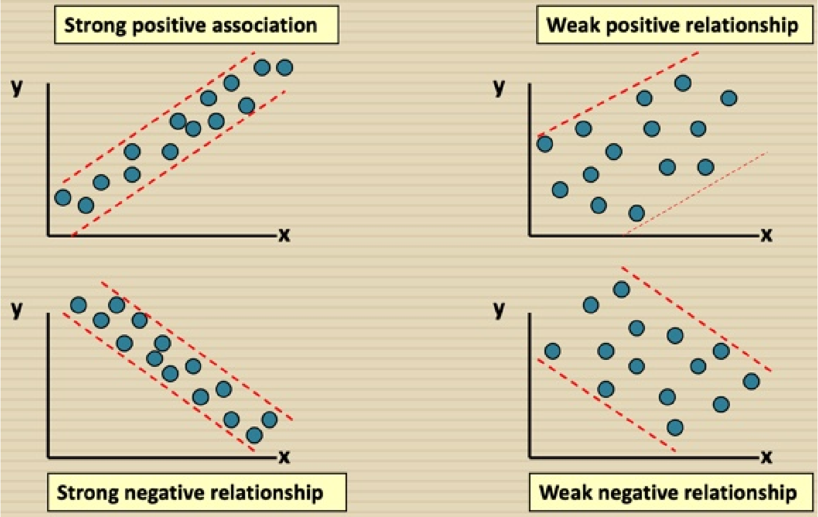
```
hours <- c(3,2,2,5,4)
cor(grades,hours)
```

```
## [1] 0.6063391
```

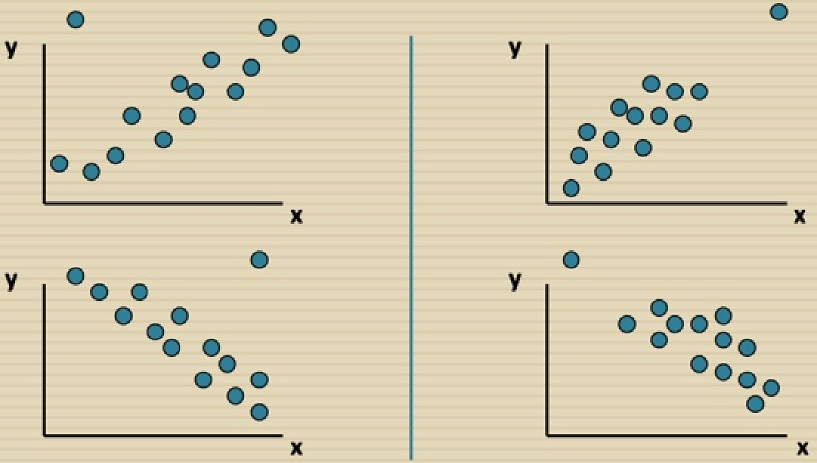
Linear vs Curvilinear Relationships



Strength of Linear Relationships



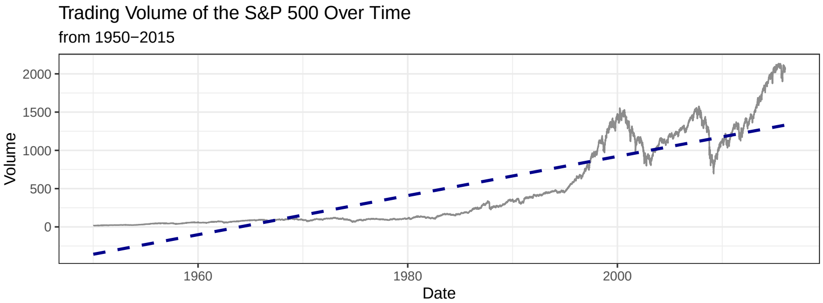
Outliers and Correlation



Time Series Graph

This is the nearly the same as the scatterplot, the only difference is that time is on the x-axis

Also, there is typically a straight line through the chart called the trend-line



Organize, Display, and Summarize Data

- Data Tables
  - organize information in context
- Tally sheets, bar and pie graphs
  - display frequencies
- Histograms
  - display shape, center, spread of univariate quantitative data
- Scatterplots and correlation
  - display and summarize direction, form, strength of bivariate quanitative data
- Trend plot Analysis
  - display and summarize direction, form, strenth of bivariate quantitative tim series data
- Equations
  - summarize two or more variables

Summary of Key Terms

- A *frequency distribution* is a table that reports the number of times each numerical value or grouping of values occurs in quantitative data.
- The *mode* is the most frequently occurring value in the data.
- A *histogram* is a bar graph display of a frequency distribution with the frequencies represented by bar heights.
- A *class* is an interval of values for a frequency distribution and the corresponding histogram bar. In software programming it is referred to as the bin. The bin is defined by *class width* and *cut points*, which are the beginning and ending values of each class.
- A frequency distribution is *symmetric* if the left half of the histogram is the mirror image of the right half.
- A frequency distribution is *unimodal* if the histogram has one mode or modal class and is bimodal if the histogram bars group around two separate modes.
- A unimodal distribution is *skewed right* if its histogram has a long tail to the right of its modal class and *skewed left* if its tail lies only to the left.

R and RStudio

- Why R?
  - Open source
  - Easier to learn than python
    - Not whitespace sensitive
    - Easy to install packages
    - Pairs well with RStudio
- RStudio
  - In my opinion, it is the best graphical user interface (GUI) available
  - Allows for generation of many styles of reports with code embedded
  - Easier to follow along in the coding process

Installing R and RStudio

1. First, download and install R, which is available at https://www.r-project.org
2. After downloading and installing R, RStudio provides an interactive environment for using R, available at https://www.rstudio.com

Help in R

- There are a bunch of cheatsheets available in RStudio, just click the **Help** at the top of the window then **Cheatsheets**
- Also there is a package called *swirl* that can be used

```
install.packages("swirl")
library(swirl)
```

Other Software

We are only going to cover R in this course, however, there are many other languages that are important. Two main languages that are imperative for any individual in this field are:

- SQL: https://mode.com/sql-tutorial/
- python: https://mode.com/python-tutorial/
    - Quantitative Economics: https://quantecon.org/

Optional: Version Control

An important process for any individual is to ensure that you have the ability to track your code or any work. For this course, we will be using *git*. This will act as your version control throughout the course.

If you are using a Mac, open the terminal and run these two lines of code:

- /bin/bash -c “$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)”
- brew install git

If you are using Windows:

- https://git-scm.com/download/win