ISE 599
Special Topics applied predictive Analytics

Mayank Kejriwal

Research Assistant Professor/Research Lead

Department of Industrial and Systems Engineering

Information Sciences Institute

USC Viterbi School of Engineering

kejriwal@isi.edu

Analytics: some broad views

- From Oxford Dictionary:
  - the **systematic** computational analysis of **data or statistics**
  - information resulting from the **systematic** analysis of **data or statistics**
- From Techopedia:
  - Analytics is the **scientific process** of discovering and **communicating** the meaningful patterns which can be found in data. It is concerned with **turning raw data into insight for making better decisions**. Analytics relies on the application of statistics, computer programming, and operations research in order to quantify and gain insight to the meanings of data. It is especially useful in areas which record a lot of data or information.
- From 'Innovating with Analytics', MIT Sloan Management Review:
  - "…there is a strong correlation between driving competitive advantage and innovation with analytics and a company's effectiveness at managing the **information transformation cycle**, that is: **capturing data, analyzing information, aggregating and integrating data, using insights to guide future strategy and disseminating information and insights**."

Analytics is an emerging power

- Analytics ‘Have’ and ‘Have-nots’ exist at the levels of departments, companies and personnel
  - Almost every modern **industry** has some company investing in analytics

At MillerCoors, analytics is a central part of a corporate program to change the way the organization conducts business.

Analytics is inherently applied

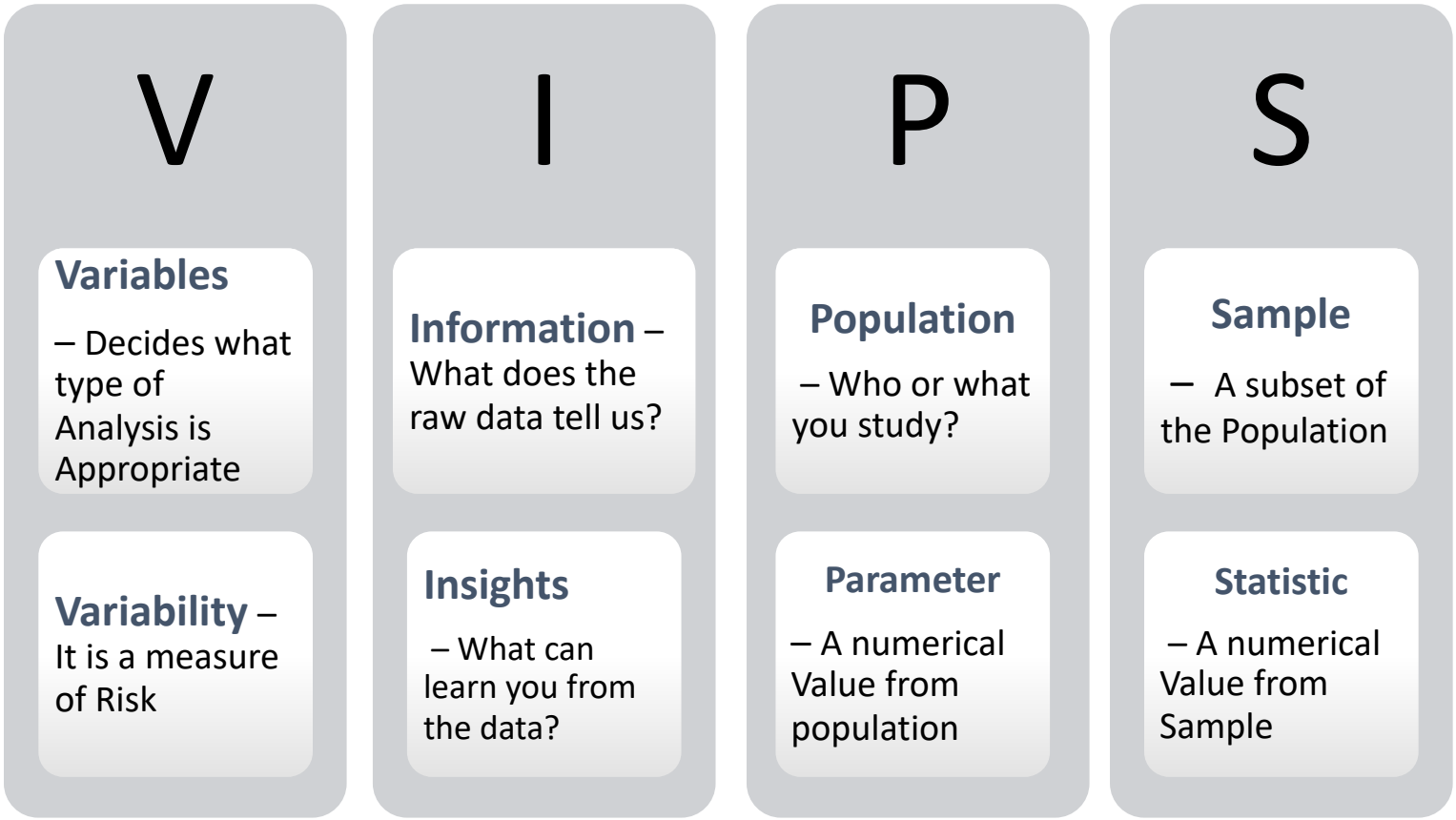
- In industry, analytics forces you to ask who your customer is (which in turn forces you to ask what the *problem* is)
  - Inevitably, you must be able to connect the dots between the customer for your 'analytical product' and the actual customer

Former Apple CEO Steve Jobs, *"When you start looking at a problem and it seems really simple …, you don't really understand the complexity of the problem. And your solutions are way too oversimplified. Then you get into the problem, and you see it's really complicated. And you come up with all these convoluted solutions … That's where most people stop. The really great person will keep on going and find … the* ***key, underlying principle of the problem*** *and come up with a* ***beautiful elegant solution*** *that works."*

How is this course organized?

statistics

I sincerely acknowledge *my* probability/statistics professor(s) for teaching me much of these


V
Variables
– Decides what type of Analysis is Appropriate
Variability – It is a measure of Risk
I
Information – What does the raw data tell us?
Insights
– What can learn you from the data?
P
Population
– Who or what you study?
Parameter
– A numerical Value from population
S
Sample
– A subset of the Population
Statistic
– A numerical Value from Sample

Lesson 1: Statistics (+visualizations) are the first line of attack in most applied analytics pipelines

Basic Statistical Concepts

- Population: The *total* collection of elements of interest in a given study.
  - All students in USC
  - all possible employees of a company

- Variable: A measurable characteristic that takes on different values.

- Parameter: A number that provides a quantitative description of some characteristic of a *population*. Generally, the average or proportion
  - proportion of people who rent
  - average GPA

- Sample: A subset of a population.
  - 1,000 individuals in a health survey
  - 5 students selected at random

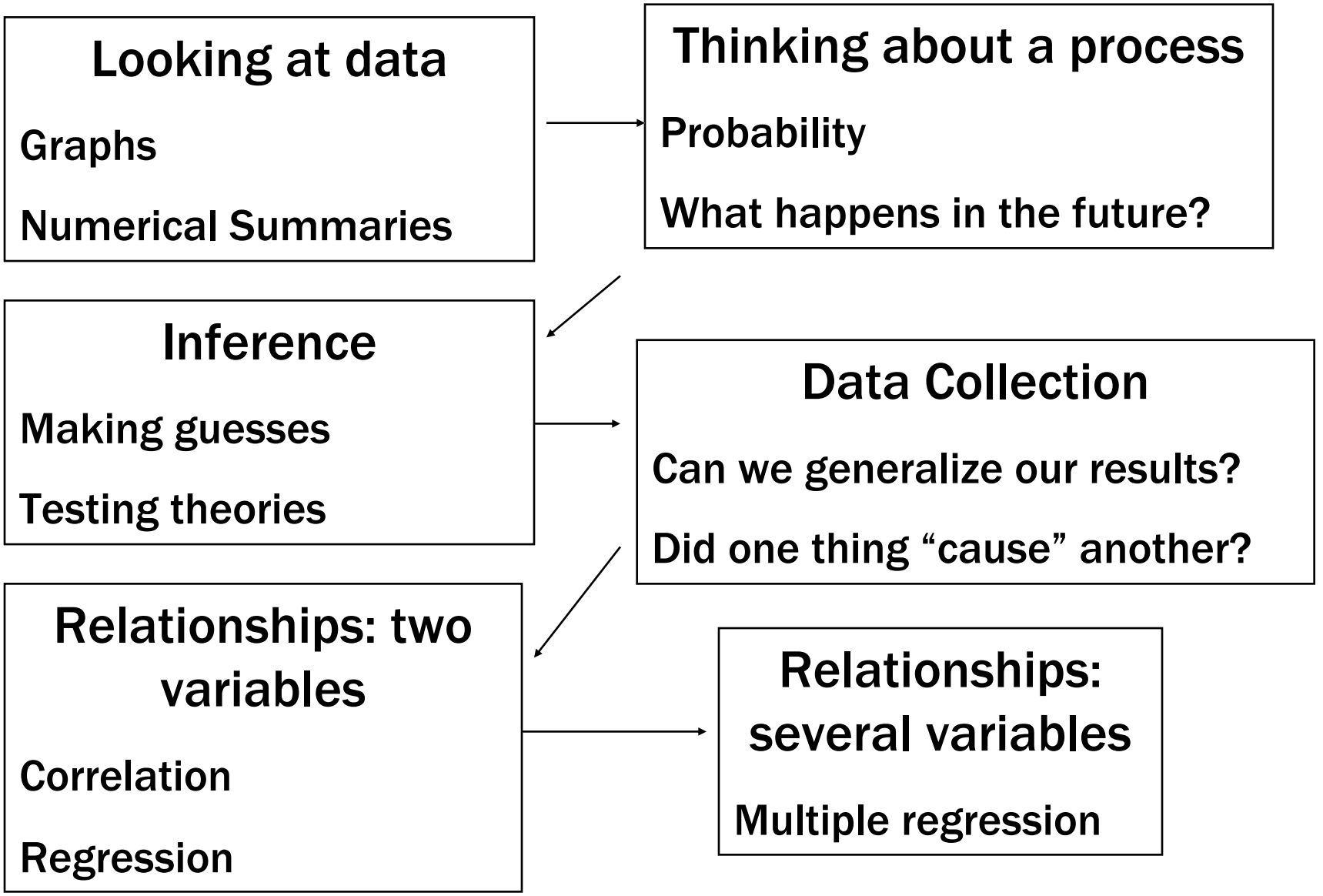
- Statistic: A numerical quantity that provides information concerning some characteristic of a *sample*. Generally, the average or proportion
  - proportion of home-owners who are seeking assistance for health reasons in the sample of 1,000
  - average age of the students in the sample of 5

- Observation: A recorded value of a variable for a particular individual.
  - this guy is 5’ 9” tall
  - this person is a Republican

- Data: A set of observations (i.e. the numerical values recorded for all individuals in a sample).

- Variable: A measurable characteristic that takes on different values.
  - Quantitative variable: differs in amount (e.g. yearly pay)
  - Qualitative variable: differs in kind (e.g. party affiliation)

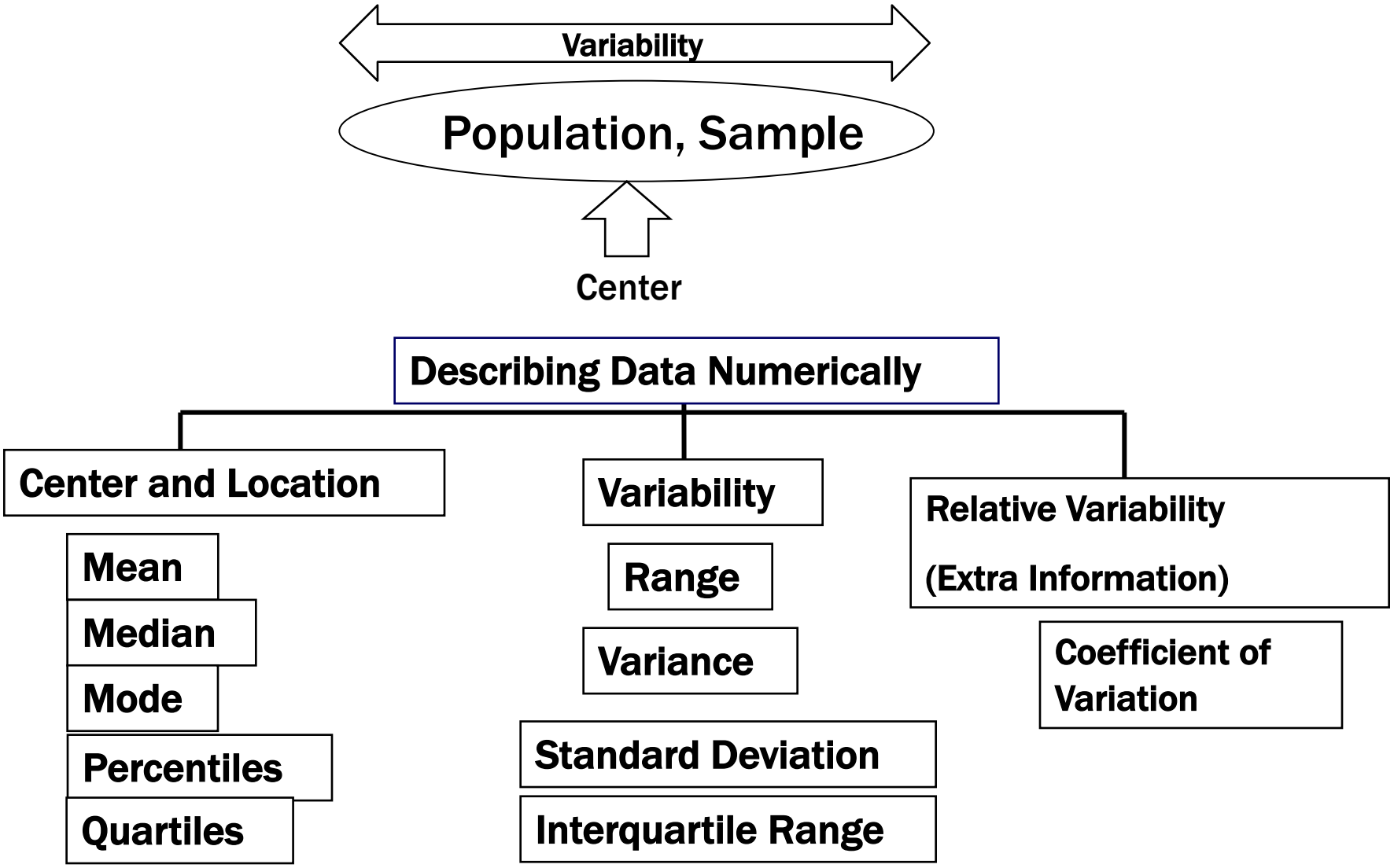
- Statistical inference: The use of *sample* information to infer something about *population*.

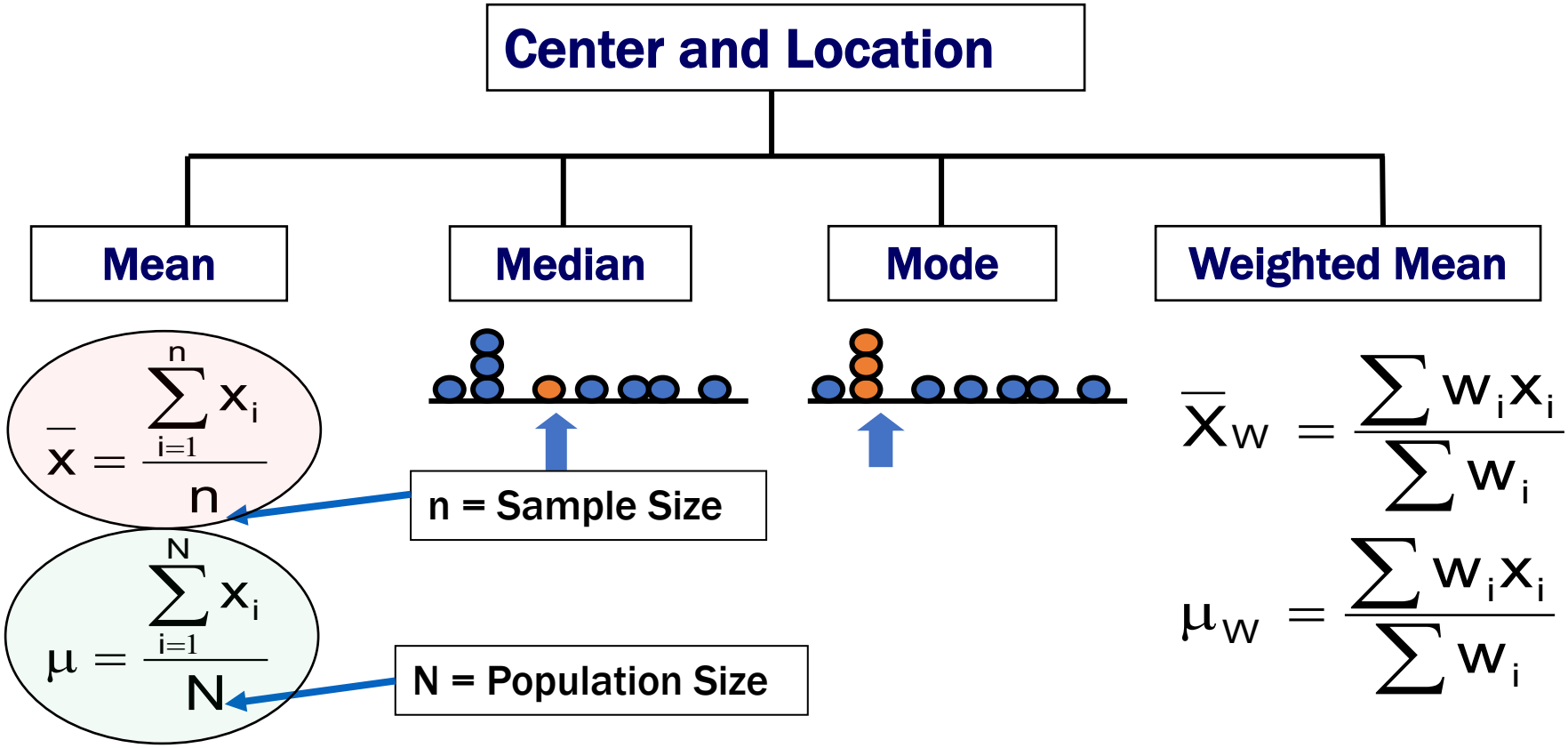

Looking at data
Graphs
Numerical Summaries
Thinking about a process
Probability
What happens in the future?
Inference
Making guesses
Testing theories
Data Collection
Can we generalize our results?
Did one thing "cause" another?
Relationships: two variables
Correlation
Regression
Relationships: several variables
Multiple regression

Data Types

- Nominal : simply classification.
  - *party affiliation*
  - *brand preference*

- Ordinal : values are ordered (large numbers indicate greater amounts).
  - *rate this coffee (1 = "Great",…..,5 = "Awful")*
  - *on a scale of 1 to 10, what do you think of President Reagan?*


Variability
Population, Sample
Center
Describing Data Numerically
Center and Location
Mean
Median
Mode
Percentiles
Quartiles
Variability
Range
Variance
Standard Deviation
Interquartile Range
Relative Variability
(Extra Information)
Coefficient of Variation


Center and Location
Mean
Median
Mode
Weighted Mean
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
n = Sample Size
N = Population Size
$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$
$\mu_w = \frac{\sum w_i x_i}{\sum w_i}$

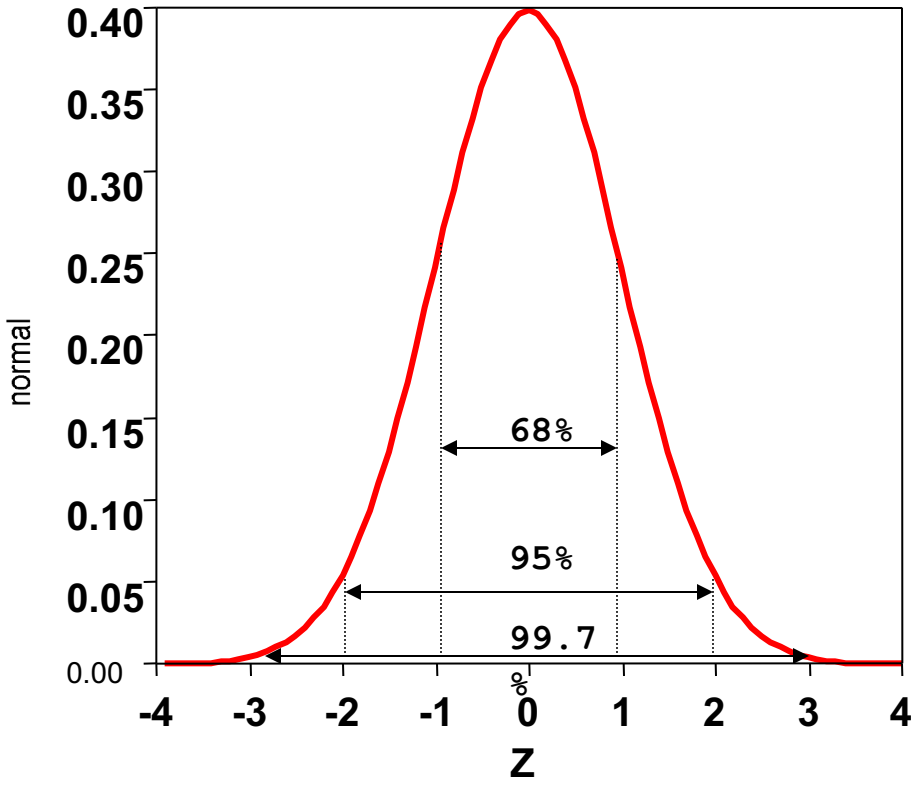
Empirical rule

If the normal curve fits well then (approximately):

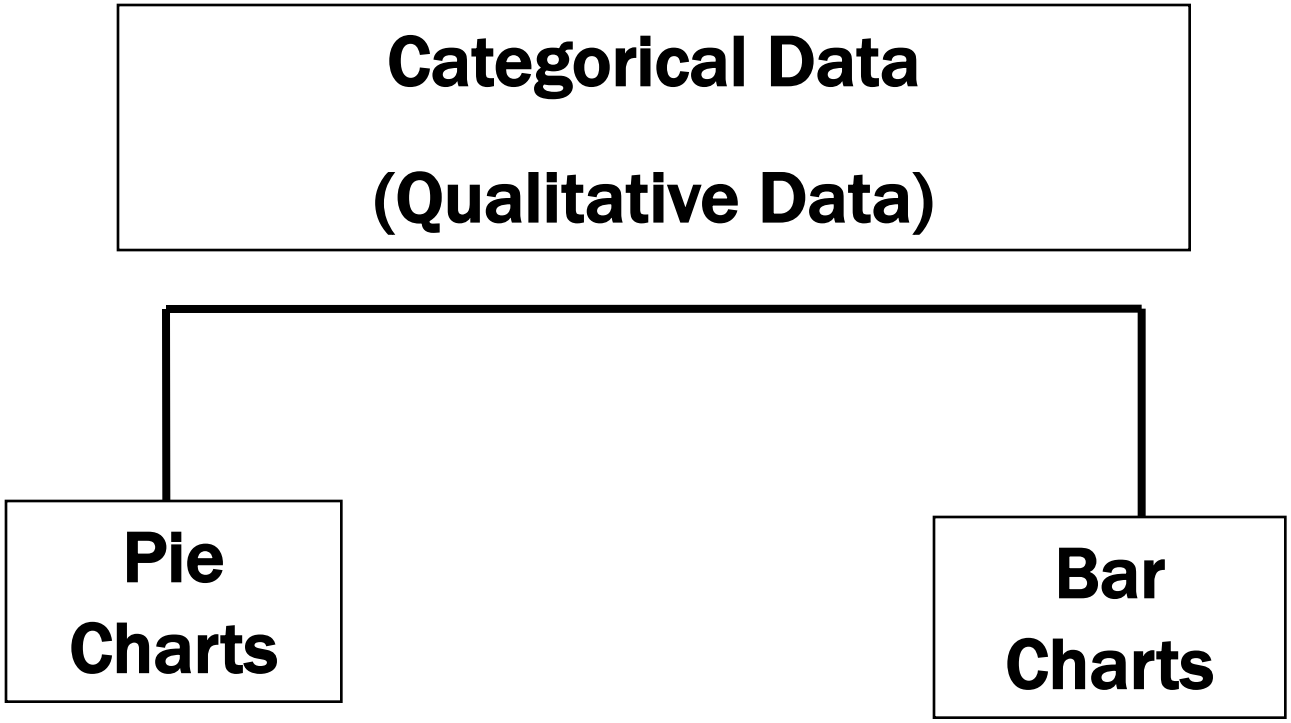
68% of the data is within 1 SD (Standard Deviation) of the mean.

95% within 2 SD.

99.7% within 3 SD.

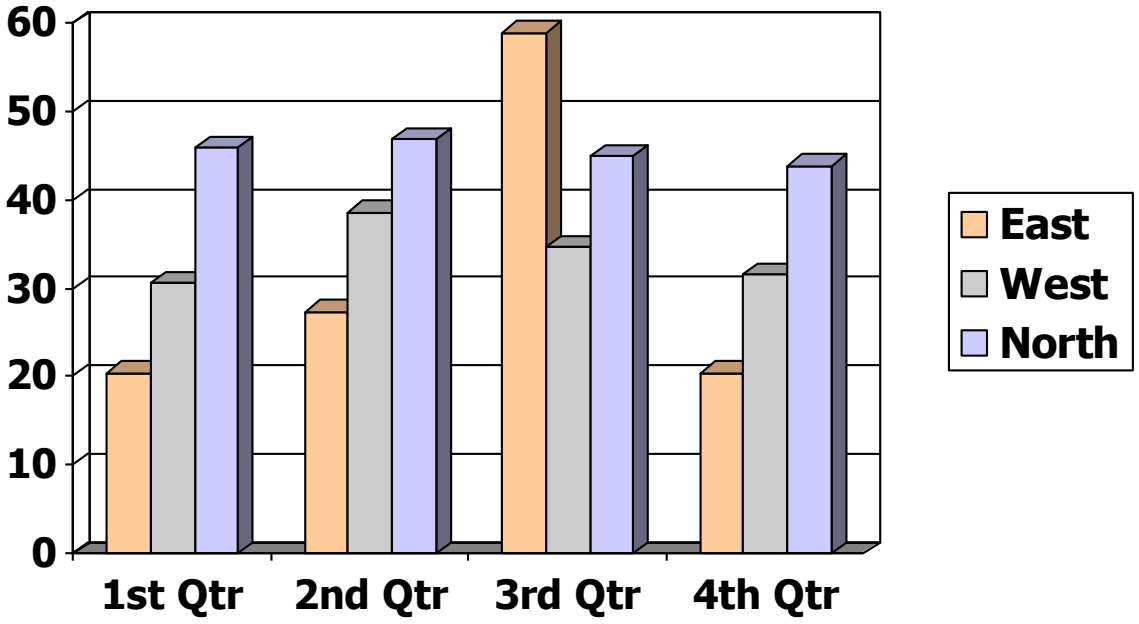


Graphing Categorical Data

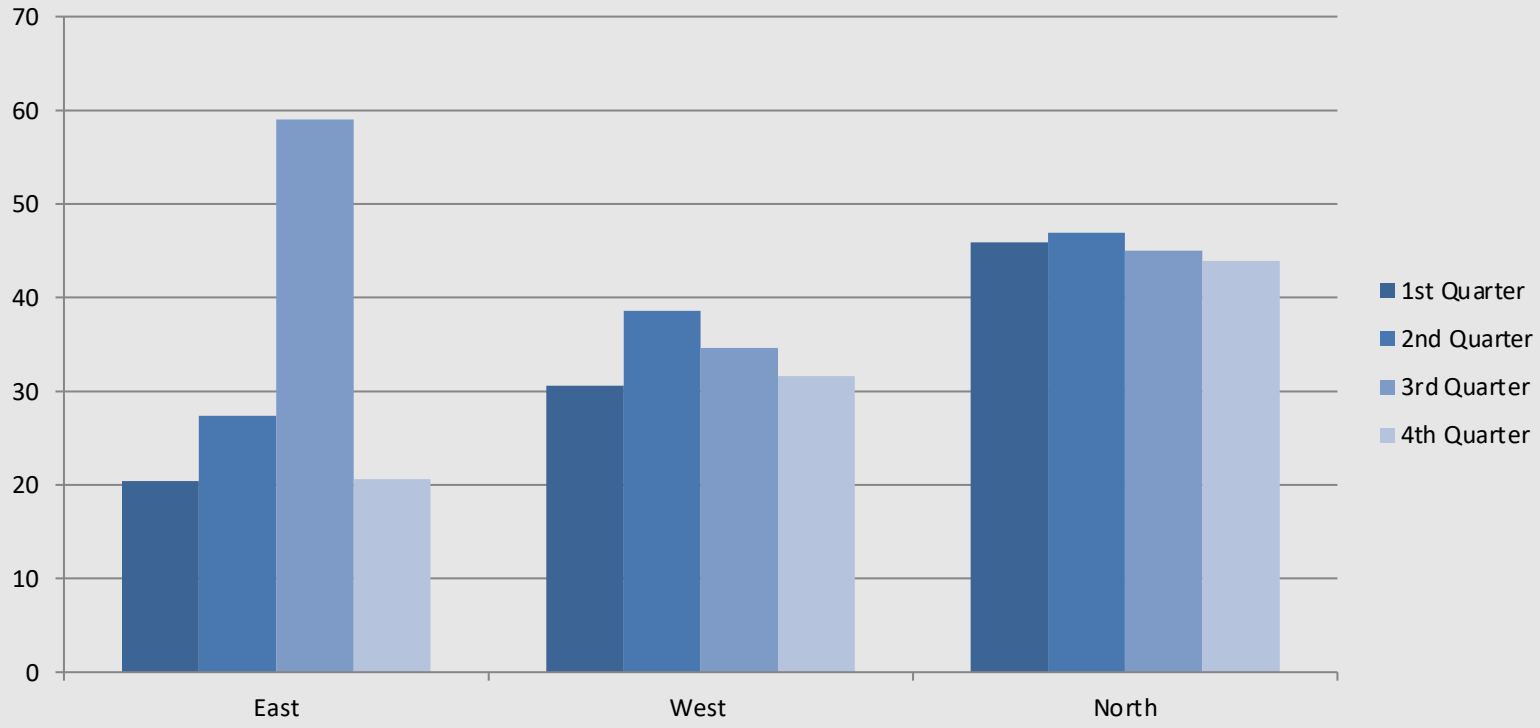


What if we have two categorical variables?

	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
East	20.4	27.4	59	20.6
West	30.6	38.6	34.6	31.6
North	45.9	46.9	45	43.9



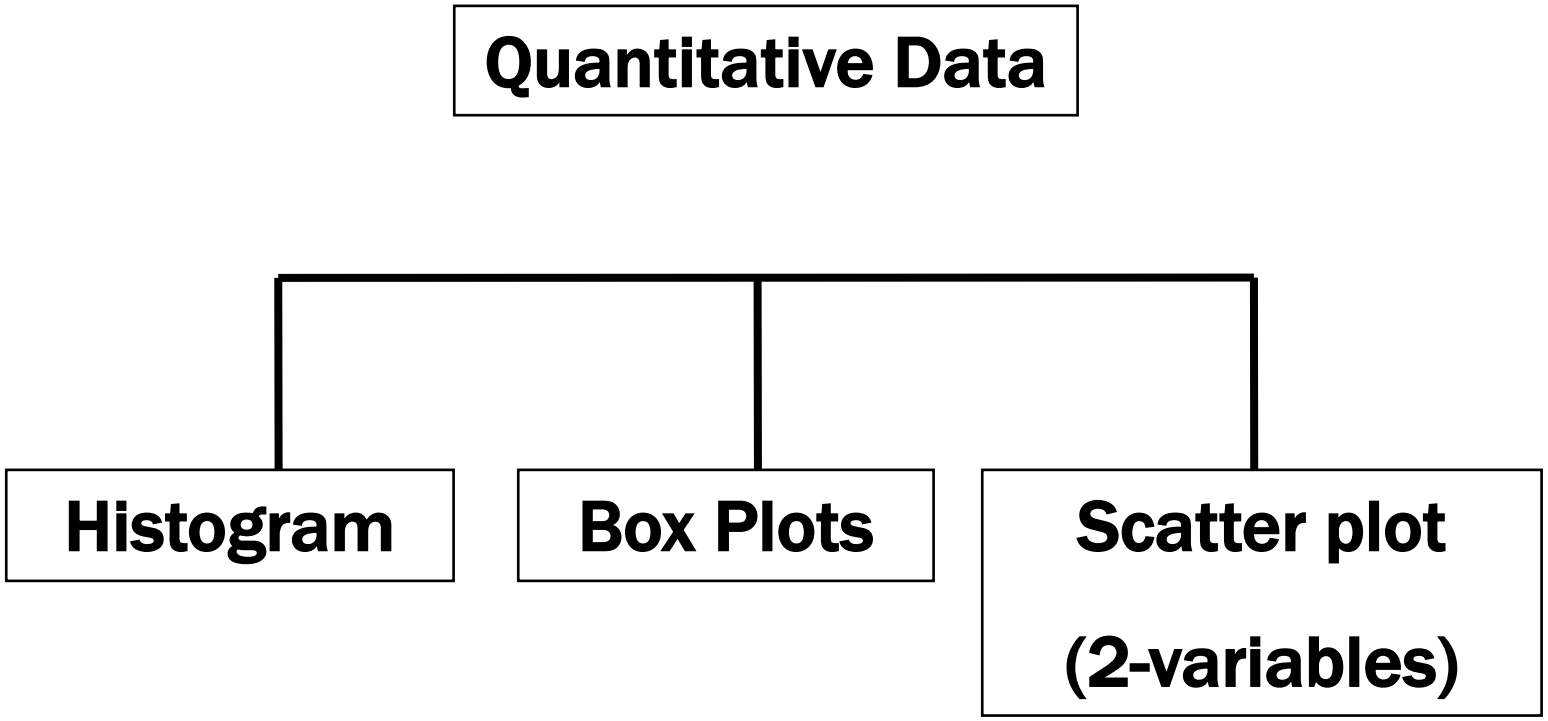
Alternate view (which is better?)



Can you identify if (and why) the following situation might be misleading?

- A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective

Graphing Quantitative Data



Confounds

- Confounding occurs when the experimental controls do not allow the experimenter to **reasonably eliminate plausible alternative explanations** for an observed relationship between independent and dependent variables.

Can you identify if (and why) the following situation might be misleading?

- The more churches in a city, the more crime there is. Thus, churches lead to crime.

Correlation vs. Causation

- The use of a controlled study is the most effective way of establishing causality between variables. In a controlled study, the sample or population is split in two, with both groups being comparable in almost every way. The two groups then receive different treatments, and the outcomes of each group are assessed.

Can you identify if (and why) the following situation might be misleading?

- 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages.

When making comparisons…

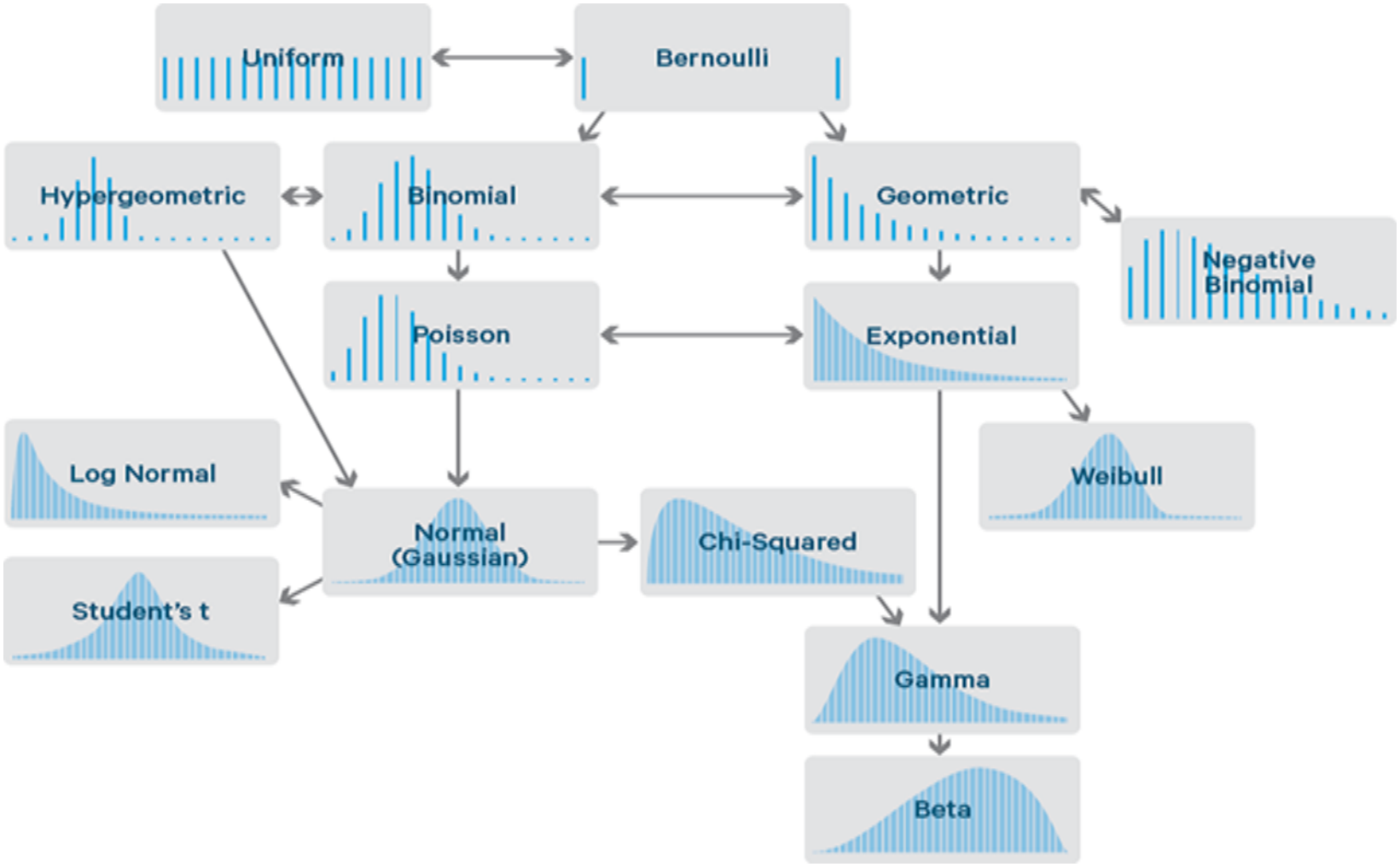
- …Make sure to choose a suitable frame of reference!
- Single biggest cause of misleading (sometimes intentional) statistics, especially in business and science

Lesson 1: Statistics (+visualizations) are the first line of attack in most applied analytics pipelines

“There are three kinds of lies – lies, damned lies, and statistics”

- To be an intelligent consumer of statistics, your first reflex must be to question the statistics you encounter
  - The more statistical knowledge and experience you have, the more adept you will be at doing the questioning
  - Like everything in life, skepticism should also be exercised in moderation
    - [Do not be like] Metrodorus of Chios: “None of us knows anything, not even this, whether we know or we do not know; nor do we know what ‘to not know’ or ‘to know’ are, nor on the whole, whether anything is or is not”

What about probability?



Statistical Tests

Null and Alternative Hypotheses

- Statistical hypothesis: claim about a parameter of a population.
- Null hypothesis (H_0): specifies a default course of action, preserves the **status quo.**
- Alternative hypothesis (H_a): contradicts the assertion of the null hypothesis,

Ha: It is the Research Hypothesis →

- **What you want to test**
- **The question you want to investigate**
- **The statement for which you are collecting evidence**

Example problem

The manager of a health maintenance organization has set as a target that the mean waiting time of non-emergency patients **will not exceed** 30 minutes (**Status Quo**). In spot checks, the manager finds the waiting times of 36 patients; the patients are selected randomly on different days. Assume that the **population standard deviation of waiting times is 10 minutes, the sample mean is 35 minutes (What is given? Status Quo or Research Hypothesis?)**

a. What is the relevant parameter to be tested?

μ = mean waiting time of ALL non-emergency patients

b. Formulate null and research hypotheses.

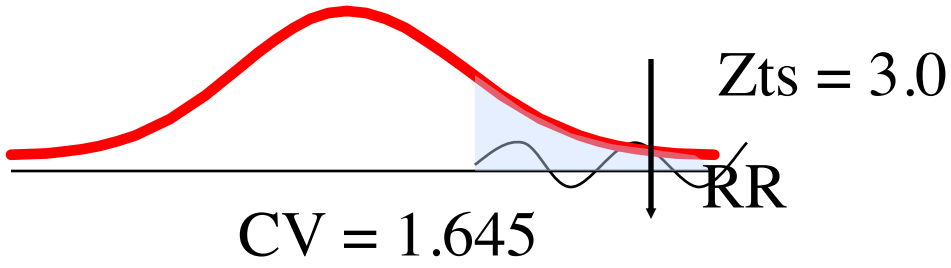
Cont'd

c. State the test statistic and the rejection region corresponding to $\alpha = .05$.

$$\text{Test Statistic}: \qquad Z_{ts} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$Z_{ts} = \frac{35 - 30}{10 / \sqrt{36}} = \frac{5 * \sqrt{36}}{10} = 3.0$$

$$\text{Rejection Region}: \qquad \text{Reject } H_0 \text{ if } Z > 1.645$$



Takeaways

- Think about statistics like scales in music: no one really ‘enjoys’ it, but without mastering it, you cannot be a good musician
- In the classroom, we state the problem in an obvious way but in the real world modeling the problem is half the battle!