STATISTICS AND PROBABILITY

01 ANALYZING CATEGORICAL DATA

* Individual: A single member or unit of a population or sample being studied.
* Variable: A characteristic or attribute that can take on different values among the individuals.
* Categorical variable: A variable that represents distinct groups or categories (e.g. gender, color, class...)
* Bar chart: A graphical representation of categorical data using rectangular bars where the length or height of each bar corresponds to the value it represents.
* Venn Diagrams: A visual representation of mathematical or logical relationships between different sets, using overlapping circles (or other shapes) to show common and distinct elements.
* Two-way Table: A table that displays the frequency distribution of two categorical variables, showing how the categories of one variable relates to the categories of another.
* When analyzing two-way relative tables, pay attention to whether it is column-relative or row-relative or both. For example, if it's only column-relative, sum of the values on the same row does not equal to 1.00 (100%).
* Distribution in Two-way Tables.
  * Joint Distribution: Two (or more) random variables occurring together (e.g. 20 out of 200 students studied 21 to 40 minutes per day and answered 60-79% of all questions correctly.)
  * Marginal Distribution: Ignoring (marginalizing out) other variables and giving the distribution of remaining variables (e.g. 70 out of 200 students answered 40-59% of all questions correctly.)

* Conditional Distribution: The distribution of one random variable given that another random variable takes a specific value (e.g. 14.28% of the students who studied more than 60 minutes per day answered 60-79% questions correctly)

02 DISPLAYING AND COMPARING QUANTITATIVE DATA

* Data can be represented in many ways, and each method highlights different aspects, making patterns and trends easier to understand.
* A frequency table organizes data by listing each category or value along with its corresponding count (frequency). A frequency dot plot visually represents this data using dots stacked above a number line, where each dot represents one occurrence of a value. Both tools help summarize and display the distribution of a dataset.
* A histogram is a graphical representation of data that uses bars to display the frequency of values within specified intervals (bins). Unlike bar graphs the bars in a histogram touch each other, emphasizing the continuous nature of quantitative data. It helps visualize the distribution, shape, and spread of a dataset.
* A stem-and-leaf plot organizes numerical data by splitting each value into a stem (leading digit(s)) and a leaf (the trailing digit). The stems are listed vertically, and the leaves extend horizontally, showing the distribution while preserving the original data points. It's useful for displaying small datasets while maintaining exact values.
* Left-tailed distribution graph means the data is skewed to the left, right-tailed distribution graph means the data is skewed to the right, and a distribution graph with no tails means the data is symmetrical.
* Clusters: Groups of data points concentrated closely together in a distribution.

* Gaps: Unusually empty ranges where no data points appear between values.
* Peaks: The highest frequencies in distribution, representing modes or local maxima
* Outliers: Extreme values that fall for outside the overall pattern of the data.
* Line Graphs display data points connected by straight lines, showing trends or changes over time. It's ideal for visualizing continuous data, and comparing multiple datasets (using different colored lines).
* Line graphs can be misleading if:
  1. The x-axis is truncated (doesn't start at zero), making small changes dramatic.
  2. The x- or y-axis is stretched/compressed, distorting the rate of change
  3. Uneven intervals are used, creating false patterns.

03 SUMMARIZING QUANTITATIVE DATA

* Statistics is the science of collecting, analyzing, interpreting, and presenting data.
  * Descriptive statistics summarizes and describes the main features of a dataset using measures of central tendency and measures of spread.
  * Inferential statistics draws conclusions or makes predictions about a population (the entire group) based on sample (a subset of that population) data using hypothesis testing, confidence intervals, and regression analysis.
* Measures of central tendency help us answer the question, "What is the center/ middle / typical value / average of this data?"
  * Mean is the sum of all values in a dataset, divided by the number of values. It's the "balancing point", because the total distance from it to the data points below it is equal to the total distance from it to the data points above it.
    * Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ * Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  * Median is the middle value in an ordered dataset (the mean of the middle two if total number of values is even).
  * Mode is the most frequent value in a dataset.

* Removing an outlier significantly affects the mean, pulling it toward the remaining data, while the median remains relatively stable since it depends only on the middle values. Scaling an outlier pulls the mean strongly in its direction but has little to no effect on the median. Because of these reasons, mean is meaningful only if the dataset is not skewed and there are no extreme outliers. If any of these conditions exists, it's better to use the median. If neither are applicable (e.g. data is categorical), the mode is used.
* Measures of Spread help us to answer the question, "How spread apart is the data?"
  * Range is the difference between the maximum and the minimum data points. Range and its half (mid-range) are highly sensitive to outliers.
  * Interquartile Range (IQR) is the amount of spread in the middle 50% of a dataset which equals to the distance between the first quartile (Q_1) and the third (Q_3), where Q_1 is the median of the data points to the left of the median of the ordered dataset, and Q_3 is the median of the data points to the right of the median of the ordered dataset. Since it ignores outliers, IQR is a better measure of spread than the range for data with outliers.
  * Variance measures how far the data points are spread out from the mean of the set. The formulas for the population and sample are different:
    * Variance (for population) $= \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$
    * Variance (for sample) $= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
    * When the sample mean ($\bar{x}$) is calculated, $(x_i-\bar{x})$ are smaller than the true deviations $(x_i-\mu)$, which causes the sum of squared deviations to be smaller than it should be, leading to underestimation of the

variance. To adjust this, we divide by $(n-1)$ instead of n. $(n-1)$ arises because one degree of freedom is lost. If we were estimating more parameters (e.g. in linear regression) we might divide by $(n-p)$ where p is the number of estimated parameters.

* The standard deviation is the square root of the variance. Because it shares the same unit of measurement as the original dataset, it provides a more intuitive and practical measure of spread than variance

* If the data is skewed (and/or has extreme outliers), the standard deviation (or variance) is not the best measure to evaluate the spread because it is based on the mean, which itself is not the best measure to evaluate the central tendency in these conditions. For these types of data the IQR is a better measure of spread.
* A Box-and-Whisker Plot displays the five-number summary of a dataset (① The minimum ② First quartile (box edge) ③ Median (line inside the box) ④ Third quartile (box edge) ⑤ The maximum) with potential outliers marked individually.
* The most common method to classify a data point as an outlier is to use thresholds based on IQR.
  * Lower outlier threshold = $Q_1 - 1.5\ IQR$ (1.5 is the most common scalar)
  * Higher outlier threshold = $Q_3 + 1.5\ IQR$
* Mean Absolute Deviation (MAD) = $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$. MAD is less sensitive to extreme values than variance (and standard deviation) because it uses absolute differences instead of squares. However, it lacks mathematical properties exploited in advance statistics, such as differentiability.