**Statistics class notes**

**📖 Part 1: Introduction to Statistics**

* ✅ **Definition:** Statistics is a field of study focused on **collecting, organizing, analyzing, and interpreting data** to make informed decisions.
* ✅ **Key Terminology:**
  + **Population:** The entire group of individuals or items you are interested in studying (e.g., all students in a school).
  + **Sample:** A smaller, manageable subset of the population that is selected for study (e.g., 50 students from that school).
  + **Parameter:** A numerical value that describes a characteristic of the **population** (e.g., the average height of *all* students).
  + **Statistic:** A numerical value that describes a characteristic of a **sample** (e.g., the average height of the *50 sampled* students). It is used to estimate the population parameter.

**📖 Part 2: Types of Data**

Data can be broadly classified into two main types:

**1. Quantitative (Numerical) Data**

* ✅ Represents measurable quantities. It answers "how much" or "how many."
* **Types of Quantitative Data:**
  + **Continuous:** Data that can take any value within a given range. It is *measured*.
    - **Examples:** Height, weight, temperature, time.
  + **Discrete:** Data that can only take specific, distinct integer values. It is *counted*.
    - **Examples:** Number of website visitors, number of errors in a program, word count.

**2. Qualitative (Categorical) Data**

* ✅ Represents characteristics or qualities that can be grouped into categories.
* **Types of Qualitative Data:**
  + **Nominal:** Categories with **no natural order** or ranking.
    - **Examples:** Hair color, country of origin, product type.
  + **Ordinal:** Categories with a **meaningful order** or ranking, but the intervals between them may not be uniform.
    - **Examples:** Education level (bachelor's, master's), shirt size (small, medium, large).
  + **Binary:** A special case with only **two possible outcomes**.
    - **Examples:** Yes/No, True/False, 0/1, Pass/Fail.

**📖 Part 3: Descriptive Statistics**

**1. Measures of Central Tendency**

* ✅ These measures describe the center or typical value of a dataset.
  + **Mean (Average):** The sum of all values divided by the number of values. It is highly **sensitive to outliers**.
  + **Median:** The middle value in a sorted dataset. It is **not affected by outliers**, making it a robust measure for skewed data.
  + **Mode:** The value that appears most frequently in a dataset.

**2. Measures of Dispersion (Spread)**

* ✅ These measures describe how spread out or varied the data points are.
  + **Variance:** Measures the average squared difference of each data point from the mean.
    - 🔑 **Population Variance (σ²):** Σ(xi - μ)² / n
    - 🔑 **Sample Variance (s²):** Σ(xi - x̄)² / (n - 1) (Using n-1 provides a more accurate estimate of the population variance).
  + **Standard Deviation (SD):** The square root of the variance. It measures the average distance of data points from the mean, in the same units as the data.
    - 🔑 **SD:** √Variance
  + **Interquartile Range (IQR):** Describes the spread of the middle 50% of the data.
    - 🔑 **Formula:** IQR = Q3 - Q1
    - **Using IQR to Find Outliers:** A data point is considered an outlier if it falls outside these bounds:
      * 🔑 **Lower Bound:** Q1 - 1.5 \* IQR
      * 🔑 **Upper Bound:** Q3 + 1.5 \* IQR

**3. Skewness**

* ✅ A measure of the **asymmetry** of a distribution, indicating if the data is concentrated on one end.
  + **Positive Skew (Right-skewed):** The tail stretches to the right. Most data is on the left.
    - *Example:* Most students scored low, but a few scored very high.
    - 🔑 **Rule:** Mean > Median > Mode
  + **Negative Skew (Left-skewed):** The tail stretches to the left. Most data is on the right.
    - *Example:* Most students scored high, but a few scored very low.
    - 🔑 **Rule:** Mean < Median < Mode
  + **Zero Skew:** The distribution is symmetrical (like a Normal Distribution).
    - 🔑 **Rule:** Mean = Median = Mode

**📖 Part 4: Relationship Between Variables**

**1. Covariance**

* ✅ Measures the **direction** of the linear relationship between two variables (i.e., whether they move together or in opposite directions).
* **Drawback:** The value is not standardized, making it hard to interpret the **strength** of the relationship.

**2. Correlation**

* ✅ A **standardized** version of covariance that measures both the **direction and strength** of the linear relationship.
* **Interpretation:** The value is always between **-1 and +1**.
  + **+1:** Perfect positive linear relationship.
  + **-1:** Perfect negative linear relationship.
  + **0:** No linear relationship.
* 🔑 **Formula:** Corr(X, Y) = Cov(X, Y) / (σx \* σy)

**📖 Part 5: Probability**

* ✅ A measure of the likelihood that an event will occur, ranging from 0 (impossible) to 1 (certain).

**1. Key Concepts**

* **Independent Events:** The outcome of one event does not affect the outcome of another.
  + 🔑 P(A|B) = P(A)
* **Dependent Events:** The outcome of one event affects the outcome of another.
  + 🔑 P(A|B) ≠ P(A)
* **Joint Probability (AND):** The probability of two events happening together.
  + 🔑 For independent events: P(A ∩ B) = P(A) \* P(B)

**2. Conditional Probability**

* ✅ The probability of an event (A) occurring, given that another event (B) has already occurred.
* 🔑 **Formula:** P(A|B) = P(A ∩ B) / P(B)

**3. Bayes' Theorem**

* ✅ A theorem used to **update the probability** of a hypothesis based on new evidence. It "reverses" the conditional probability.
* **Use Case:** It helps find P(Cause | Effect) when you already know P(Effect | Cause).
  + *Example:* We know the probability of a **positive test** given you **have a disease**. Bayes' Theorem helps us find the probability you **have the disease** given you got a **positive test**.
* 🔑 **Formula:** P(A|B) = [P(B|A) \* P(A)] / P(B)

**📖 Part 6: Probability Distributions**

**1. Probability Mass Function (PMF)**

* ✅ Used for **discrete** random variables.
* ✅ It gives the probability that a variable is **exactly equal** to a specific value.
* *Example:* The probability of rolling exactly a 4 on a die is 1/6.

**2. Probability Density Function (PDF)**

* ✅ Used for **continuous** random variables.
* ✅ The probability is found by calculating the **area under the curve** over a specific interval.
* 🔑 **Important:** The probability of a continuous variable taking on a single, exact value is **zero**.

**3. Cumulative Distribution Function (CDF)**

* ✅ Used for both discrete and continuous variables.
* ✅ It gives the probability that a variable will take a value **less than or equal to** a specific value x.
* 🔑 **Formula:** CDF(x) = P(X ≤ x)

**4. Bernoulli Distribution**

* ✅ Models a **single trial** with exactly **two possible outcomes**: success (1) or failure (0).
* *Example:* A single coin flip; a single customer making a purchase or not.

**5. Binomial Distribution**

* ✅ Models the **number of successes (k)** in a fixed number of **independent Bernoulli trials (n)**.
* *Example:* The probability of getting exactly 3 heads (k=3) in 10 coin flips (n=10).
* 🔑 **PMF Formula:** P(k) = (nCk) \* p^k \* (1-p)^(n-k)