

A decorative graphic in the top-left corner consisting of a grid of colored squares. The squares are arranged in a pattern that tapers to the right. The colors include light purple, medium purple, teal, and dark teal. The squares are of varying sizes and are arranged in a way that creates a sense of depth and movement.

Quantitative Data Analysis

Neba Nfonsang



Quantitative Data Analysis Outline

- Introduction
- Descriptive Statistics
 - Measure of Central Tendency (Mean, Median, and Mode)
 - Measure of Variability (Range, Standard Deviation, Variance)
 - Measure of Shape (Kurtosis and Skewness)
- Inferential Statistics
 - Two Sample Independent T-Test
 - Analysis of Variance (ANOVA).

Components of a Dataset

An observation

Variables

A dataset

	Height	Weight	City
1	65	112	Denver
2	71	136	Aurora
3	69	153	Boulder
4	68	142	Denver
5	67	144	Denver
6	68	123	Boulder



Data Related Words

- A **variable** is a characteristic or attribute of an individual or object that takes on different values.
- Example of variables include age, ethnicity, city, happiness level, etc.
- **Data** are values or facts associated with a variable.
- **An observation** is data collected from an individual.
- **A dataset** is a collection of observations for several individuals or objects.

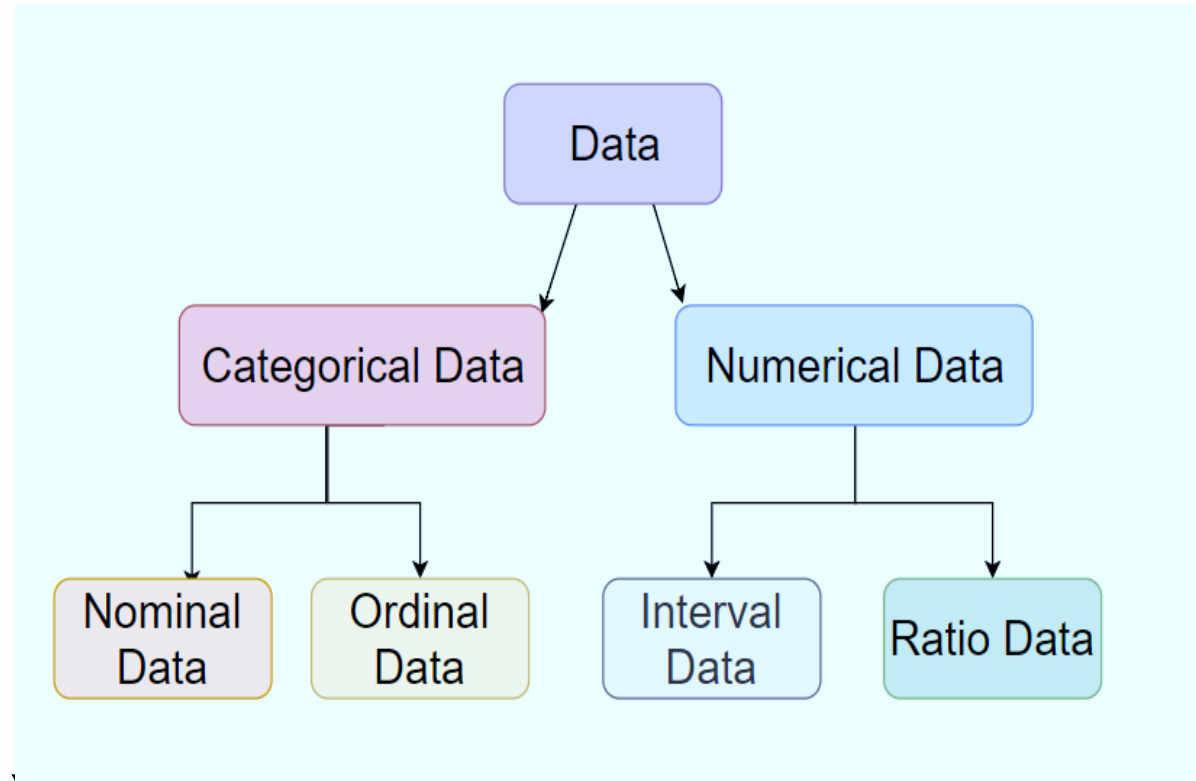


Types of Data

- Data can be broadly classified into **categorical data** or **numerical data**.
- Categorical data consist of names or labels such as “male” or “female” belonging to a categorical variable such as gender.
- Numerical data consist of numbers that can be processed mathematically or statistically.
- Numerical data can be added, subtracted, averaged, multiplied or even divided to get meaningful results.

Levels of Measurement (Data Types)


- Categorical data can be further split into:
 - nominal data (names or labels)
 - ordinal data (labels that can be ranked)
- Numerical data can be divided into:
 - Interval data (no meaningful zero)
 - Ratio data (has a no meaningful zero)





Data Types and Level of Measurement

Variable	Data type	Level of Measurement
City	Categorical	Nominal
Gender	Categorical	Nominal
Depression level on a 3-point Likert scale (low, moderate, high)	Categorical (sometimes treated a numerical if range is larger)	Ordinal (sometimes considered interval data)
Time	Numerical	Interval
GPA	Numerical	Interval
Height	Numerical	Ratio
Math Score	Numerical	Ratio



Complete the Data Types and Measurement Levels

Variable	Data type	Level of Measurement
Social Security Number		
Nationality		
Family size		
Weight		
Education level (Bachelor's master's, PhD)		
Phone number		
Eye color		
IQ (intelligence quotient)		



Descriptive Statistics



Measure of Central Tendency

- Measure of central tendency involves measuring the center of the data distribution.
- Mean, mode and median are used to measure the center of the data distribution.
- Why do we care about the center of your distribution?
 - The center of the data distribution is a summary of the data.
 - If the data is normally distributed, the center represent the point around which most of the data lie.



Measure of Central Tendency

- **Mean** is the average value:
 - ☐ Calculated as sum of all data points divided by number of data points
- **Median** is the middle value of the sorted data.
- **Mode** is the value with the highest frequency in the data.

- Given the following data:

2, 3, 3, 3, 4, 5, 6

- What is the:
 - ☐ Mean?
 - ☐ Mode?
 - ☐ Median?

Measure of Variability

- Variability measures how spread the data is.
- Range, standard deviation and variance are common measures of variability.
- **Range** is the difference between the maximum and minimum value in the data.
- Variance (δ^2) measures on average, the deviation of the data points from the mean.
- $$\delta^2 = \frac{(x - \bar{x})^2}{n-1}$$
- Standard deviation is the square root of the variance (has the same unit as the data)

Measure of Variability

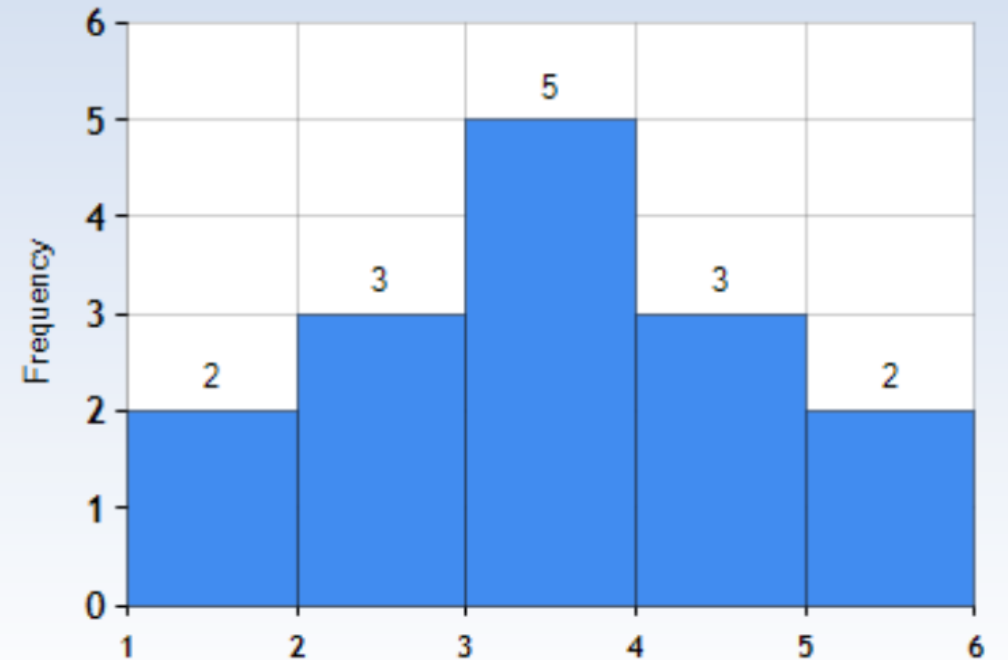
Data points	Deviation (data – mean)	Deviation squared (data – mean) ²
2	$2 - 4 = -2$	4
3	$3 - 4 = -1$	1
5	$5 - 4 = 1$	1
6	$6 - 4 = 2$	4
Total		10
Variance		$10/3 = 3.33$
Standard Dev		$\sqrt{3.33} = 1.83$

Mean = $(2 + 3 + 5 + 6)/4 = 16/4 = 4$
Degree of freedom = $n - 1 = 4 - 1 = 3$

n = number of data points = 4

Measure of Shape

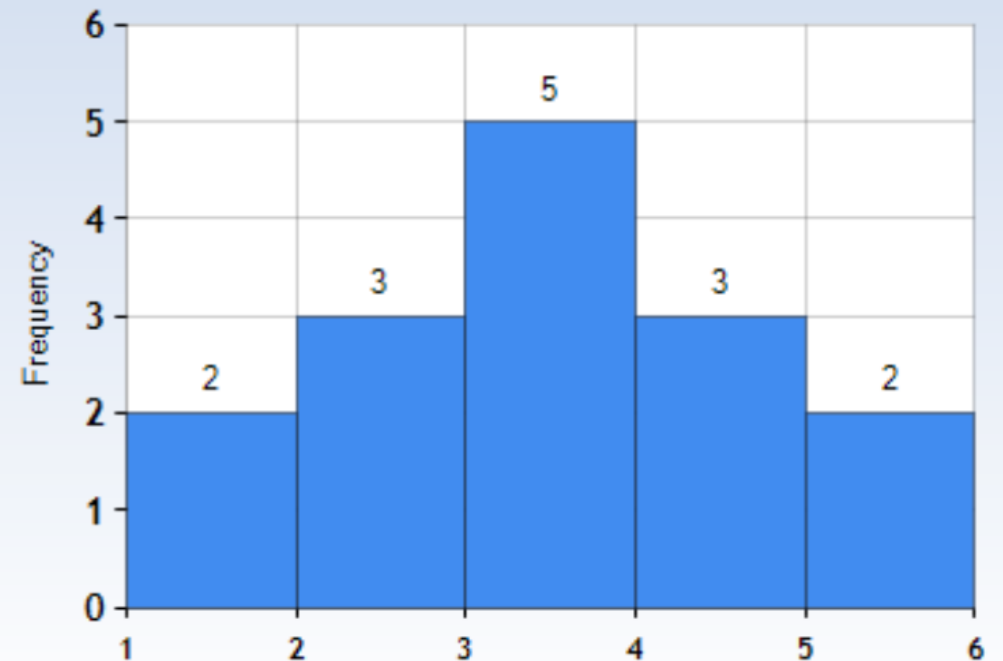
- The measure of shape describes the shape of the distribution of data.
- The distribution of data consist of all sorted data points and their frequencies.
- The distribution of data can be visualized using a histogram or a boxplot.



Measure of Shape

- Skewness and kurtosis are numerical measures of the shape of the data distribution.
- **Skewness** is the measure of symmetry of data distribution.
- **Kurtosis** measures the thickness of the tail of a distribution

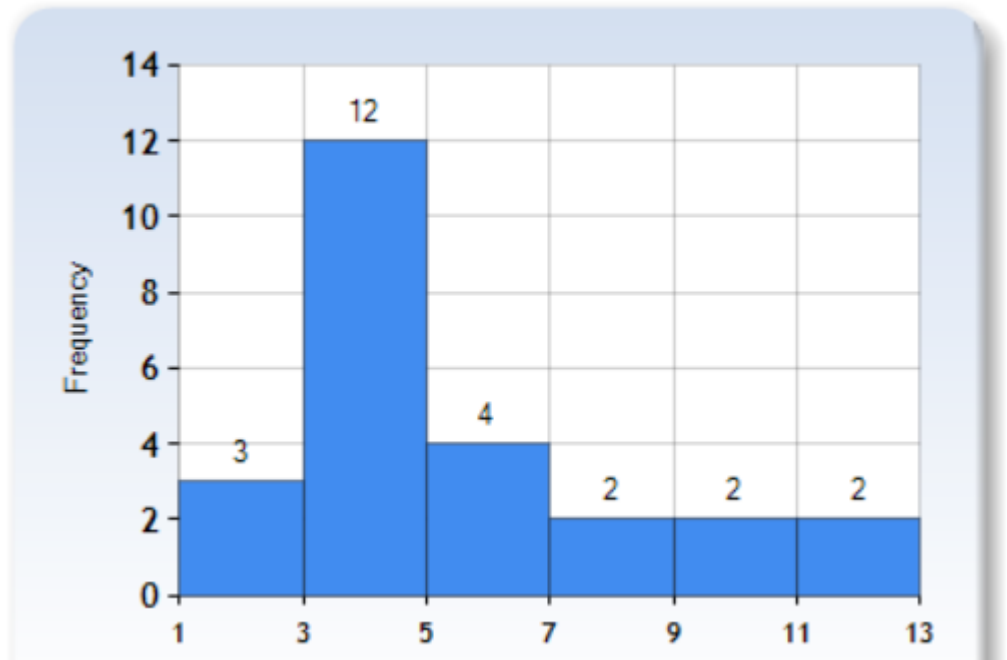
Is the distribution as shown in the histogram skewed or symmetric?



Measure of Shape

- A data distribution is said to be skewed to the left if the tail of the distribution is longer on the left.
- A data distribution is said to be skewed to the right if the tail of the distribution is longer on the right.

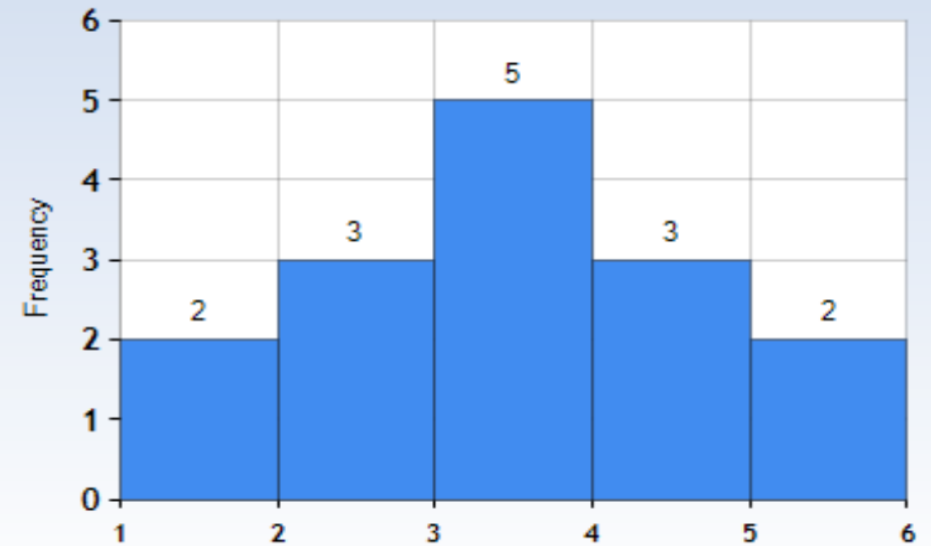
Is this distribution of data left-skewed or right-skewed?



Measure of Shape

- When the distribution of data is neither skewed to the left or to the right, the distribution is symmetric.
- A **normal distribution** is a symmetric distribution that looks like a bell-shape.

Does the data as shown in the histogram appear to be normally distributed?





Measure of Shape: normality

- Most inferential statistical test such as the t-test and analysis of variance require the data to be normally distributed.
- Visual inspection of a histogram or boxplot can be used to confirm the normality of data.
- Other plots such as qqplot or boxplot could also be used to establish normality.
- If the skewness or kurtosis of the data is between -1 and 1, it can also be said that the data appears to be normally distributed.



Inferential Statistics



Inferential Statistics

- Inferential statistics tries make inference or draw conclusion about the population using sample data.
- Examples of inferential statistical tests include:
 - T-test
 - Analysis of Variance
 - Chi-squared Test



T-Test

Goal:

- An independent sample t-test tries to investigate if there is a statistical significant difference between two groups with respect to a continuous variable.
- That is, do two groups differ significantly on a continuous variable?
- The t-test actually compares the means of two groups to find out if the mean of one group is significant different, higher or lower than the mean of the other group.



T-Test

An example of a research question for a t-test:

- Is there a significant difference in math score of a hybrid college class and a traditional college class?

Another way of asking this question:

- Does a hybrid approach to learning have a statistical significant effect on math scores?



T-Test

Treatment and control group:

- In the hybrid and traditional approach example, students who participated in the hybrid class are in the treatment group while those in a traditional class are in a control group.

■ **Does the intervention work?**

- With the hybrid class and traditional class example, we are trying to see if the hybrid method (new intervention) works or is better than the traditional teaching method.



T-test

The independent variable:

- The independent variable is teaching method (intervention or treatment) and has two levels or values (hybrid and traditional).

The dependent variable:

- The dependent variable or outcome is math score of students.



T-Test Assumptions

- Certain assumptions need to be met for the results of t-test to be valid for generalization or inference.
- **Independence assumption:** The observations or scores should be independent. This assumption is usually assumed met.
- **Normality Assumption:** The data should follow a normal distribution. A histogram or skewness can be used to check this assumption.
- **Homogeneity of Variance:** The variances of the two groups should be the same. You can use Levene's test to check for this assumption.



T-Test Assumptions (continues)

- **Homogeneity of Variance**

Test: The null hypothesis for this test is that the variances of the groups are equal.

- To check for homogeneity of variance, you want the p-value to be greater than 0.05. This indicates that the variances are equal.

- If the **variances are unequal** (p-value is less than 0.05) you can run a different version of t-test called the **Welch's t-test**.

- In some software, when running a t-test, you are provided options to include whether variances are equal or unequal.



Run the T-test

- After checking for the assumptions, run the t-test if the assumptions are met.
- The null hypothesis for the t-test is that there is no statistical significant difference between the groups on the dependent variable (alternative hypothesis: there is a significant difference...).
- When you run the t-test, the two values you should be looking for in the results include:
 - The test statistic (t-statistic)
 - The p-value of the t-statistic



How to Interpret the Results of a T-test

- If the p-value of the t-statistic is less than 0.05, you can reject the null hypothesis in support of the alternative hypothesis.
 - Then conclude that there is a statistical significant difference between the groups with respect to the dependent variable in your research question.
- If the p-value of the t-statistic is greater than 0.05, fail to reject the null hypothesis. That means the null hypothesis is supported.
 - Then, conclude that there is no statistical significant difference between the groups with respect to the independent variable.



Reporting the t-statistics

- The report of the results need to include:
 - The purpose of the test (technically your research question reworded).
 - Whether independence, normality, and homogeneity of variance assumptions are met or not.
 - Whether the results of are significant or not.
 - If the results of the t-test are significant, report the t-statistics and p-value in a formal manner.
 - If results of the t-test are significant, report the means and standard deviations of the groups.



Analysis of Variance (ANOVA)

Goal:

- The goal of ANOVA is to test if there is a statistical significant difference among three or more groups with respect to a dependent variable.
- It is an extension of the t-test to more than two groups.
- If there are three groups, the ANOVA test compares the mean of three different pairs of groups to find out if there is a significant difference between the means of at least one pair of groups.



ANOVA

An example of a research question for ANOVA:

- Is there a statistical significant difference in the math score of college students in an online class, a traditional class, and a hybrid class?
- Do the math score of college students differ across online, traditional and hybrid learning environments?



ANOVA

- The research question compares three teaching methods or learning environments on math score.
- We want to understand the effect of learning environments (online, hybrid and traditional) on the math score of students.
- The independent variable is the learning approach (could also be viewed as teaching method).
- The dependent variable is the math score of students.



ANOVA

- Same assumptions for the t-test should be checked before running the ANOVA.
- However, the ANOVA test is robust to normality. That means, even if your data is not normality distributed, you can still run an ANOVA and get valid results.
- The interpretation of ANOVA test results is similar to that of a t-test but note that the test statistics for ANOVA is the F-statistics (not t-statistics).
- Reporting your result is similar to how you report the t-test but note that you will report an F-statistics and a corresponding p-value.



ANOVA

- If the results of the ANOVA test are significant, you need to run a follow up test.
- The follow up test is usually called a post hoc test or multi-comparison test.
- The post hoc test tries to find exactly which groups differ.
- The specific groups that differ will also be reported as well as the means of the groups and the standard deviations for those groups.