FREQUENCY DISTRIBUTION  
(PHÂN PHỐI TẦN SỐ)

**I. Definition:** A frequency distribution is a powerful tool in statistics that helps organize and summarize data. It essentially shows how often different values or categories occur within a dataset. This visualization makes it easier to understand the underlying patterns and trends in the data.

**- Frequency Tables:** Can be grouped / ungrouped

For example: AGE **(**20-29, 30-39, 40-49 / 24, 32, 38, 41, 43, 47)

- **Frequency Graphs:**

+ Histogram / Polygon / Boxplot [metric data]

+ Bar Chart [SPSS terminology] / Pie Chart [categorical data]

**II. Types of Frequency Distributions:**

**1. *Ungrouped Frequency Distribution***: Suitable for smaller datasets or when dealing with discrete data (e.g., number of children in a family). Each data point has its own frequency.

|  |  |
| --- | --- |
| **Score on test** | **Number of students** |
| 5 | 3 |
| 10 | 4 |
| 15 | 5 |
| 18 | 4 |
| 20 | 4 |
| Total: 20 | |

***2. Grouped Frequency Distribution*:** Used for larger datasets or continuous data (e.g., heights of students). Data is divided into classes, and the frequency represents the number of data points within each class.

|  |  |
| --- | --- |
| **Score on test** | **Number of students** |
| 0-5 | 3 |
| 6-10 | 4 |
| 11-15 | 5 |
| 16-20 | 8 |
| Total: 20 | |

Σ**ƒ = 12 [n=12]**

Σ**ƒX = 18**

Proportion = (ρ=

|  |  |  |
| --- | --- | --- |
| **Number of siblings [X]** | **ƒ (frequency) (Tần số)** | **ƒX (ƒ multiply X)** |
| 0 | 3 | 0 |
| 1 | 5 | 5 |
| 2 | 1 | 2 |
| 3 | 2 | 6 |
| 4 | 0 | 0 |
| 5 | 1 | 5 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of siblings [X]** | **ƒ (frequency) (Tần số)** | **ƒX (ƒ multiply X)** | ρ= | ρ(100) |
| 0 | 3 | 0 |  | 25% |
| 1 | 5 | 5 |  | 41.7% |
| 2 | 1 | 2 |  | 8.3% |
| 3 | 2 | 6 |  | 16.7% |
| 4 | 0 | 0 |  | 0% |
| 5 | 1 | 5 |  | 8.3% |
|  | 12 | 18 | 1 | 100% |

**III. Levels of Measurement of Variables**

**-** Metric Variables (Biến Định Lượng): Height in cm, Temperature in degrees Celsius, Number of siblings (Numbers mean something real).

- Categorical Variables (Biến Phân Loại): Gender, Nationality, Location, Diet (Numbers are arbitrary and used as codes).

**IV. Features of Distributions (Metric Data) Summary (Đặc Điểm của Phân Phối (Dữ Liệu Định Lượng)**

- Central Tendency: Represents the center of the distribution using measures like mean, median, and mode, indicating the most typical or representative value of the group.

- Variability: Assesses how spread out or clustered the scores are, commonly measured by standard deviation or interquartile range (IQR).

- Shape: Distributions can be symmetric (mirror image on both sides) or skewed (with scores tapering off at one end), indicating positive or negative skewness.

**V. Formula for calculation of the mean (Trung Bình)**

**1. Definition:** Mean is the average value of a set of numbers.

**2. Population Mean calculation**:

𝝁 represents the mean of the population

Σ is a shorthand way of writing "sum of "

𝑿 represents a data value

N is the total number of population data values

𝝁 (MU): population mean

**3. Sample Mean calculation:**

𝑴 represents the mean of the sample

Σ is a shorthand way of writing "sum of "

𝑿 represents a data value

n is the total number of sample data values

**VI. The median (Trung vị):** the median provides us with the location of the midpoint

in the distribution => middle number of data set

Data set: 7, 7, 10, 14, 15, 23, 32 => Median 14 because 14 is the middle number of data set

**VII. Differences between Mean and Median**

**- Mean:** Affected by extreme values ​​(outliers) and often does not accurately reflect the center of the data in the drift distribution.

- **Median**: Not affected by extreme values, so is just a better number for drifting dispersions or outliers.

**VIII. The mode (Mốt)**

The mode is the score or category that appears most frequently in a dataset and is a useful measure of central tendency, especially for categorical data. Unlike the mean and median, the mode can represent an actual score in the data and can have multiple values, resulting in unimodal, bimodal, or multimodal distributions. It is particularly relevant for nominal scales and can also be useful for ordinal data, although the median is often preferred in such cases.

Data set: 7, 7, 14, 10, 15, 23, 32 => Mode 2 because the frequency of number 7 is 2

**IX. Symmetric[al] and Skewed distributions:**

1. **Symmetric distribution (Phân phối đối xứng):** Symmetrical distributions have a balanced shape where the mean, median, and mode are all located at the center.
2. **Skewed distribution (Phân phối lệch):** skewed distributions have an uneven shape, causing the mean, median, and mode to be positioned differently**.**

* **Positive Skew**: Tail extends to the right; order of central tendency is mode < median < mean.
* **Negative Skew**: Tail extends to the left; order of central tendency is mean < median < mode.

**X. Percentiles and Percentile Rank**

**- Percentiles**: The percentage of values in a distribution that are at or below a specific value. For example, if a value is at the 85th percentile, then 85% of the values in the distribution are at or below that value.

**- Percentile Rank**: The percentage of individuals who have a score at or below a specific value. If a score has a percentile rank of 16%, it means that 16% of individuals have a score equal to or lower than that score.