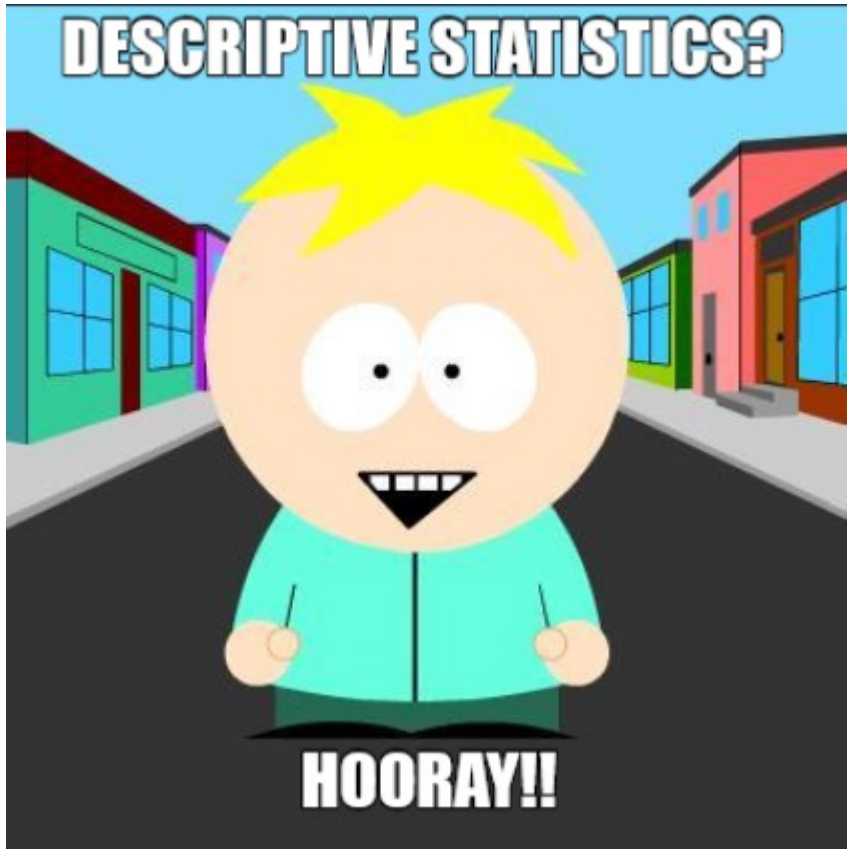
Basics to R

Descriptive and inferential statistics, and data visualization

October 11th, 2022

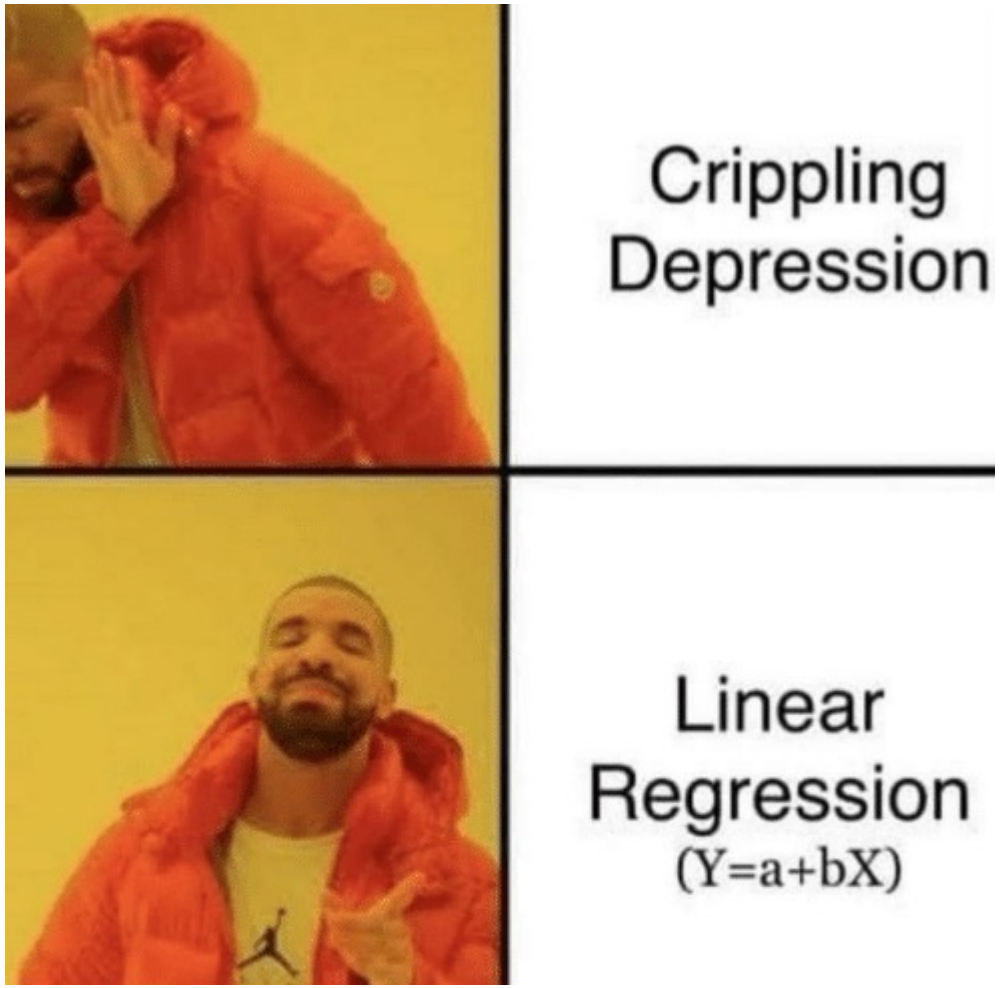
Lecture Part 1: Descriptive Statistics

- What are descriptive statistics?
- Why are they important?
- What do I need to look at?

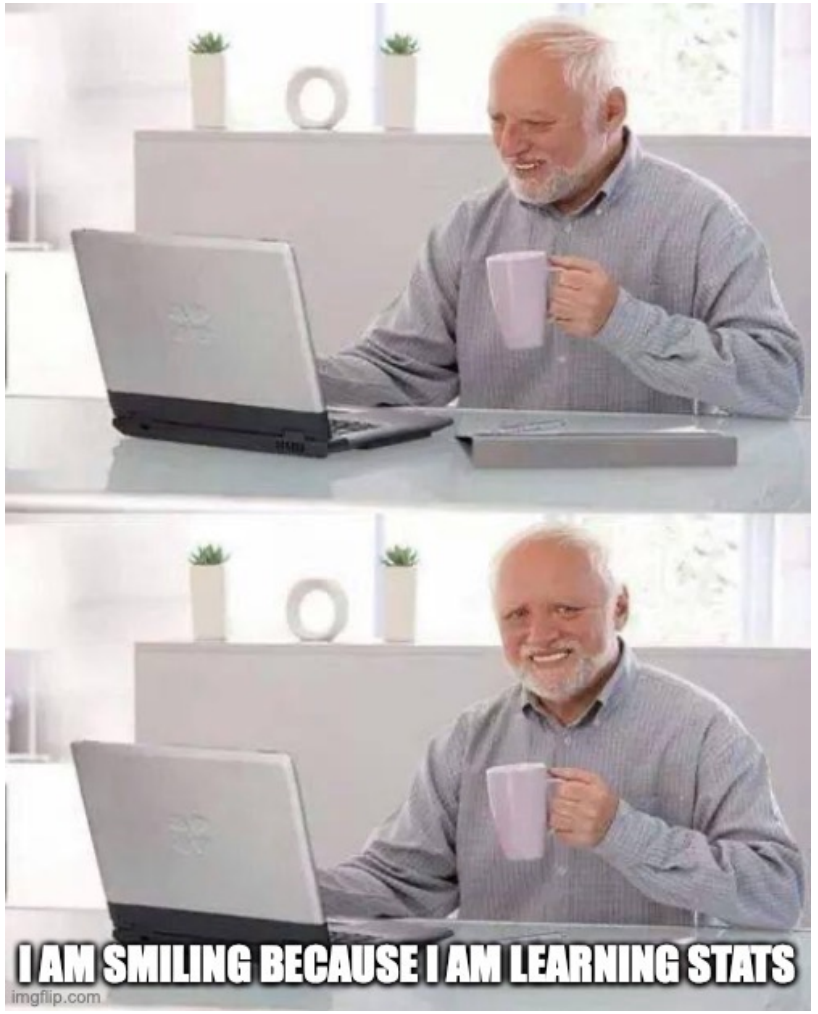


Lecture Part 2: Inferential Statistics

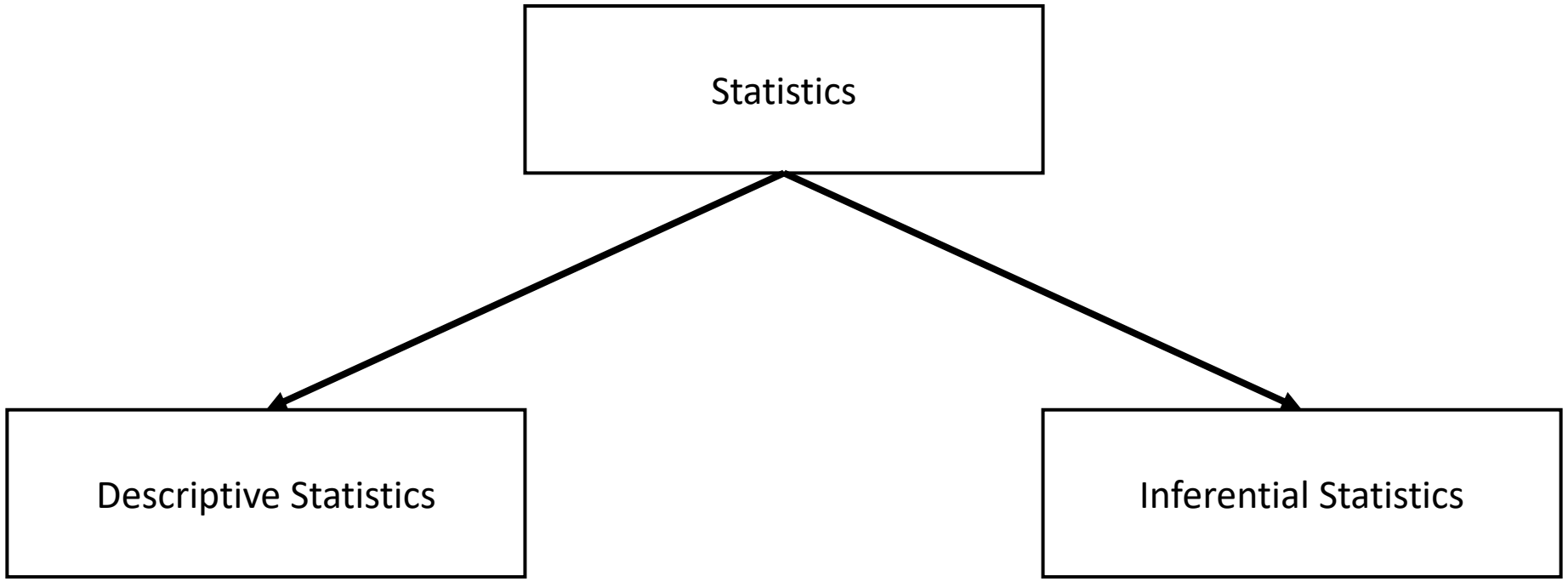
- What are inferential statistics?
- The General Linear Model
- Correlation, Linear Regression, and ANOVA
- What do I need to look at?



Part 1: Descriptive Statistics

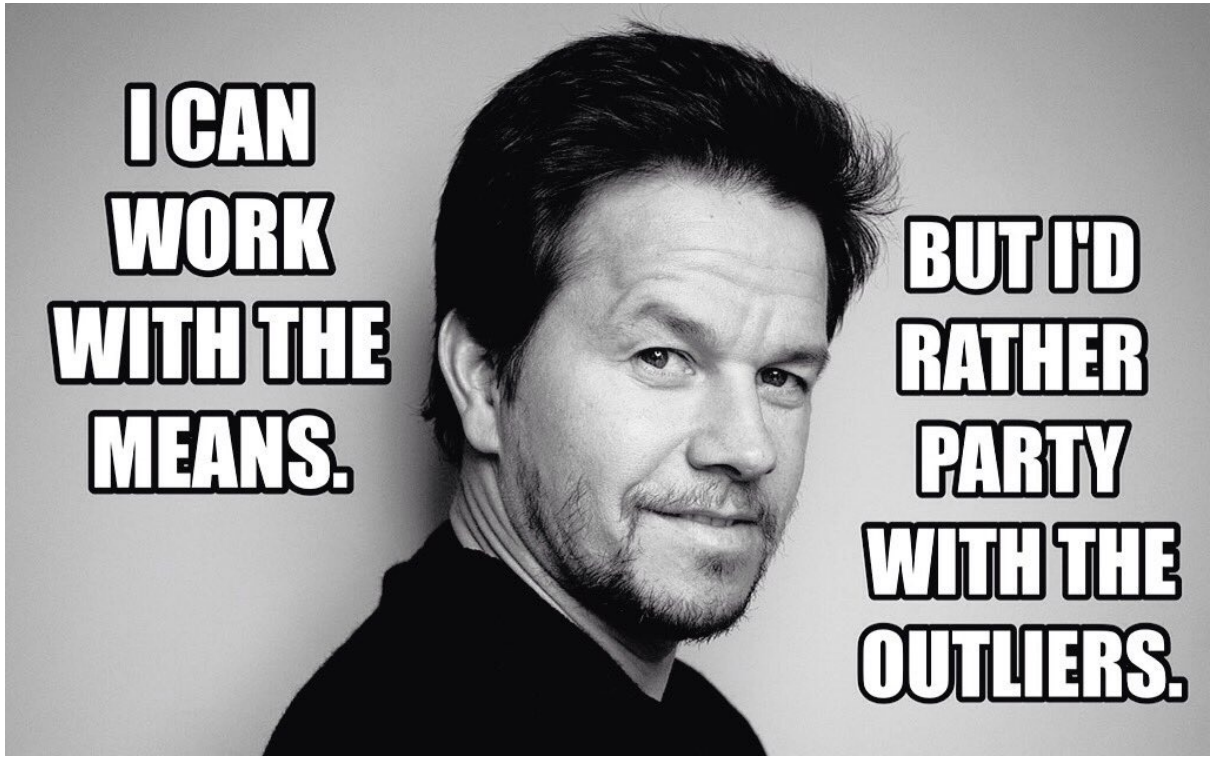


What are descriptive statistics?



What are descriptive statistics?

- Simple summaries about the sample
- Can be quantitative (e.g., the mean) or visual (e.g., histograms)
- Common descriptive statistics:
  - **Measures of central tendency:** mean, mode, median
  - **Measures of variability:** standard deviation, variance, range, IQR
  - **Modality**
  - **Skew**
  - **Kurtosis**



But why should I care?

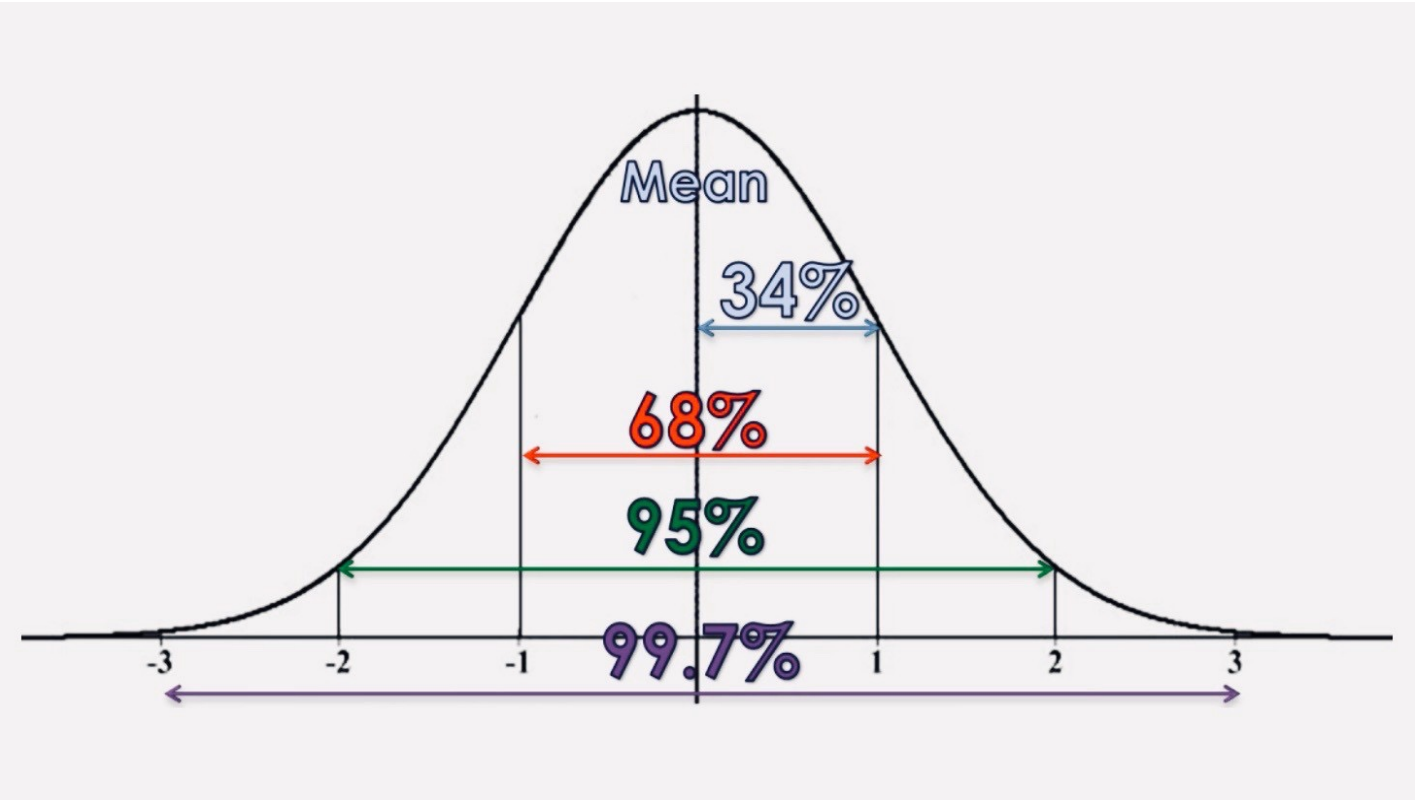
- Know thy data, know thyself
- All inferential statistics are based on assumptions
  - If you're data don't meet the assumptions, then you're conclusions may be false
- Skew, kurtosis, variance, and distribution

Variance

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

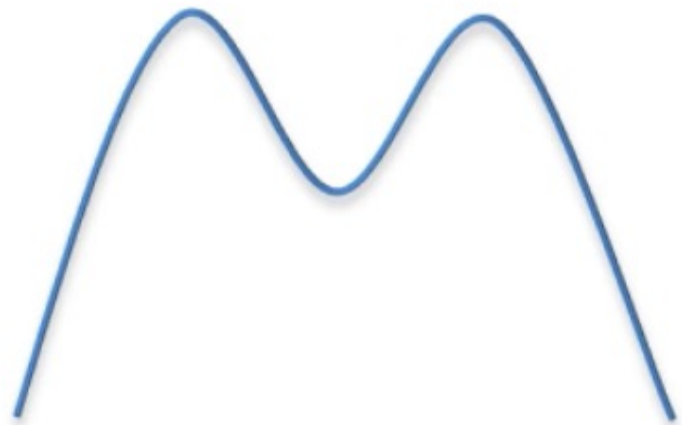
Standard Deviation

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

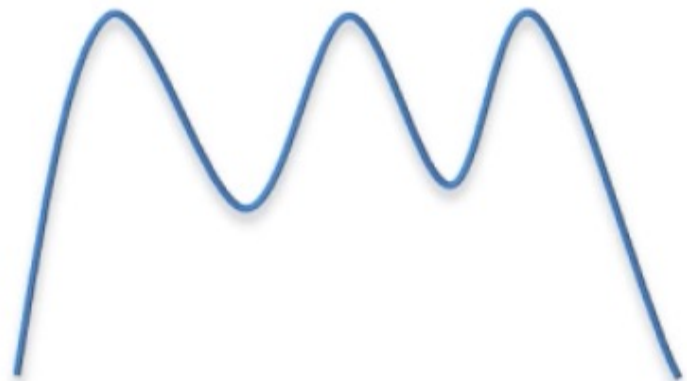


Unimodal

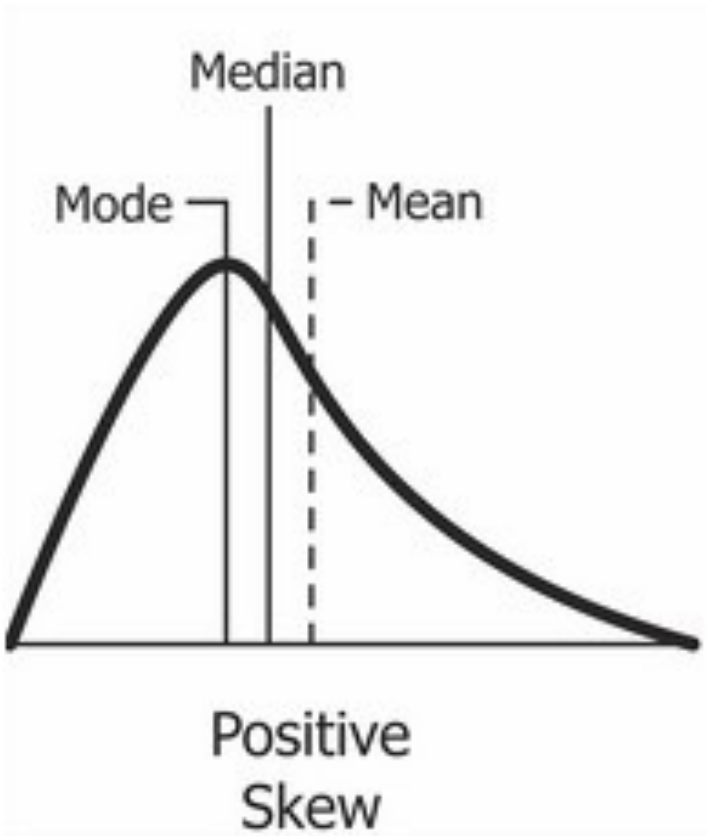
Bimodal



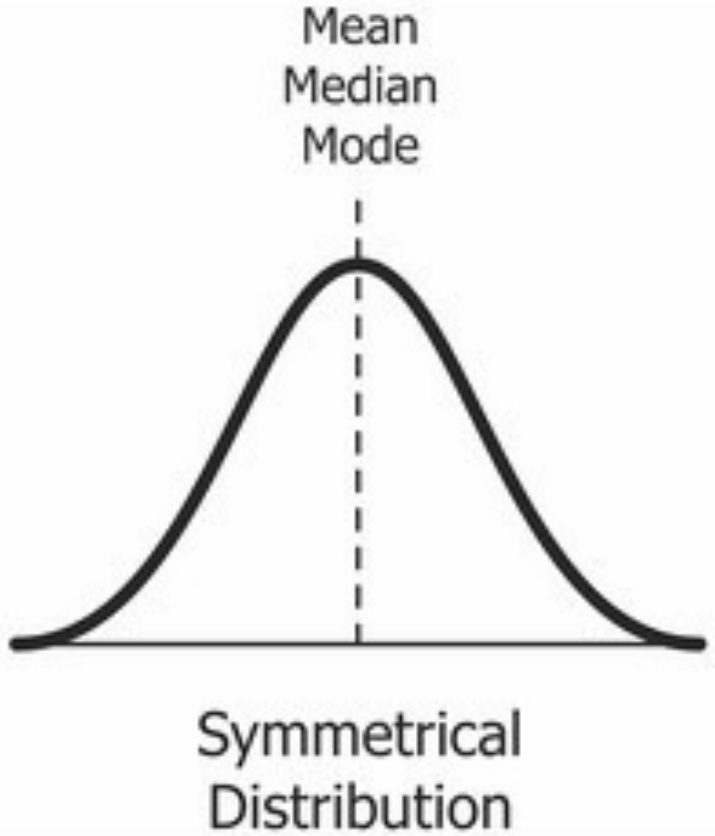
Multimodal



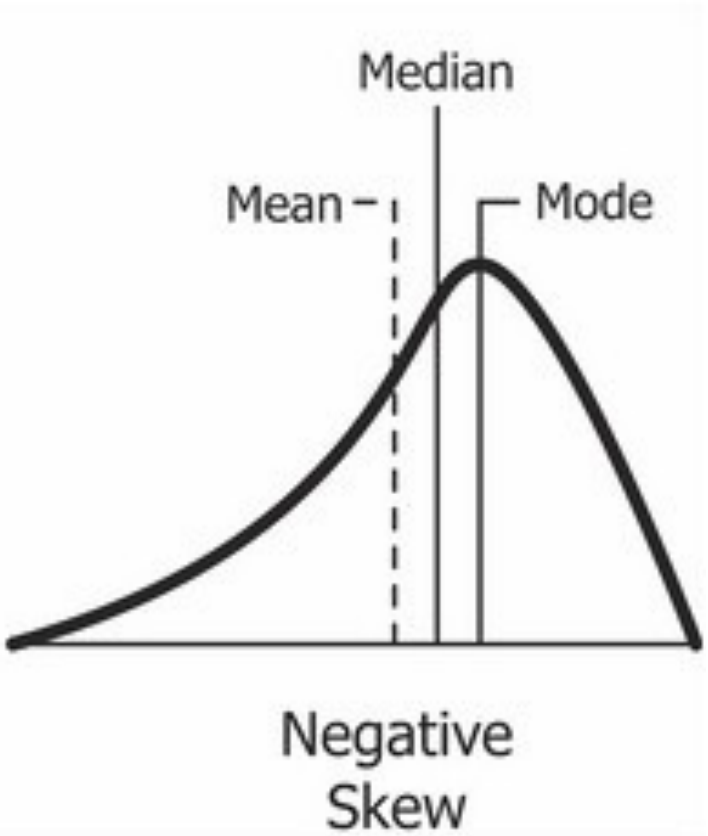
AKA *The Dream Distribution*



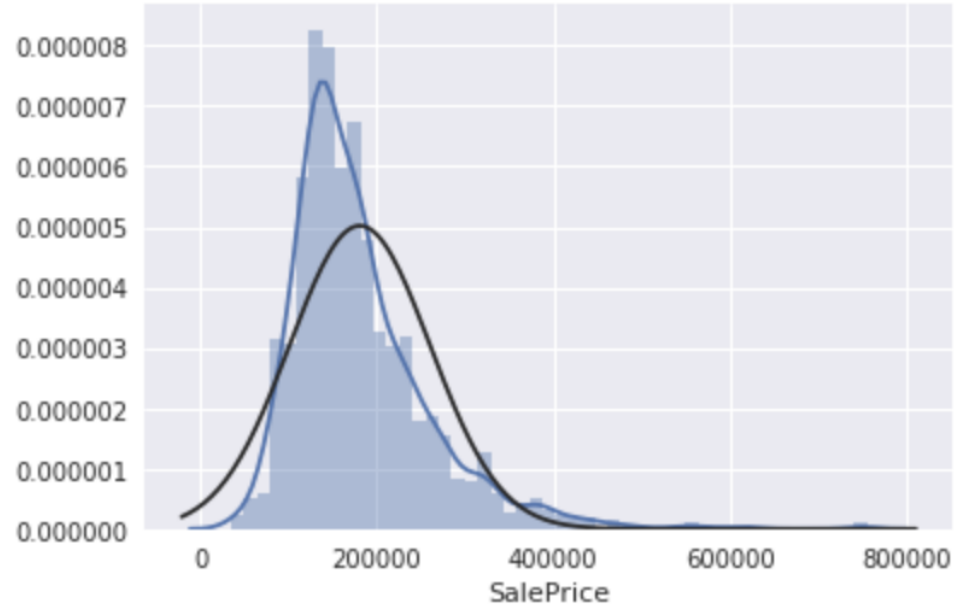
Positive Skew

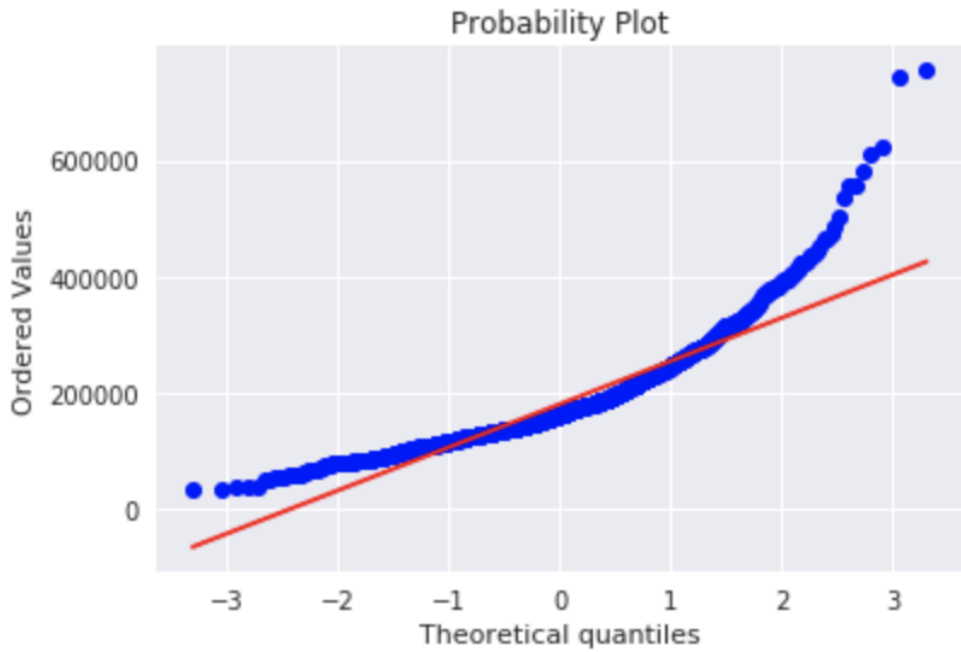


Symmetrical Distribution



Negative Skew

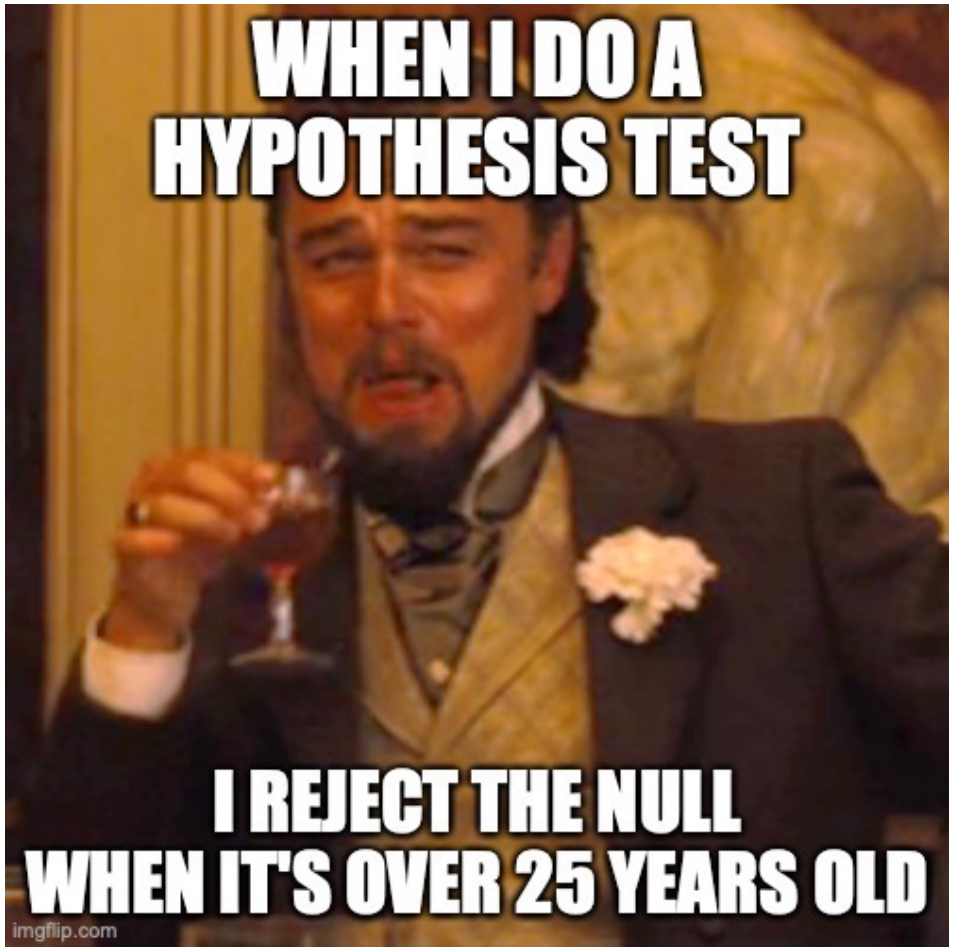

0.000008
0.000007
0.000006
0.000005
0.000004
0.000003
0.000002
0.000001
0.000000
0
200000
400000
600000
800000
SalePrice


Probability Plot
600000
400000
200000
0
Ordered Values
−3
−2
−1
0
1
2
3
Theoretical quantiles

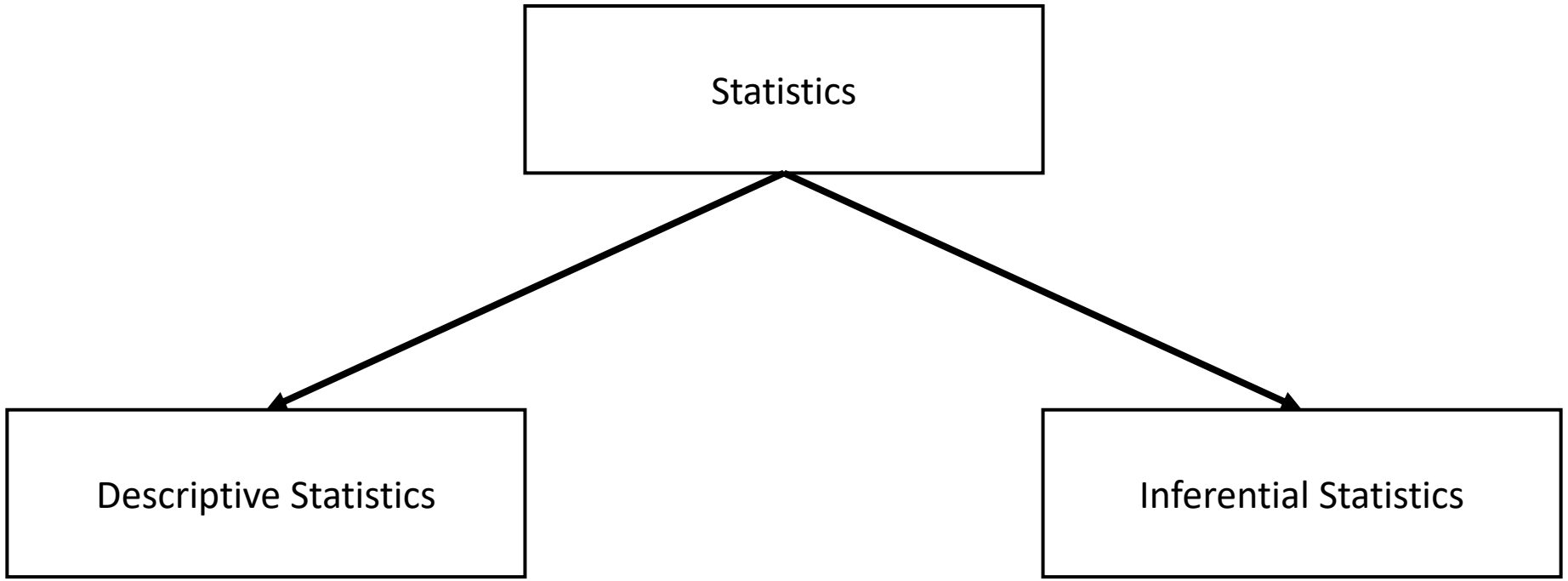
Summary

- Descriptive statistics are not very interesting
- Not looking at descriptive statistics is bad
- Ryan Gosling stats memes are great ways to end a PowerPoint

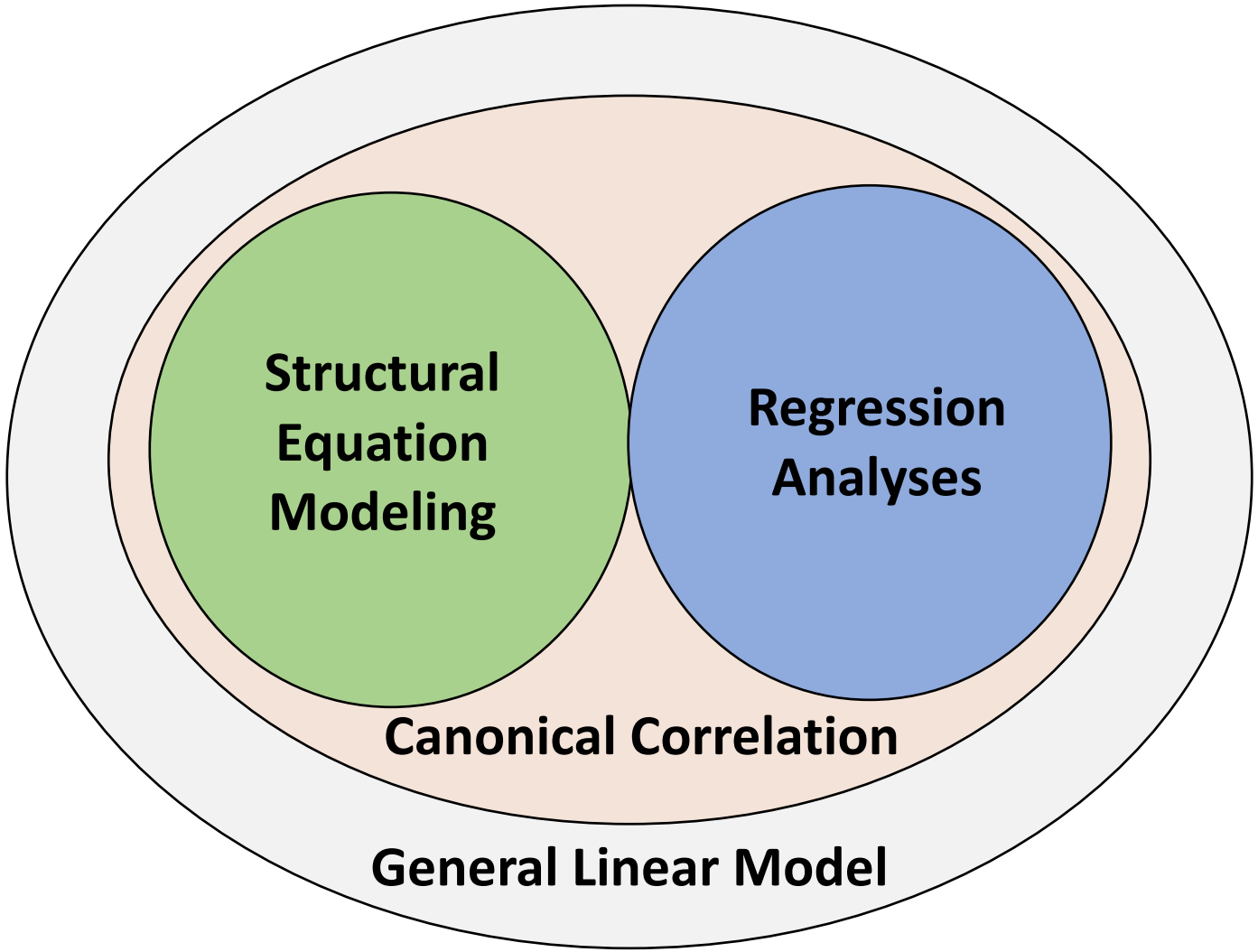
Part 2: Inferential Statistics



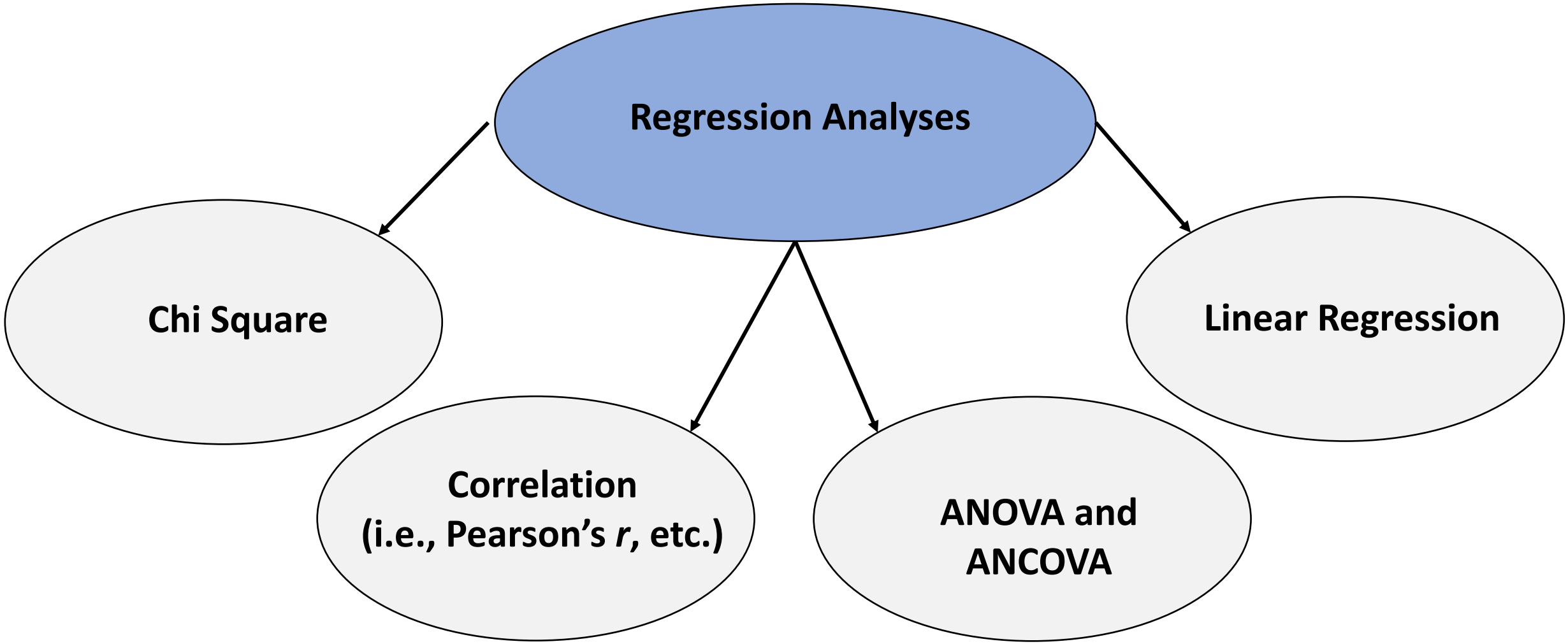
What are inferential statistics?



What are inferential statistics? A very brief introduction to the General Linear Model



What are regression analyses?



What are the assumptions of regression analyses?

- Linearity
- Independence
- Normality
- Equivariance (i.e., Homoscedasticity)

Assumption #1: Linearity

The regression model is linear in parameters

$Y = \alpha + (\beta 1 * X1) + (\beta 2 * X2)$

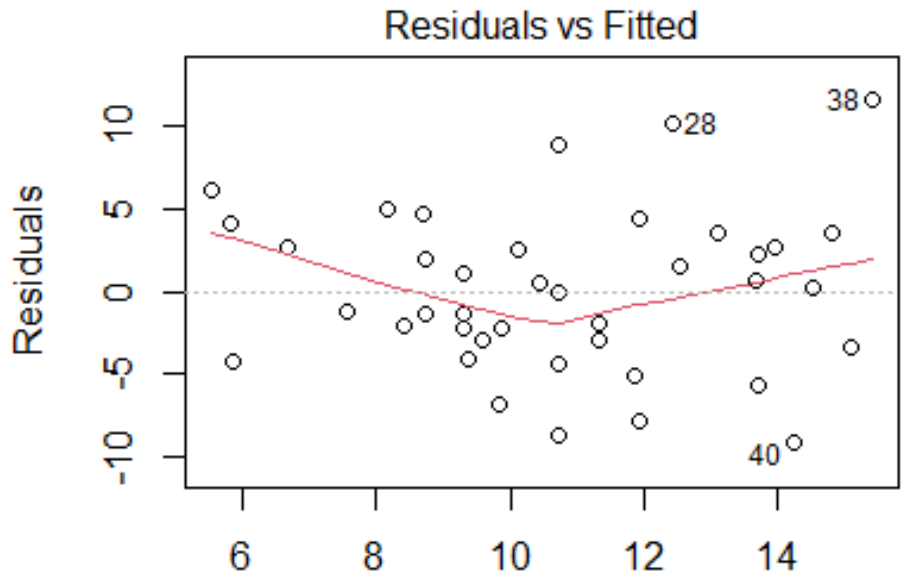
$Y = \alpha + (\beta 1 * X1) + (\beta 2 * X2^2)$

$Y = \alpha + (\beta 1 * X1) + (\beta 2 * X2^2) + (\beta 2 * \ln(X2))$

$Y \neq \alpha + (\beta 1 * X1) + (\beta 2 * X2)^2$

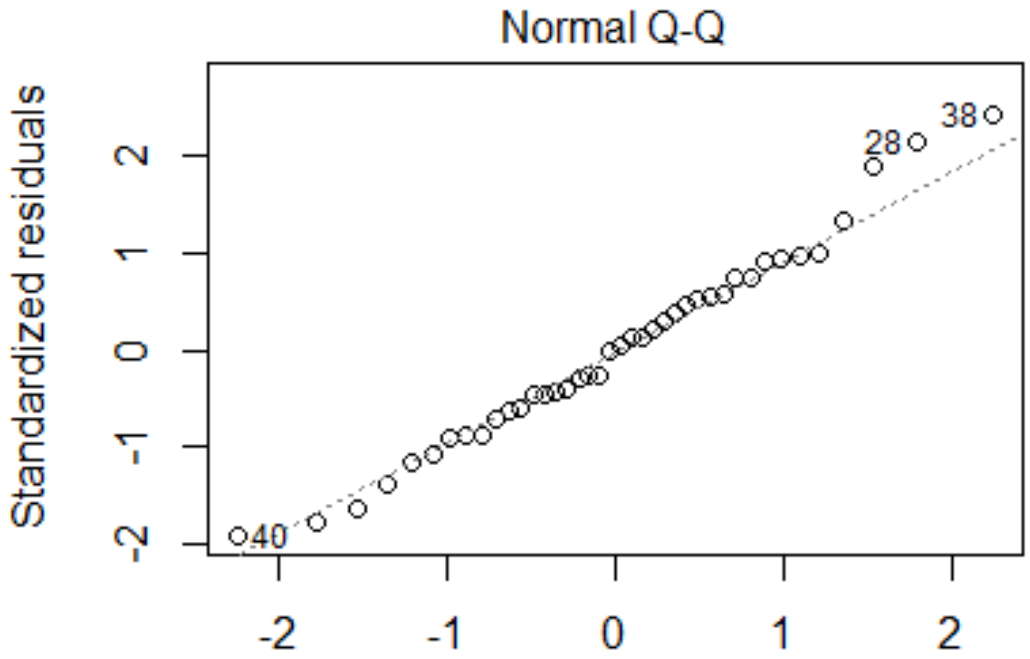
Assumption #2: Independence

Observations are independent of each other



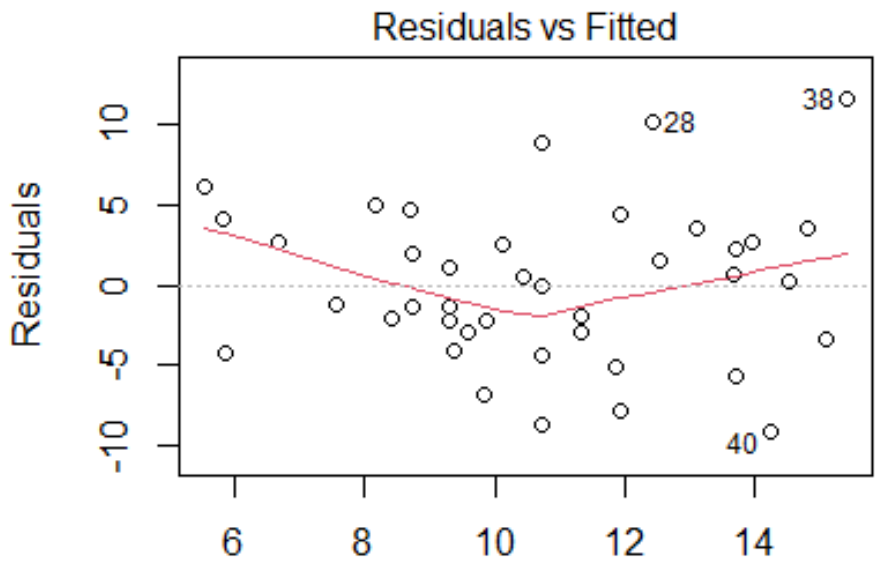
Assumption #3: Normality

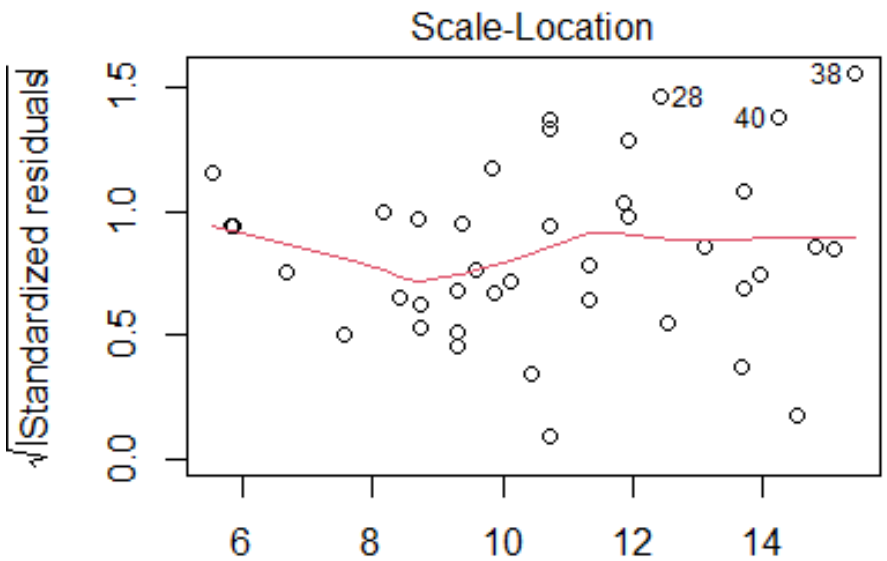
Residual errors are normally distributed



Assumption #4: Equivariance or Homoscedasticity

<u>The variance of residuals should not increase with fitted values of the response variable.</u>





Summary

- The fundamental basis of most statistics is rooted in the same mathematics
- This means that how you write your R code is pretty much the same for many common analyses
- Ryan Gosling stats memes always work

Practical Outline

Using R and R studio and a practice data set we will:

1. Explore basic descriptive statistics of our data
2. Examine basic inferential statistics of our data
3. Visualize both descriptive and inferential statistics