# 575 STAT

# Biostatistics

**Text Book:** Biostatistics.

TENTH EDITION

# Biostatistics

## A Foundation for Analysis in the Health Sciences

Wayne W. Daniel | Chad L. Cross

# Schedule of Assessment Tasks for Students During the Semester

| Assessment | Examination | Week due | Proportion of Final Assessment |
|---|---|---|---|
| 1 | Mid-term exam | $7^{th}$ - $12^{th}$ | 50% |
| 2 | Project | $14^{th}$ | 10% |
| 3 | Final exam | After $15^{th}$ Week | 40% |

## Definition: Statistics

Statistics is a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting and drawing conclusions based on that data.

# Biostatistics

When the data is obtained from the biological sciences and medicine, we use the term "biostatistics".

# Types of Statistics

## 1- Descriptive Statistics

Descriptive Statistics are used to summarize or describe the important characteristics of a known set of data.

**For example:** Let us consider everyone in this room. Each one of us is a source of data. A characteristic of this data may be degree program, age, height, sex, marital status.

# Types of Statistics

## 2- Inferential Statistics

Inferential Statistics goes beyond the description. It involves the use of sample data to make inferences about a larger set of data from which the sample was chosen.

**For example:** If we consider this class as a sample of KSU students and calculated the average age of the class. We could then infer that the average age of all KSU students is the same as our sample.

# Population and Sample

## Definition:

A **population** is the complete collection of elements (scores, people, measurements) to be studied.

A **sample** is a subcollection of elements drawn from the population.

# Population and Sample

**Population**

**(Some Unknown Parameters)**

**Example: KSU Students**

**(Height Mean)**

**N=Population Size**

**Sample = Observations**

**(We calculate Some Statistics)**

Example: 20 Students from KSU

**(Sample Mean)**

**n = Sample Size**

# Population Size and Sample Size

The number of elements in the population is called the population size and is denoted by $N$.

The number of elements in the sample is called the sample size and is denoted by $n$.

# Notes:

- Let $X_1, X_2, \ldots, X_N$ be the population values (in general, they are unknown)
- Let $x_1, x_2, \ldots, x_n$ be the sample values (these values are known)
- Statistics obtained from the sample are used to estimate (approximate) the parameters of the population.

# Why do we study and analyze subcollections (samples) of a population

**Definition (Parameter)**

It is a numerical characteristics of a population that summarize the data for the entire population.

**Definition (Statistic)**

It is a numerical characteristics of a sample.

# Definition (Variables)

A variable is a characteristic, feature or factor that varies from one individual to another in a population.

# Classification of Variables

## Quantitative Variables

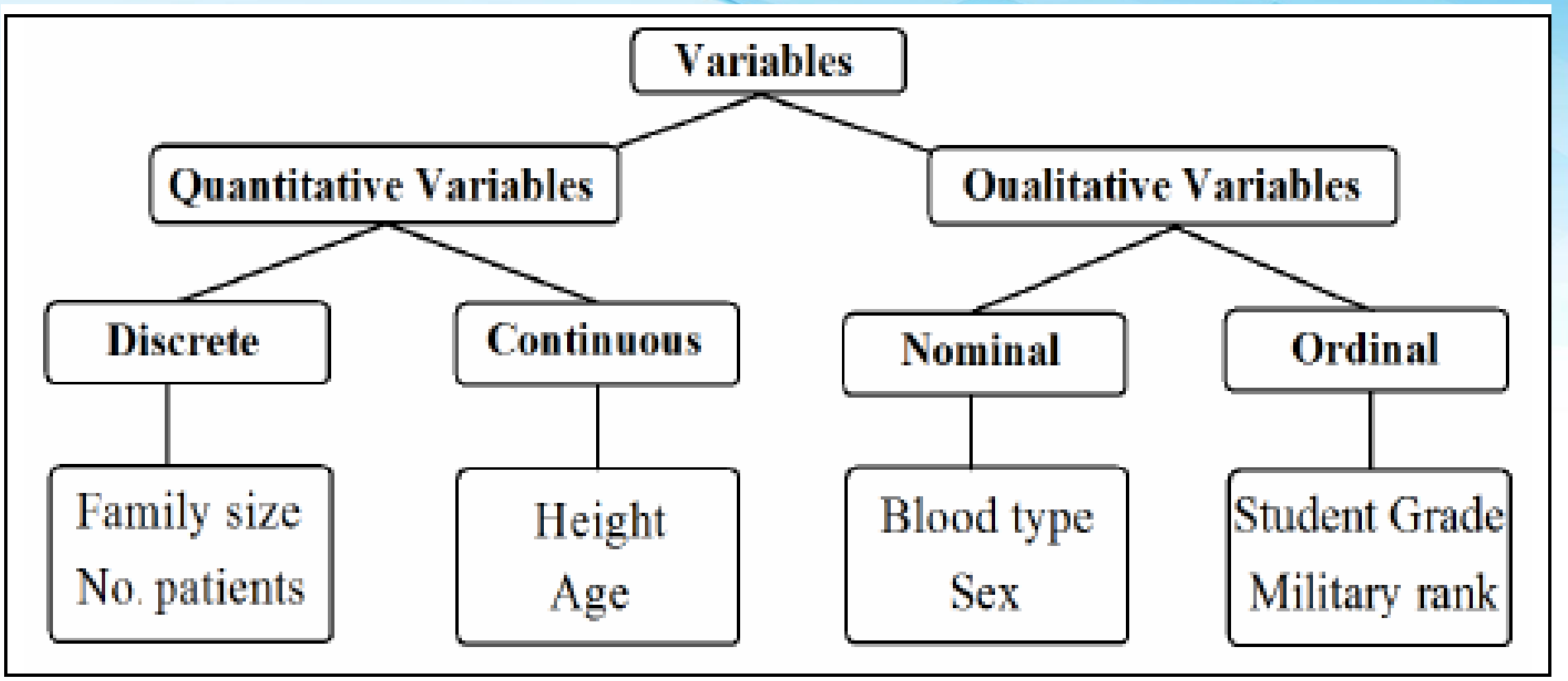
### Discrete variables

(The number of cars in a parking lot – The number of patients in a hospital )

### Continuous variables

(height - weight - time it takes to get to school)

## Qualitative Variables


Variables
Quantitative Variables
Oualitative Variables
Discrete
Continuous
Nominal
Ordinal
Family size
No. patients
Height
Age
Blood type
Sex
Student Grade
Military rank

# Chapter 2

## Strategies for Understanding the Meaning of Data

## 1. Introduction:

Summarization techniques involve:

1) frequency distributions
2) descriptive measures

## 2. The Ordered Array:

***Example:*** The following values represent a list of ages in a study on smoking cessation:

55 46 58 54 52 69 40 65 53 58

The ordered array is:

40 46 52 53 54 55 58 58 65 69

# Grouped Data: The Frequency Distribution

## To group a set of observations:

1) select a suitable set of contiguous,
2) non-overlapping intervals.

**Example:** The following table gives the hemoglobin level (g/dl) of a sample of 50 men.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 17.0 | 17.7 | 15.9 | 15.2 | 16.2 | 17.1 | 15.7 | 17.3 | **13.5** | 16.3 |
| 14.6 | 15.8 | 15.3 | 16.4 | 13.7 | 16.2 | 16.4 | 16.1 | 17.0 | 15.9 |
| 14.0 | 16.2 | 16.4 | 14.9 | 17.8 | 16.1 | 15.5 | **18.3** | 15.8 | 16.7 |
| 15.9 | 15.3 | 13.9 | 16.8 | 15.9 | 16.3 | 17.4 | 15.0 | 17.5 | 16.1 |
| 14.2 | 16.1 | 15.7 | 15.1 | 17.4 | 16.5 | 14.4 | 16.3 | 17.3 | 15.8 |

Intervals:

13.0 – 13.9 , 14.0 – 14.9 , 15.0 – 15.9 , 16.0 – 16.9 , 17.0 – 17.9 , 18.0 – 18.9

Variable = X = hemoglobin level (continuous, quantitative)

Sample size = n = 50, Max= 18.3, Min= 13.5

| Class Interval | Tally | Frequency |
|---|---|---|
| 13.0 – 13.9 | 𝍸 | 3 |
| 14.0 – 14.9 | 𝍸 | 5 |
| 15.0 – 15.9 | 𝍸 𝍸 𝍸 | 15 |
| 16.0 – 16.9 | 𝍸 𝍸 𝍸 \| | 16 |
| 17.0 – 17.9 | 𝍸 𝍸 | 10 |
| 18.0 – 18.9 | \| | 1 |

# The grouped frequency distribution for the hemoglobin level of the 50 men is:

| Class Interval (Hemoglobin level) | Frequency (no. of men) |
|---|---|
| 13.0 – 13.9 | 3 |
| 14.0 – 14.9 | 5 |
| 15.0 – 15.9 | 15 |
| 16.0 – 16.9 | 16 |
| 17.0 – 17.9 | 10 |
| 18.0 – 18.9 | 1 |
| Total | $n$=50 |

## Notes:

1. Minimum value ∈ first interval.
2. Maximum value ∈ last interval.
3. The intervals are not overlapped.
4. Each value belongs to one, and only one, interval.
5. Total of the frequencies = the sample size = $n$

## Mid-Points of Class Intervals:

$$\text{Mid-point} = \frac{\text{upper limit} + \text{lower limit}}{2}$$

**For example**, Mid-point of the 1st interval = (13.0+13.9)/2 = 13.45.

- Mid-point of a class interval is considered as a typical (approximated) value for all values in that class interval.
- There are no gaps between true class intervals.

## True Class Intervals:

- d = gap between class intervals
- d = lower limit – upper limit of the preceding class interval
- true upper limit = upper limit +d/2
- true lower limit = lower limit - d/2

| Class Interval | True Class Interval | Mid-point | Frequency |
|---|---|---|---|
| 13.0 – 13.9 | 12.95 - 13.95 | 13.45 | 3 |
| 14.0 – 14.9 | 13.95 - 14.95 | 14.45 | 5 |
| 15.0 – 15.9 | 14.95 - 15.95 | 15.45 | 15 |
| 16.0 – 16.9 | 15.95 - 16.95 | 16.45 | 16 |
| 17.0 – 17.9 | 16.95 - 17.95 | 17.45 | 10 |
| 18.0 – 18.9 | 17.95 – 18.95 | 18.45 | 1 |

## Cumulative frequency ( CF ):

CF of the 1st class interval = frequency.

CF of a class interval = frequency + CF of the preceding class interval.

## Relative frequency and Percentage frequency:

Relative frequency = frequency / n

Percentage frequency = Relative frequency × 100%

| Class Interval | Frequency | Cumulative Frequency | Relative Frequency | Cumulative Relative Frequency | Percentage Frequency | Cumulative Percentage Frequency |
|---|---|---|---|---|---|---|
| 13.0 – 13.9 | 3 | 3 | 0.06 | 0.06 | 6% | 6% |
| 14.0 – 14.9 | 5 | 8 | 0.10 | 0.16 | 10% | 16% |
| 15.0 – 15.9 | 15 | 23 | 0.30 | 0.46 | 30% | 46% |
| 16.0 – 16.9 | 16 | 39 | 0.32 | 0.78 | 32% | 78% |
| 17.0 – 17.9 | 10 | 49 | 0.20 | 0.98 | 20% | 98% |
| 18.0 – 18.9 | 1 | 50 | 0.02 | 1.00 | 2% | 100% |

- The number of people whose hemoglobin levels are between 17 and 17.9 = 10
- The number of people whose hemoglobin levels are less than or equal to 15.9 = 23
- The number of people whose hemoglobin levels are less than or equal to 17.9 = 49

- The percentage of people whose hemoglobin levels are between 17 and 17.9 = 20%
- The percentage of people whose hemoglobin levels are less than or equal to 14.9 = 16%
- The percentage of people whose hemoglobin levels are less than or equal to 16.9 = 78%