2024
SUMMARY

COMMUNITY MEDICINE

DR. AHMAD AL-QAWASMI

ARKAN ACADEMY

Biostatistics

- **Variable:** a characteristic of interest that can have different values (measured, record, analyzed)
  - Independent variable is the presumed cause, and dependent variable is the presumed effect
    - ✓ Smoking is independent variable → Lung cancer is a dependent variable
    - ✓ Sugar pill placebo is independent variable → Unchanged BP is a dependent variable

- A quantitative variable (can be measured), qualitative variable (not measured but sorted)

1. Qualitative (categorical) variables

- **Nominal:** such as sex, exam result, blood group, eye color, ethnicity, marital stats, housing style
  - Nominal level of measurement: Cannot perform meaningful operation on this data
  - Involve names, labels or qualities, car type, colors, zip code, religion
  - The *lowest level of measurement*, which is a qualitative classification with no quantitative value
  - Appropriate statistics: **mode**, **frequency** BUT we cannot use mean
- **Ordinal:** such as response to treatment (poor, fair, good), severity of disease (mild, moderate, severe) and income status (low, middle, high), where the ordinal level of measurements:
  - Indicates order not exact quantity (exact differences between ranks cannot be specified)
  - Involves using numbers to designate ordering on an attribute
  - Appropriate statistics: **mode**, **frequency** and **median** BUT we cannot use mean

2. Quantitative (numerical) variables

- **Discrete:** number of family members, heart beats or hospital admissions in a day
- **Continuous** (no gap between possible values): Height, weight, age, BP, serum cholesterol, BMI
  - Continuous data can be transformed into classification in qualitative data
- **Interval level of Measurement:** Data is placed in meaningful intervals and order (ranked)
  - Real (actual) numbers and the difference between the ranks can be specified (numbers indicate both the ordering and the distance between the values)
  - Appropriate statistics: **mode**, **frequency**, **median** and **mean**
  - Example: Temperature, year of birth, order of racers
- **Ratio level of Measurement:** Is the highest level where data can be categorized, ranked and the difference between ranks can be specified and has a zero point (total absence of the quantity)
  - Age, weight, height, pulse rate, time
  - Appropriate statistics: **mode**, **frequency**, **median** and **mean**
- The goal of the researcher is to use the highest level of measurement possible

	Category Names	Meaningful Order	Equal Distance	True Zero & Ratios
Nominal	✓	—	—	—
Ordinal	✓	✓	—	—
Interval	✓	✓	✓	—
Ratio	✓	✓	✓	✓

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

- **Descriptive Statistics:** makes data *more understandable* by organizing, summarizing, and displaying data
- **Inferential Statistics:** Reports the confidence of a sample statistic in predicting a population parameter
  - Parameter: A descriptive measure computed from population data (population mean μ, standard deviation σ, variance σ^2)
  - Statistic: computed from sample data (sample mean $\bar{x}$ and standard deviation s, variance s^2)

Raw Data: Data collected without processing
Organized Data: Arranged in ascending/descending order

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Describe data characteristics	Make predictions and generalizations
Data focus	Summarizes sample data only	Draws conclusions about a population
Technique used	Measures of central tendency and dispersion	Hypothesis testing, regression, confidence intervals
Interpretation	Limited to known data	Extends findings to unknown data

Descriptive Measures

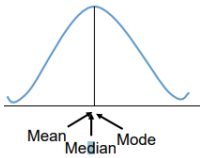
- **Measures of location** include measures of central tendency (Mean, median, mode) and non-central tendency (Quantiles, quartiles, quintiles, percentiles)
  - Measure of central tendency help find the *approximate center of the dataset*
  - Mean, mode, median are all considered as **average**

- Measures of central tendency
- **Mean:** the sum of all values divided by the number of observations.
  - Generally, **not part** of the data set and **Unique** for a given dataset
  - **Sensitive to extreme** values

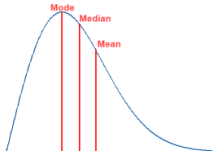
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Median:** the **middle** value when data is arranged in ascending order
  - If is odd number of values, the median is the middle value
  - If is even number of values, the median is the average of the two middle values
  - **Not affected by extreme** values, and **may be part** of the data set
- **Mode:** the **most frequently** occurring value in a dataset.
  - A dataset can have no mode, 1 mode (unimodal), 2 modes (bimodal), multiple modes (multimodal)
  - It is **always part** of the data set, and **not affected by extreme** values

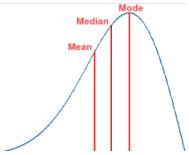
Shapes of Distributions



- Normal Distribution (Bell-shaped curve): Symmetrical (mean = median = mode)
- Right (Positively) Skewed: Long right tail (Mean > Median > Mode)
  - Mean overestimates the most common values
  - Example: Income distribution



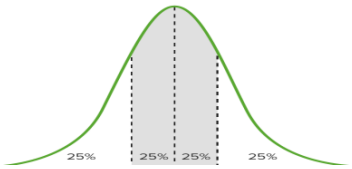
- Left (Negatively) Skewed: Long left tail (Mean < Median < Mode)
  - Mean underestimates the most common values
  - Example: Exam scores



- The **median** is a more robust statistic than mean in the presence of extreme values and skewed

- Measures of Non-Central Location: summarize data by dividing it into equal parts
- **Quartiles:** Divide the data into four equal parts
  - Q1 (First Quartile): 25% of observations are below it
  - Q2 (Second Quartile, Median): 50% of observations are below it
  - Q3 (Third Quartile): 75% of observations are below it
- Quintiles: Divide data into five equal parts
- Deciles: Divide data into ten equal parts.
- Percentiles: Divide data into 100 equal parts.
  - 25^{th} percentile = Q1, 50^{th} percentile = Q2 (median), 75^{th} percentile = Q3.

- **Measures of Dispersion** are commonly used to measure how **spread out** the data is, and include:
- Range: The difference between the largest and smallest values
  - Pros: Easy to calculate
  - Cons: Affected by extreme values
- Interquartile Range (IQR): Difference between Q3 and Q1
  - Measures spread of the middle 50% of the data
  - Not sensitive to extreme values



- Standard Deviation (σ): Measures the variation of values from the mean
  - Low standard deviation indicates data close around the mean (reliable)
  - Larger standard deviation indicates widely spread data
  - N is replaced by n-1 (*degree of freedom*) in the second equation as a *correction* for the sample size to produce *unbiased* estimator

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$$

Stock x	Deviation $x-\mu$	Squared $(x-\mu)^2$
56	−5	25
58	−3	9
61	0	0
63	2	4
67	6	36
$\Sigma x = 305$	$\Sigma(x-\mu) = 0$	$\Sigma(x-\mu)^2 = 74$

- Variance: The **square** of the standard deviation (always positive)
- Standard Error: Measures how accurately a sample mean represents the population mean
  - Decreases (more accurate) with increasing sample size

$$\text{Standard Error} = \frac{S}{\sqrt{n}}$$

- Five-Number Summary 'minimum, Q1, median, Q3, maximum'
  - Provides a quick overview of data distribution

- Importance of Descriptive Statistics:
  - Helps in visualizing and simplifying data for better interpretation
  - Identifies patterns, trends, and distributions to aid in decision-making

ARKAN

ACADEMY

علم في كل مكان

Arkan academy

Arkanacademy

www.arkan-academy.com

+962 790408805