

# Lecture1

## Introduction & Descriptive Statistics

1

### Basic Concepts - I

- **Data**
  - Any kind of numbers
  - Statistical analyses need numbers
- **Statistics**
  - Concerned with collection, organization, and analysis of data
  - Drawing inferences about a population when only a sample of the population is studied
- **Summary**
  - Data are numbers, numbers contain information, statistics investigate and evaluate the nature and meaning of this information

2

### Basic Concepts - II

- **Sources of Data**
  - Routinely kept records
  - Surveys
  - Experiments / Research Studies
- **Biostatistics**
  - Statistics applied to biological sciences and medicine
  - Statistics including not only analytic techniques but also study design issues
- **Variables**
  - a characteristic that can take on different values for different persons, places or things
  - **Statistical analyses need variability; otherwise there is nothing to study**

3

### Basic Concepts - III

- **Types of Variables**
  - Qualitative
  - Quantitative
  - Random
    - Discrete random
    - Continuous random
- **Population**
  - Sample
  - **Statistical analysis infers from a sample the characteristics of the population**

4

### Measurements and Measurement Scales

- **Measurement**
  - assignment of representative numbers to objects or events according to a set of rules
- **Measurement Scales**
  - Nominal
    - Yes/No
  - Ordinal
    - Income (low/medium/high)
  - Interval
    - Degrees (Fahrenheit, Celsius)
  - Ratio
    - Height, weight

5

### Random Sample

- **Reason**
  - sample a 'small' number of subjects from a population to make inference about the population
  - **Essence of statistical inference**
- **Definition**
  - A sample of size  $n$  drawn from a population of size  $N$  in such a way that every possible sample of size  $n$  has the same chance of being selected
- **Sampling with and without replacement**
  - In biostatistics, most sampling done without replacement

6

## Descriptive Statistics: Measures of Location

- Descriptive measure computed from sample data - *statistic*
- Descriptive measure computed from population data - *parameter*
- Most common measures of location
  - Mean
  - Median
  - Mode

7

## mean

- Probably most common of the measures of central tendency
  - Average
- Definition

$$\bar{x} = \frac{\sum x_i}{n}$$

- Normal distribution, although we tend to use it regardless of distribution
- Weakness
  - Influenced by extreme values

8

## Median

- Frequently used if there are extreme values in a distribution or if the distribution is non-normal
- Definition
  - That value that divides the 'ordered array' into two equal parts
    - If an odd number of observations, the median will be the (n+1)/2 observation
      - ex.: median of 11 observations is the 6th observation
    - If an even number of observations, the median will be the midpoint between the middle two observations
      - ex.: median of 12 observations is the midpoint between 6th and 7th
- Comparison of mean and median indicates skewness of distribution

9

## Mode

- Not used very frequently in practice
- Definition
  - Value that occurs most frequently in data set
- If all values different, no mode
- May be more than one mode
  - Bimodal or multimodal

10

## Measures of Dispersion

- Dispersion of a set of observations is the variety exhibited by the observations
  - If all values are the same, no dispersion
  - More the values are spread, the greater the dispersion
- Many distributions are well-described by measure of location and dispersion
- Common measures
  - Range
  - Quartiles
  - Variance
  - Standard deviation
  - Coefficient of variation

11

## Range, Percentiles and Quartiles

- Definition of Range
  - The difference between the smallest and largest values in the data set
  - Heavily influenced by two most extreme values and ignores the rest of the distribution
- Definition of Percentiles
  - Given a set of n observations  $x_1, x_2, \dots, x_n$ , the pth percentile P is value of X such that p percent or less of the observations are less than P and (100-p) percent or less are greater than P
  - $P_{10}$  indicates 10th percentile, etc.
- Definition of Quartiles
  - First quartile is  $P_{25}$
  - Second quartile is median or  $P_{50}$
  - Third quartile is  $P_{75}$
- Definition of Interquartile Range
  - $IQR = Q_3 - Q_1$ .

12

## Variance

- Variance measures distribution of the values around their mean
- Definition of sample variance

$$s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$$

- Degrees of freedom
  - n-1 used because if we know n-1 deviations, the nth deviation is known
  - Deviations have to sum to zero
- Definition of sample standard deviation

$$s = \sqrt{s^2}$$

- Standard deviation in same units as mean
  - Variance in units<sup>2</sup>

13

## Coefficient of Variation

- Relative variation rather than absolute variation such as standard deviation
- Definition of C.V.

$$C.V. = \frac{s}{\bar{x}} (100)$$

- Useful in comparing variation between two distributions
  - Used particularly in comparing laboratory measures to identify those determinations with more variation
  - Also used in QC analyses for comparing observers

14

## Grouped Data Frequency Distribution

- Grouping of continuous outcome
  - Better understanding of what data show rather than individual values
  - Primarily for description, not so much for analysis where individual values would be used
  - Examples: ages, heights, weights
- Number of intervals
  - Some "rules" exist, but generally create 5-10 equal sized intervals
- Report both frequency and relative frequency (percent of total)

15

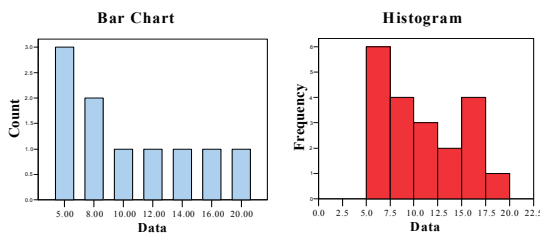
## Grouped Data Frequency Distribution

- Labeling of class limits
  - Rather arbitrary, but important to fully label limits on graphs and tables
    - I.e., 14.0 - 14.9, 15.0 - 15.9, etc.
    - Even if underlying variable is completely continuous, show break in labels
    - Can 'get away with it' for a histogram
  - Midpoints of class intervals are sometimes used

16

## Graphical Methods Bar Graphs and Histogram

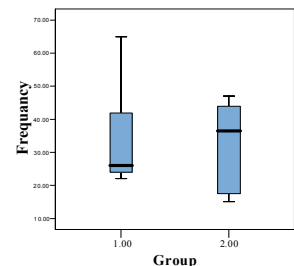
- Histogram graph of frequencies - special form of bar graph
  - Can be used to visually compare frequencies
  - Easier to assess magnitude of differences rather than trying to judge numbers



17

## Box Plots

- Descriptive method to convey information about measures of location and dispersion
  - Box-and-Whisker plots
- Construction of boxplot
  - Box is IQR
  - Line at median
  - Whiskers at smallest and largest observations



18

### Summary

- In practice, descriptive statistics play a major role
  - Always the first 1-2 tables/figures in a paper
  - Statistician needs to know about each variable before deciding how to analyze to answer research questions
- In any analysis, 90% of the effort goes into setting up the data
  - Descriptive statistics are part of that 90%

19

### Software

- Statistical software
  - SAS
  - SPSS
  - S-Plus or R
  - Statistica
  - MINITAB
  - Excel??

20