A Deep Analysis of Data with Power BI, Excel and Statistics

DR. ERNESTO LEE

In God we trust… all others, bring data…

- ANONYMOUS

Learning Path

- Why is Central Limit Theorem Central to everything that we do?
- Univariate Analysis
  - Count Plots / Histograms / Stacked Column Charts
- Bivariate Analysis
  - Numeric to Numeric
    - Pearson Correlation / Pearson R / R2 / Scatter Plot
  - Categorical to Numeric
    - T-test / p-value / ANOVA (one way) / Bar Chart
  - Categorical to Categorical
    - Cross Tab / Heat Map / Chi-square test

Central Limit Theorem

- Has 2 important parts:
  - Random samples of a population fall towards a mean
  - Sample values of a sufficient size can provide insights about the larger population
- Is important for 2 Reasons:
  - Hypothesis Testing
  - Inferences

The central limit theorem states that if we take repeated random samples from a population and calculate the mean value of each sample, then the distribution of the sample means will be approximately normally distributed, even if the population the samples came from is not normal.

- NATURE

The central limit theorem also states that the mean of the sampling distribution will be equal to the mean of the population distribution:

$$x = \mu$$

- NATURE

Demo

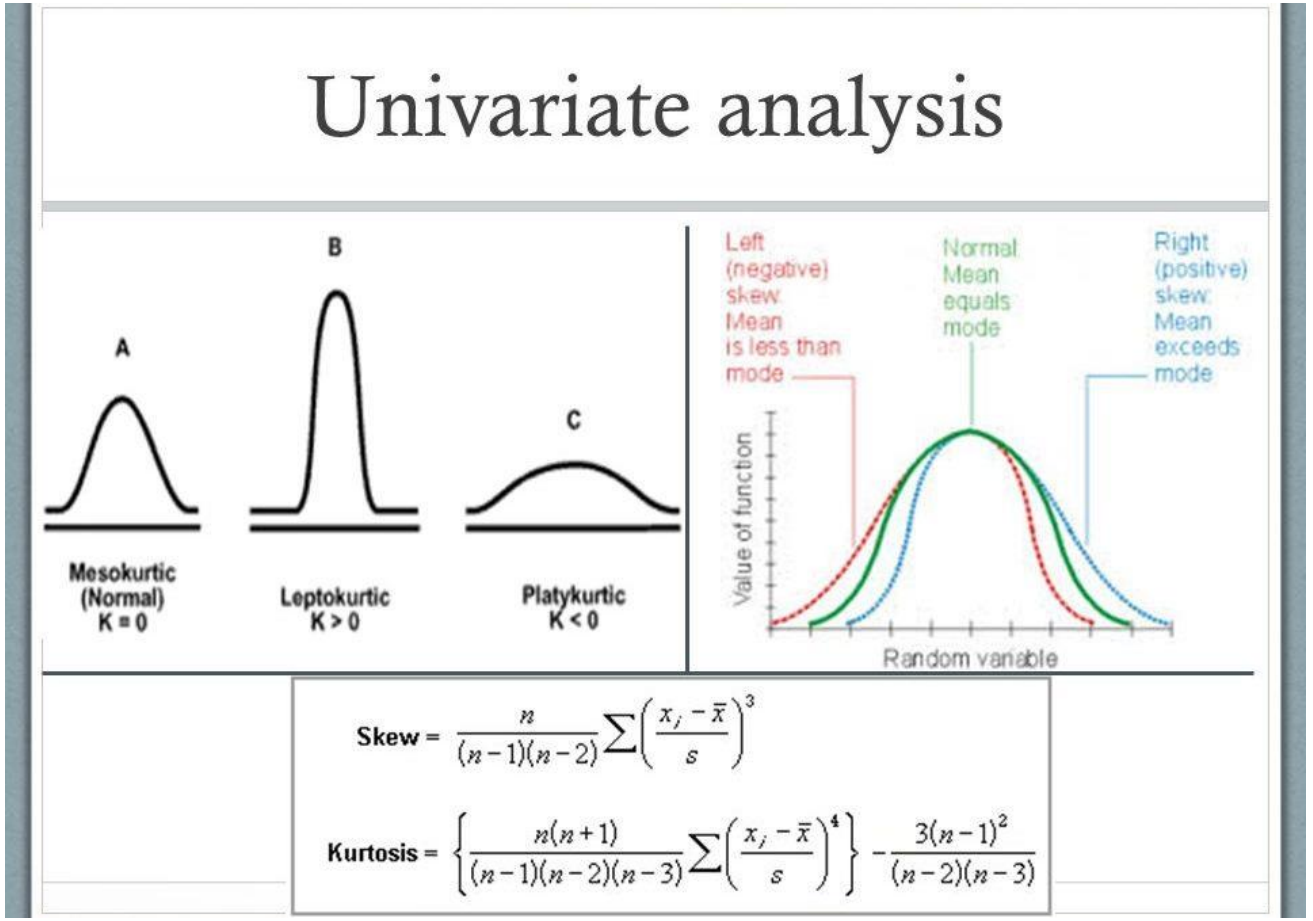
Univariate Analysis

UNIVARIATE ANALYSIS IS PERHAPS THE SIMPLEST FORM OF STATISTICAL ANALYSIS. LIKE OTHER FORMS OF STATISTICS, IT CAN BE INFERENTIAL OR DESCRIPTIVE. THE KEY FACT IS THAT ONLY ONE VARIABLE IS INVOLVED.

LOVE WIKIPEDIA

Univariate Analysis with Power BI

- Identify all of your features as quantitative or qualitative
  - Numeric (break down into Ordinal or Continuous)
  - Categorical (break down into Binary, Ordinal or Categorical)
- Utilize **Descriptive Statistics** in Power BI to pull insights from a single column of data.
- A stacked column chart will be very useful.
- Learn to bin continuous data to make it manageable.
- Remember the objective is to pull INSIGHTS from your columns



Demo

P-values

- What is the null hypothesis?
- What exactly is a p-value?
- Statistical significance of p-values

Example: Null and alternative hypothesis

You want to know whether there is a difference in longevity between two groups of mice fed on different diets, diet A and diet B. You can statistically test the difference between these two diets using a two-tailed t test.

Null hypothesis: there is no difference in longevity between the two groups.

Alternative hypothesis: there is a difference in longevity between the two groups.

What is the Bottom Line Up Front?

- The p-value will tell you if there is a statistically significant relationship between two features in your dataset. This relationship is your insight.
- The magic threshold is 0.05.
  - If **p > 0.05** then there is **NO** difference between your 2 columns of data. There is nothing statistically significant. There is no relationship between the columns (like eye color and weight).
  - If **p < 0.05** then there **IS** a statistically significant difference between your 2 columns of data (like smoker and insurance rate OR quality of water and longevity.
  - BLUF: find the values of p and you can find deep insights in your dataset.

Bivariate Analysis

BIVARIATE ANALYSIS IS ONE OF THE SIMPLEST FORMS OF QUANTITATIVE ANALYSIS. IT INVOLVES THE ANALYSIS OF TWO VARIABLES, FOR THE PURPOSE OF DETERMINING THE EMPIRICAL RELATIONSHIP BETWEEN THEM. BIVARIATE ANALYSIS CAN BE HELPFUL IN TESTING SIMPLE HYPOTHESES OF ASSOCIATION.

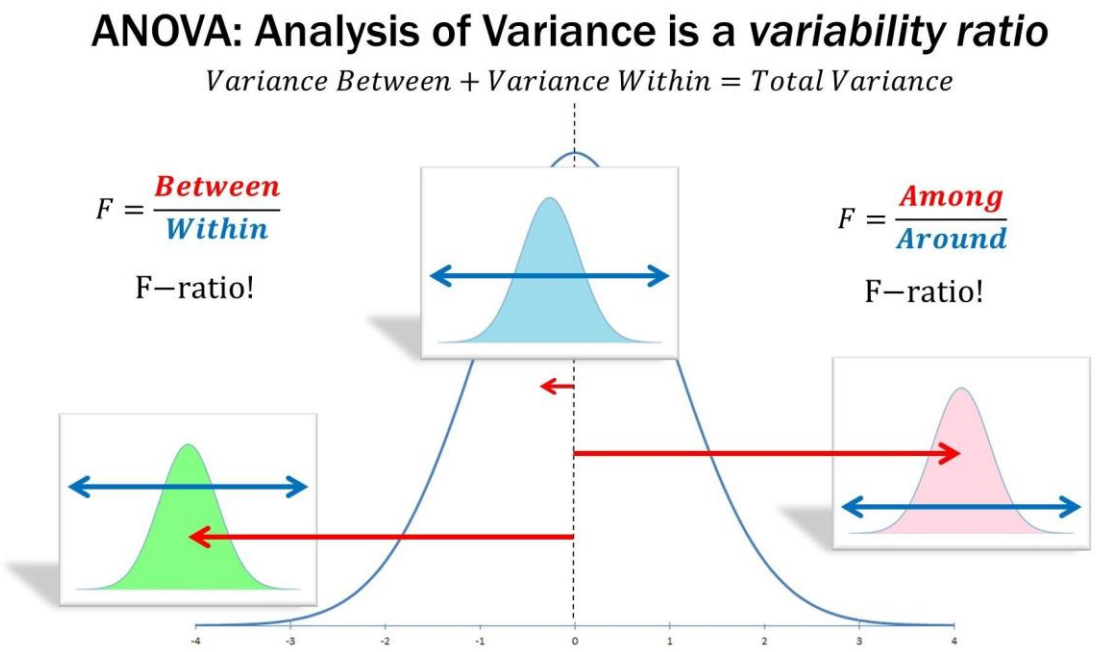
LOVE WIKIPEDIA

Bivariate Analysis with Power BI (N2N)

- Use a Scatter Plot in Power BI to visualize the relationships
- Use Person R (-1 to 1)
  - The degree of correlation
  - The sign gives you the direction
- Use R2 (0 to 1)
  - The explainability of one variables effect on another.

Demo

Bivariate Analysis with Power BI (C2N)



- Use a Bar Chart in Power BI to visualize the relationships
- Use ANOVA
  - Analysis of Variance
  - Pulls the p-value from a collection of features
- Use t-test
  - Pulls the p-value from only 2 features

Demo

Bivariate Analysis with Power BI (C2C)

- Create a matrix report in Power BI (Cross Tab report)
- Can be extended to multivariate analysis

Demo

Three types of Analysis

Three types of analysis

- Univariate analysis
  - the examination of the distribution of cases on only one variable at a time (e.g., weight of college students)
- Bivariate analysis
  - the examination of two variables simultaneously (e.g., the relation between gender and weight of college students)
- Multivariate analysis
  - the examination of more than two variables simultaneously (e.g., the relationship between gender, race and weight of college students)