Lecture 2: Text Analytics

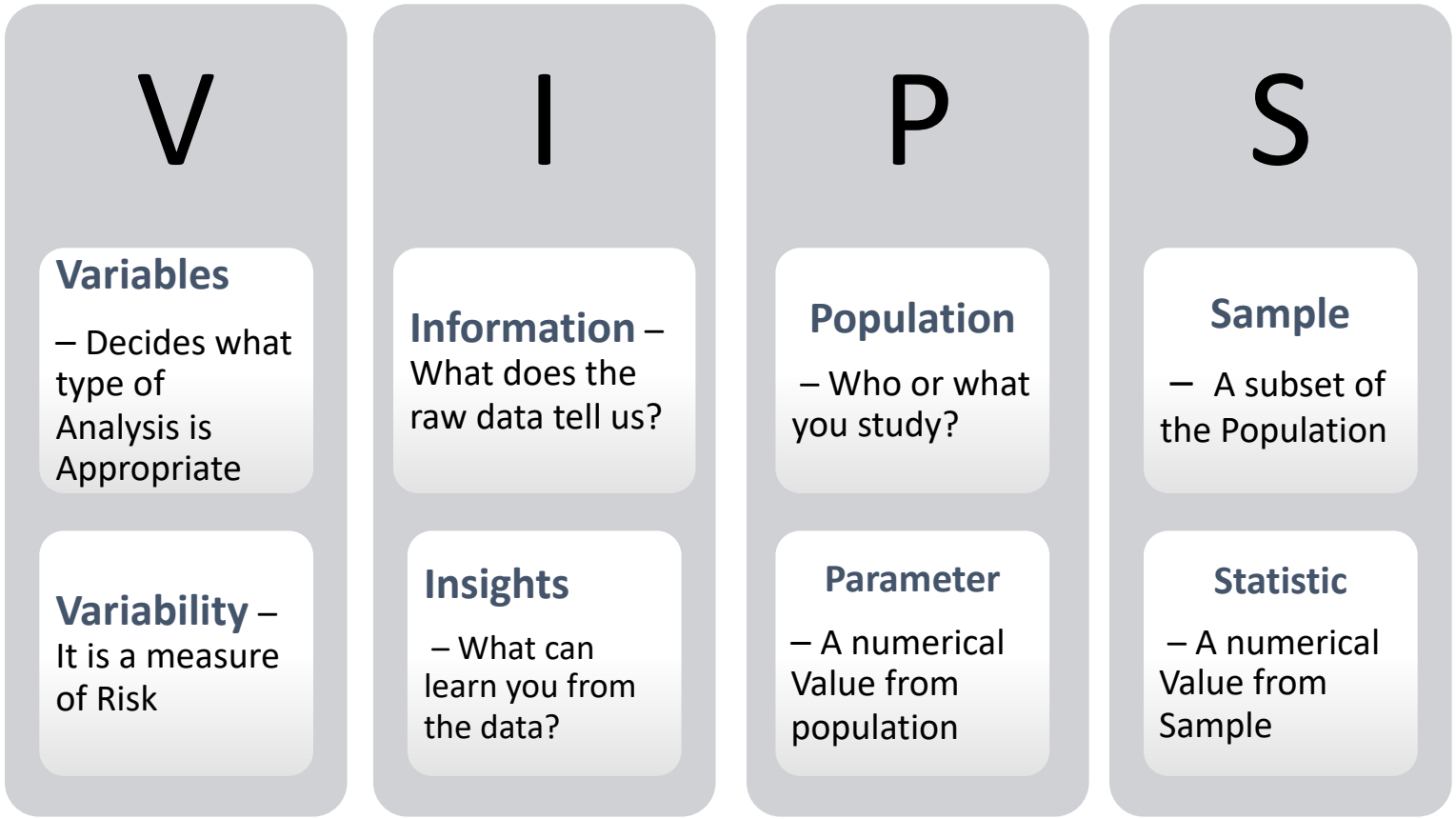
Dr. Mayank Kejriwal

Statistics: Background and preliminaries

Why statistics?

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

However, this is not a statistics course, any more than a course on machine learning (or even this course) is a ‘programming’ course. You will be expected to draw on (and re-learn) statistics as needed, for completing your assignments


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

Question from students: What else must I know?

- Difference between null and alternative hypothesis (and how to set them up given a situation)
- Statistical hypothesis testing (be sure to brush up on Student's t tests and Z test)
- What don't I need to know (or remember)?
  - Too many formulas (I'll allow you to look them up usually, but you should have a few at the tips of your fingers, such as the Normal Distribution)

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.

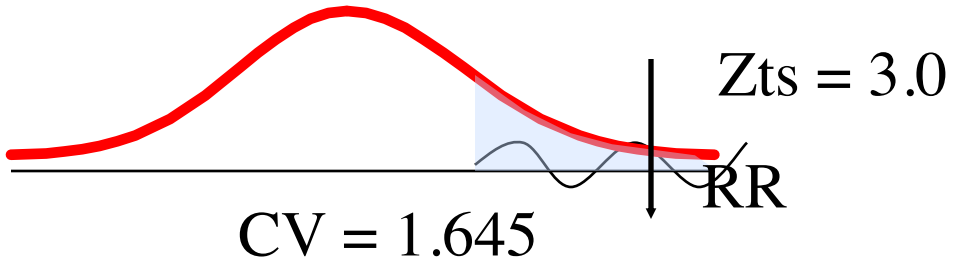
Cont'd

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

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

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

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



Statistics in real life

Uber

Overview of data generation, modeling and interpretation in statistical perspectives

Randomized Experiments
(With randomization over treatment and control groups)

Classic Experiments (Non-recurring)

- Univariate Tests
- A/B Tests

Continuous Experiments (Recurring)

- Statistical Techniques
  - Allocation %
    Eg. Thompson Sampling
  - Rollout %
    Eg. Power Based via Sequential Tests
    Eg. Risk Based
- Model-Based Techniques
  - Contextual MAB
  - Bayesian Optimization

Observational Studies
(Pure observation with no randomization)

Different Methods To Estimate Associated Lifts

- Synthetic Control
- A/B-like:
  How to construct a weighted "control" group
- Regression

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

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.

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?

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

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.

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

Related: Spurious Correlation

- Spurious Correlation, or spuriousness, is when two factors **appear** casually related but are not.
  - This is subtle…spurious correlations are **actual (statistical)** correlations but have no business being correlated
- Spurious correlations tend to occur because confounds haven't been eliminated
  - However, confounds are more general, while spurious correlations, as the name suggests, specifically arises with correlation
- Spurious Correlation can often be caused by small sample sizes or arbitrary endpoints.

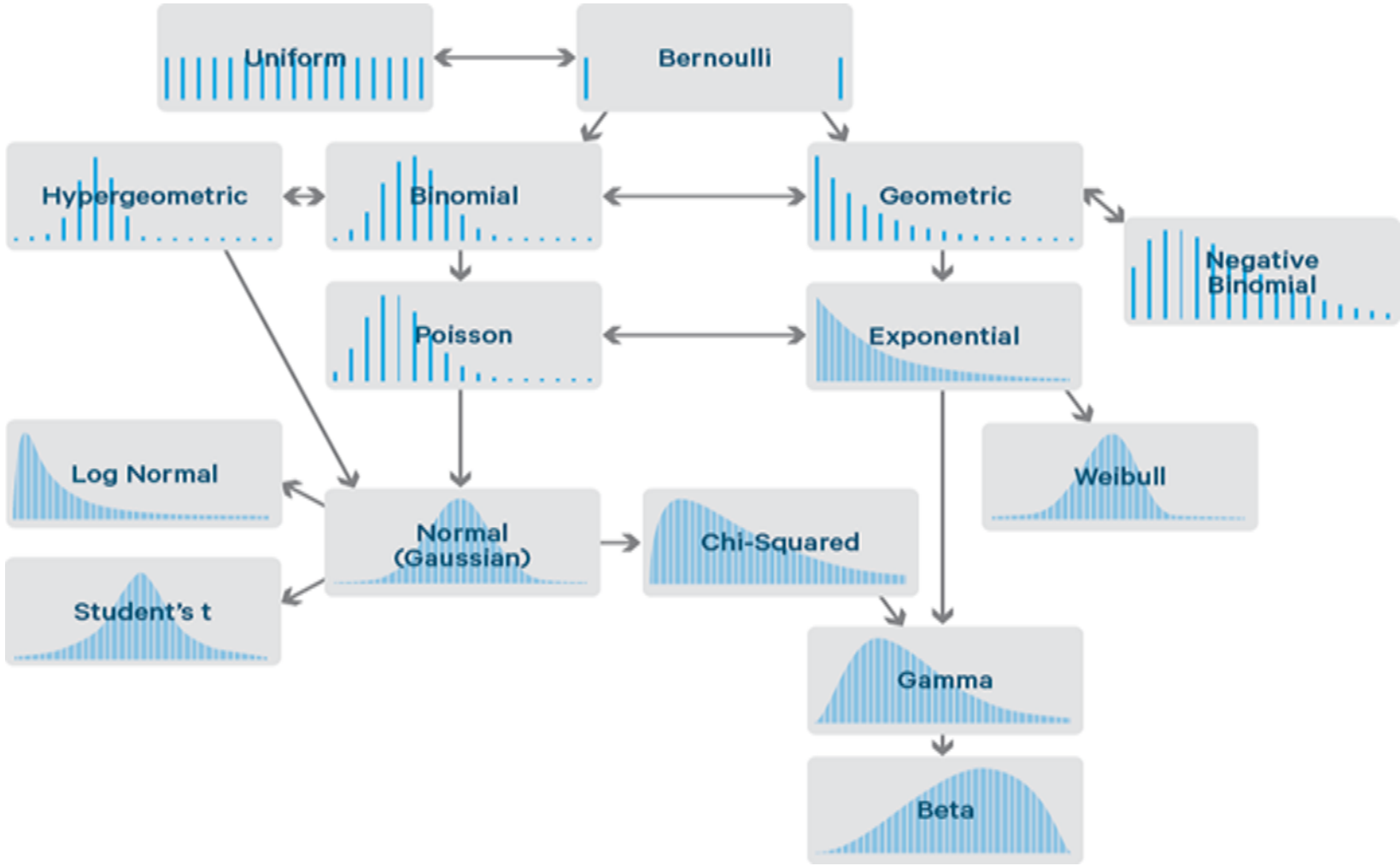
Some examples

- For the ‘male species on Wall Street’ (taken from Investopedia):
  - The skirt length theory in the 1920s
  - ‘Hemline index’ first suggested in 1925 by George Taylor of the Wharton School of Business
  - Skirt hemlines are higher when the economy is performing better, and longer during downturns
  - Accurate in 1987 (designers switched from miniskirts to floor-length skirts just before the market crashed) and also 1929
- Other ‘unconventional’ (are all spurious?) economic indicators include men’s underwear, haircuts, dry-cleaning and fast food

“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?



Takeaways

- Think about probability & 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!