Statistical Techniques in Hospital Management
QUA 537

Dr. Mohammed Alahmed
Ph.D. in BioStatistics
alahmed@ksu.edu.sa
(011) 4674108

Course Description

- This course introduces biostatistical methods and applications, covering descriptive statistics, probability, and inferential techniques necessary for appropriate analysis and interpretation of data relevant to **health sciences**.
- Use the statistical software package (**SPSS**).

Course Objectives

- Familiarity with basic biostatistics terms.
- Ability to summarize data and do basic statistical analyses using SPSS.
- Ability to understand basis statistical analyses in published journals.
- Understanding of key concepts including statistical hypothesis testing – critical quantitative thinking.
- Foundation for more advance analyses.

Course Evaluation

• Assignments and attendance	15%
• Midterm exam	25%
• Project	20%
• Final exam	40%

Course Contents

1. Descriptive statistics
2. Introduction to the SPSS Interface
3. Probability and Probability distributional
4. One-sample inference
5. Two-sample inference
6. Analysis of Variance, ANOVA
7. Non Parametric methods
8. Chi-Square Test
9. Regression and Correlation analysis

Introduction: Some Basic concepts

What is Biostatistics ?

- A **portmanteau** word made from biology and statistics.
- The application of **statistics** to a wide range of topics in **biology**, particularly from the fields of **Medicine and Public Health**.

What is Statistics ?

Statistics is a field of study concerned with:

1. **Collection, organization, summarization and analysis of data.** (**Descriptive Statistics**)

1. **Drawing of inferences about a body of data when only a part of the data is observed.** (**Inferential Statistics**)

Statisticians try to interpret and communicate the results to others.

Descriptive Biostatistics

Methods of producing quantitative and qualitative summaries of information in public health:

- **Tabulation and graphical presentations.**
- **Measures of central tendency.**
- **Measures of dispersion.**

DATA

- The raw material of **Statistics** is data.
- We may define data as **figures**.
- Figures result from the process of **counting** or from taking a **measurement.**

For example:

- When a hospital administrator counts the number of patients **(counting).**
- When a nurse weighs a patient **(measurement)**

Sources of Data

Populations and Samples

Before we can determine what statistical tools and technique to use, we need to know if our information represents a **population** or a **sample**

A **sample** is a **subset** which should be **representative** of a population.

Types of Data or Variable

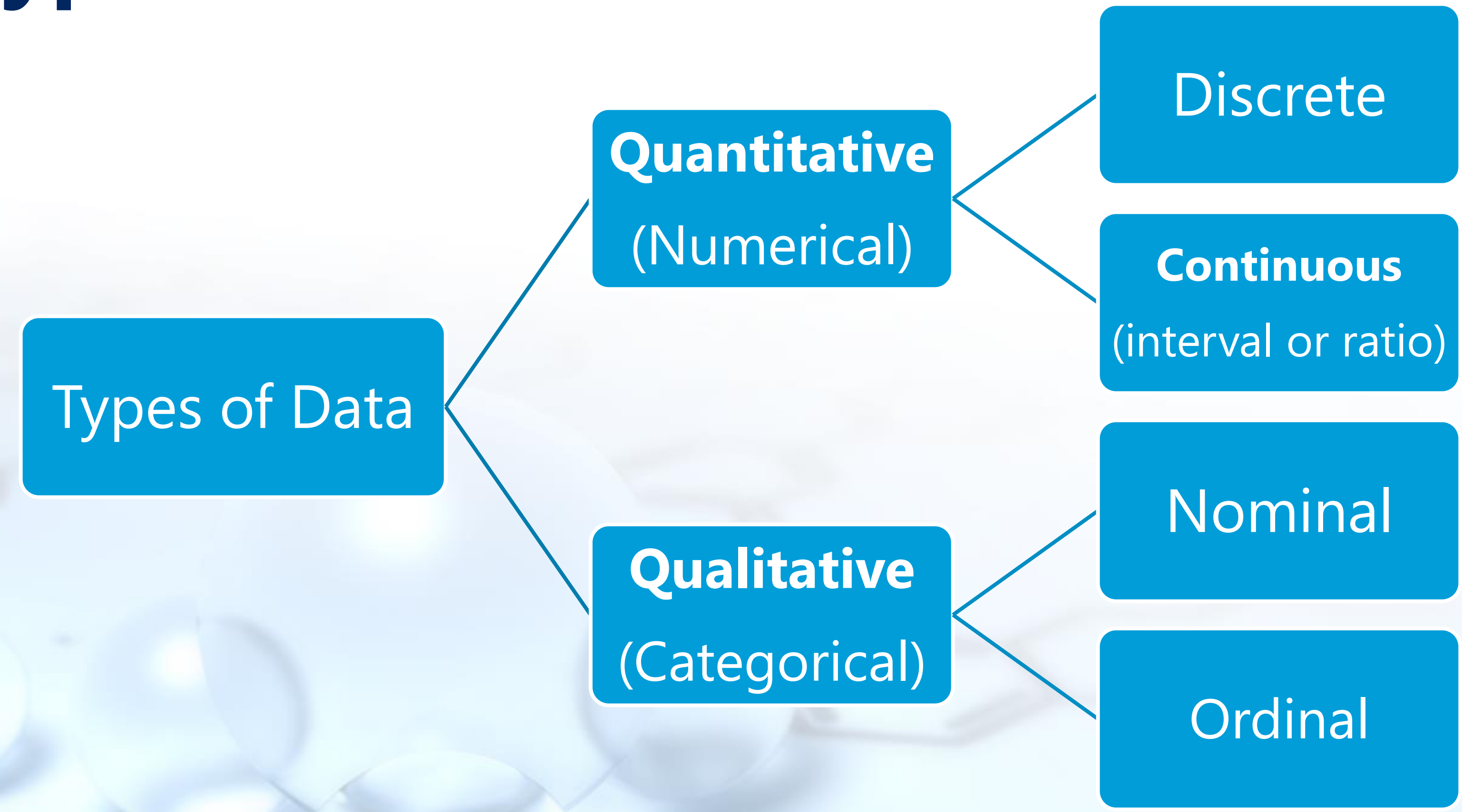
Data are made up of a set of variables.

A variable is a **characteristic** that takes on different **values** in different persons, places, or things.

For example:

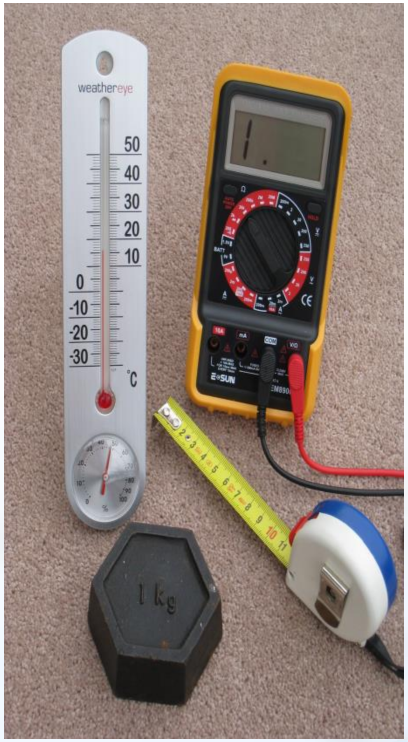
- Heart rate
- The heights of adult males
- The weights of preschool children
- The ages of patients

Types of Data or Variable



Scales of Measure

Scales	Description	Example
Nominal	qualitative classification of equal value	gender, race, color, city
Ordinal	qualitative classification which can be rank ordered	socioeconomic status of families, Education levels
Interval	Numerical or quantitative data can be rank ordered and sizes compared	temperature
Ratio	Quantitative interval data along with ratio. A ratio scale possesses a meaningful (unique and non-arbitrary) zero value	time, age.



Dr. Mohammed Alahmed

Methods of Data Presentation

- Tabulation Methods.
- Graphical Methods.
- Numerical Methods.

Tabulation Methods

Tabular presentation (simple – complex)

- **Simple frequency distribution Table**

Name of variable (Units of variable)	Frequency	%

- Categories		
Total		

- Distribution of 50 patients at the surgical department of King Khalid hospital in May 2013 according to their ABO blood groups

Blood group	Frequency	%
A	12	24
B	18	36
AB	5	10
O	15	30
Total	**50**	**100**

Frequency Distribution tables

Distribution of 50 patients at the surgical department according to their age.

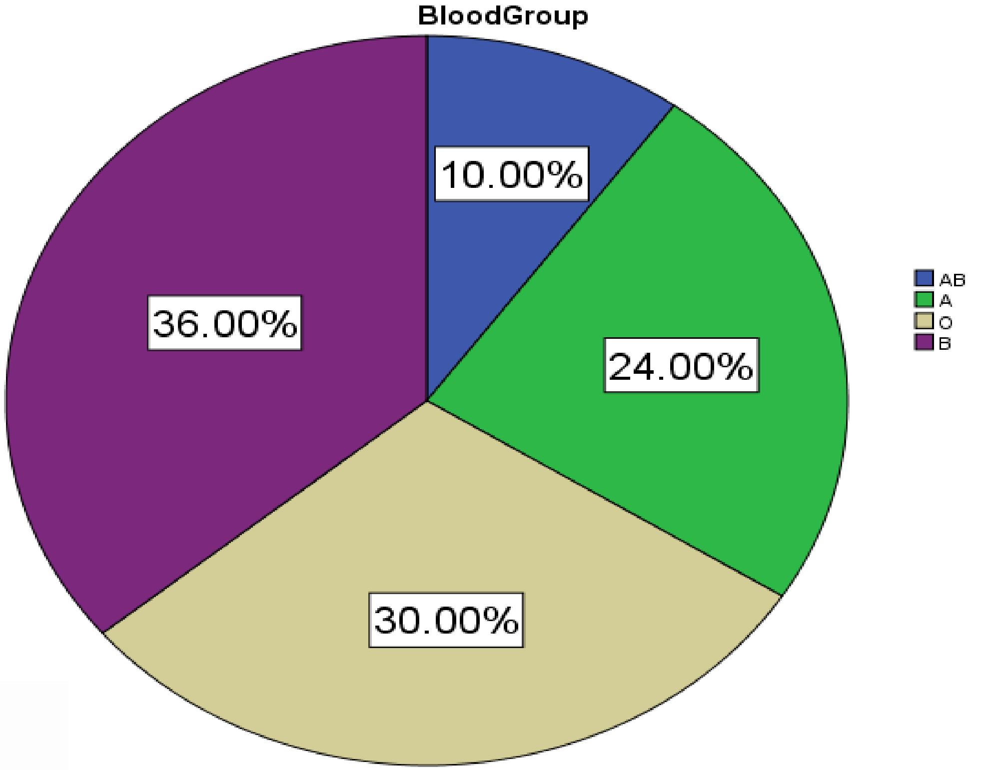
Age (years)	Frequency	%
20 -	10	20
30 -	14	28
40 -	18	36
50 -	8	16
Total	**50**	**100**

Complex frequency distribution Table

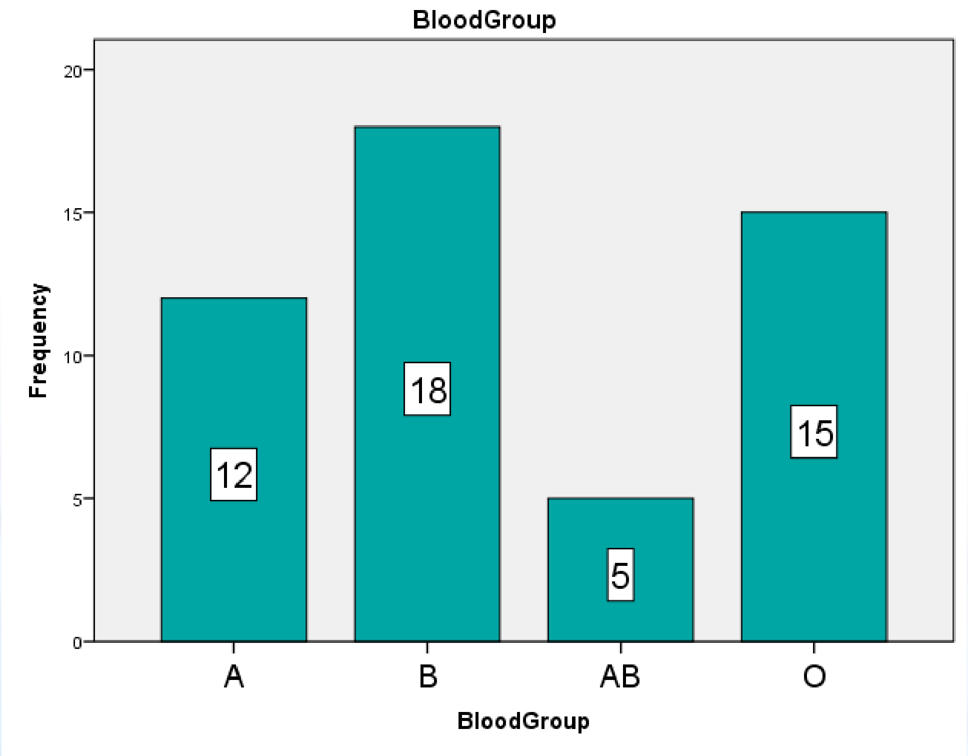
Smoking	Lung cancer				Total
	positive		negative		
	No.	%	No.	%	
Smoker	15	65.2	8	34.8	23
Non smoker	5	13.5	32	86.5	37
Total	20		40		60

Graphical Methods

- **Pie Chart**



• Bar Chart



• Two variables bar chart

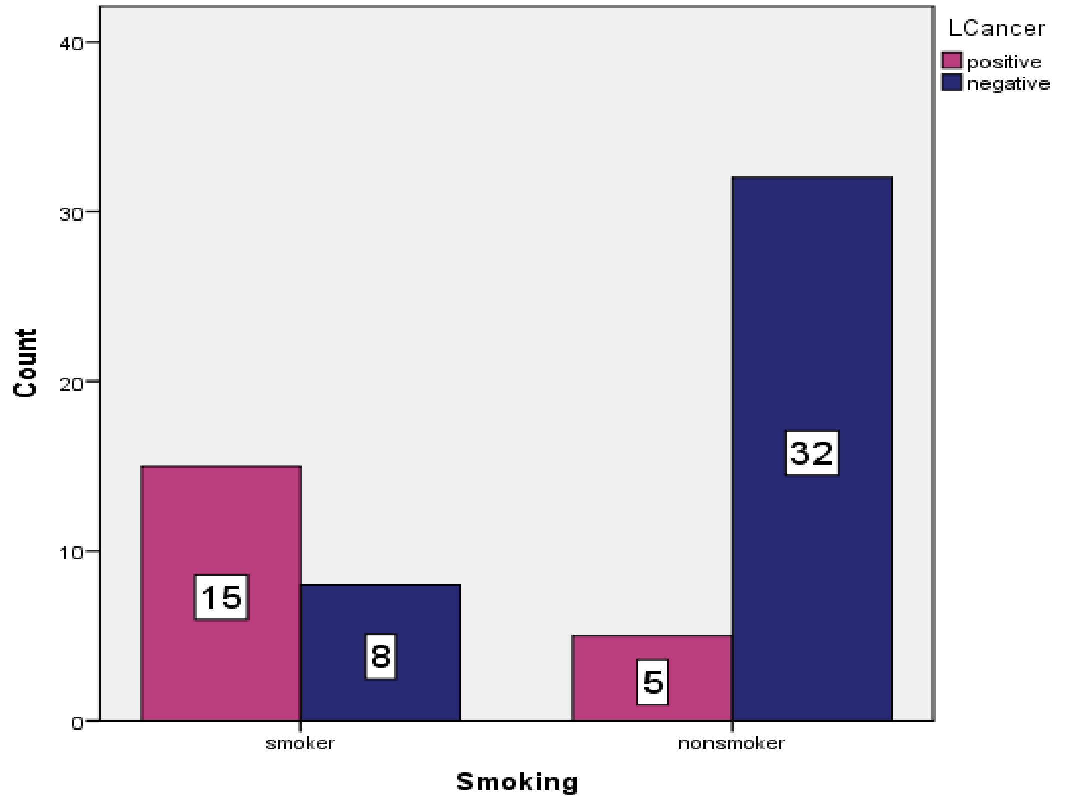
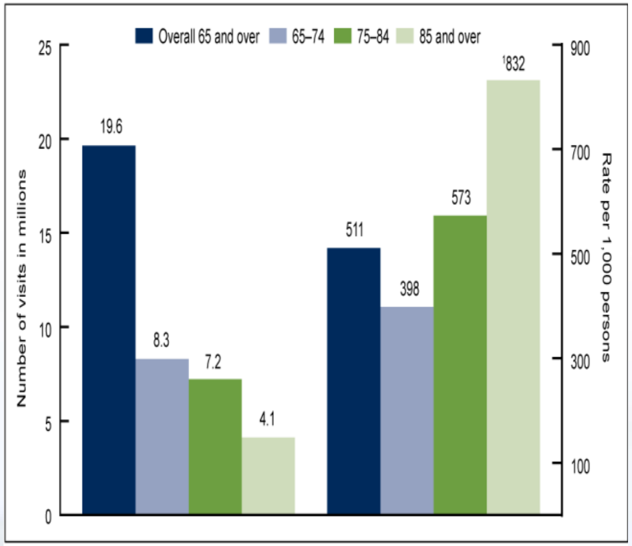
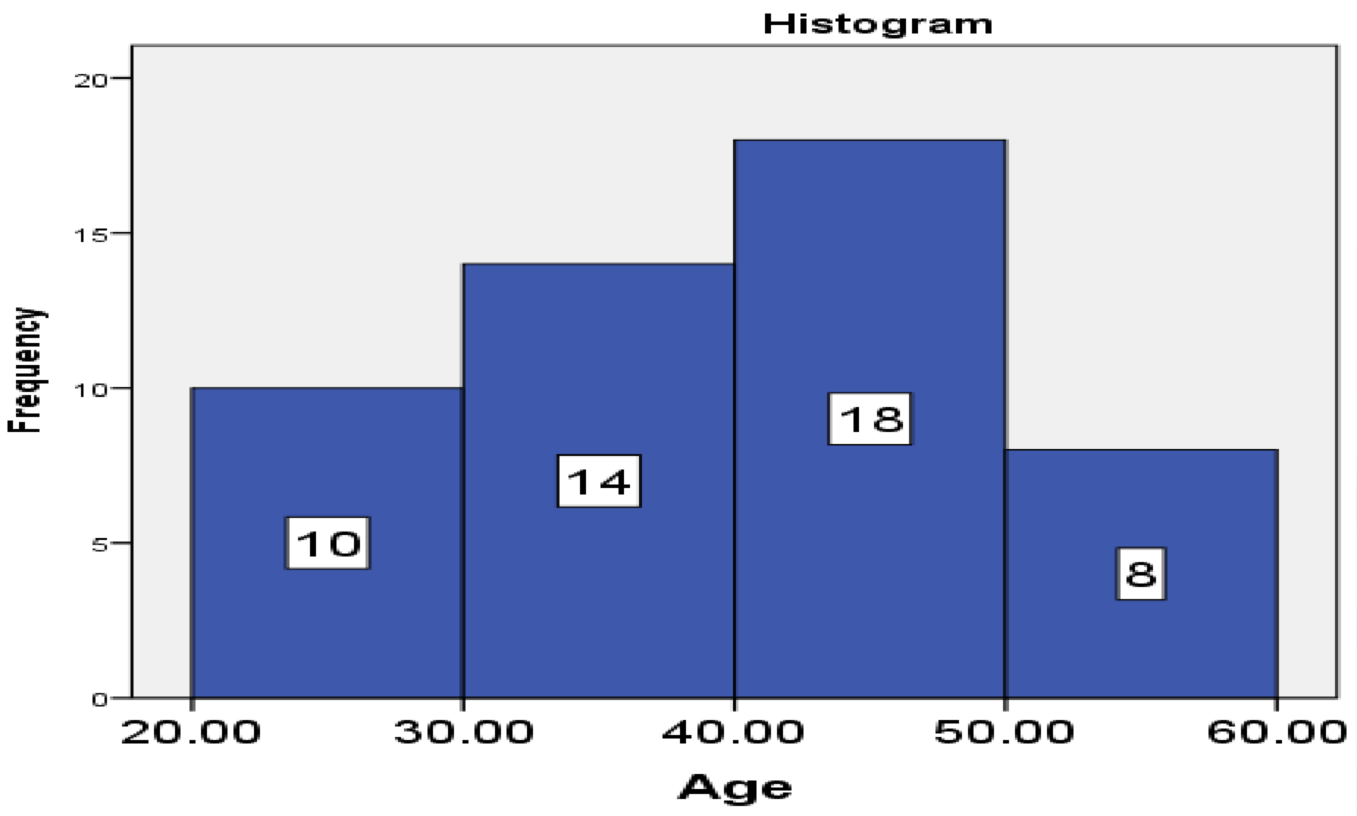


Figure 1. Number and rate of emergency department visits for persons aged 65 and over: United States, 2009–2010



[1]Linear trend shown is significant ($p < 0.05$) based on a weighted least-squares regression test.
NOTES: Figures are based on 2-year averages. A sample of 10,279 emergency department visits were made by persons aged 65 and over, representing an annual average weighted total of 19.6 million visits. Visit rates are based on the July 1, 2009, and July 1, 2010, set of estimates of the civilian noninstitutionalized population of the United States as developed by the U.S. Census Bureau's Population Division.
SOURCE: CDC/NCHS, National Hospital Ambulatory Medical Care Survey, 2009–2010.

- # Histogram



Histogram
20
15
10
5
0
Frequency
10
14
18
8
20.00
30.00
40.00
50.00
60.00
Age

Stem-and-leaf plot

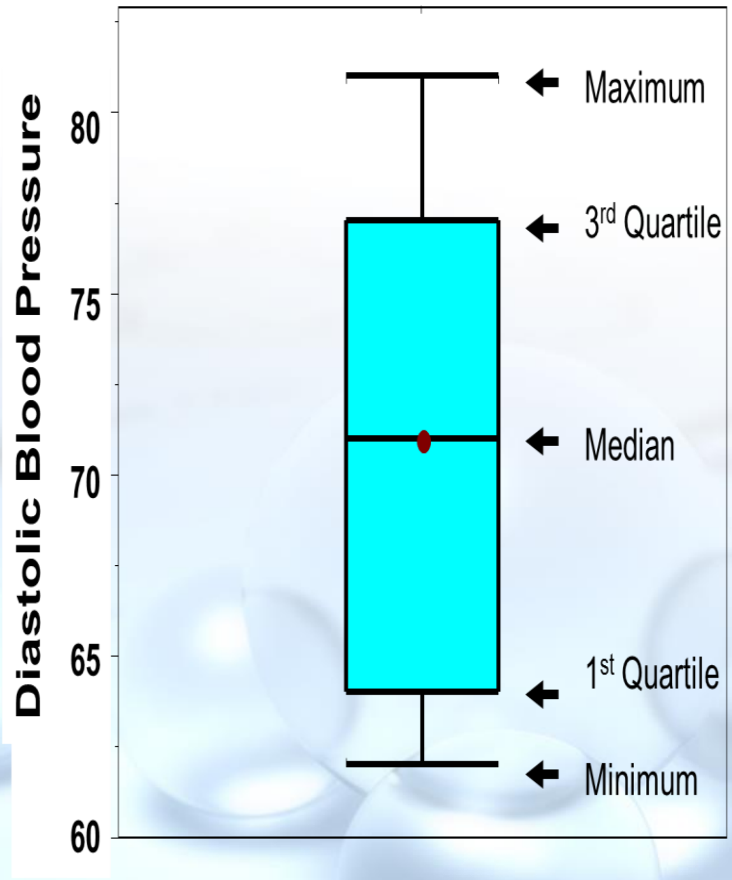
Stem-and-leaf of birthwgt N = 100
Leaf Unit = 1.0

1	3	2
1	4	
2	5	8
5	6	478
5	7	
15	8	3556788999
26	9	12344568889
45	10	0123444445567888899
(17)	11	00122235555556889
38	12	01112222344445567788
18	13	222334557888
6	14	0146
2	15	5
1	16	1

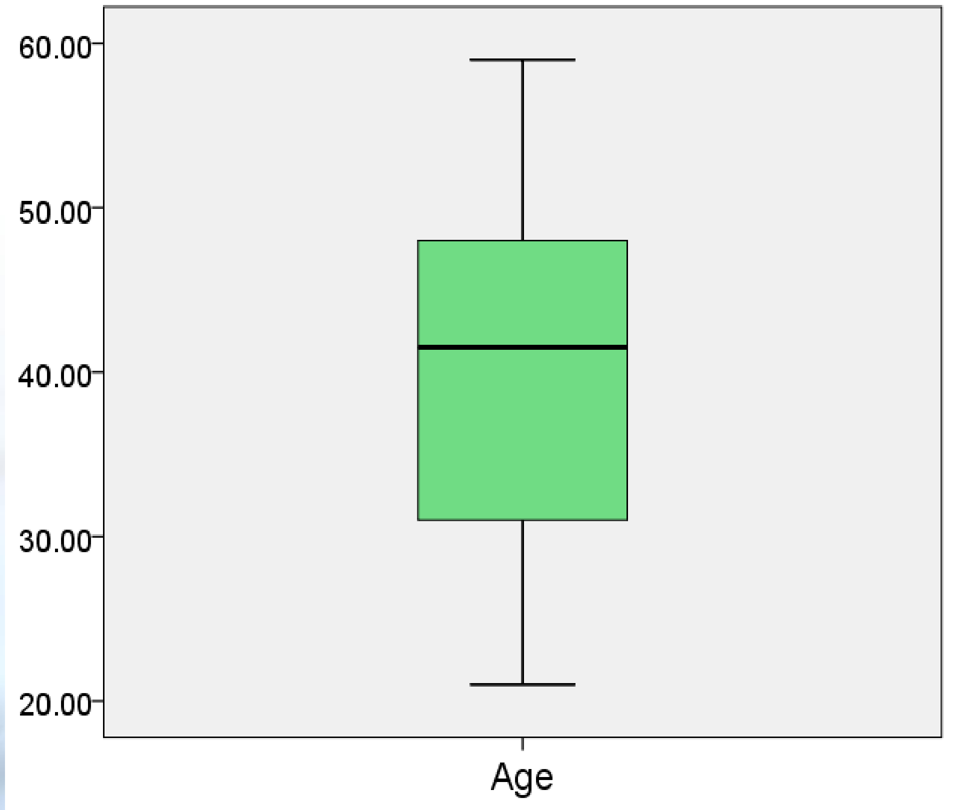
A stem-and-leaf plot can be constructed as follows:

1. Separate each data point into a stem component and a leaf component, respectively, where the stem component consists of the number formed by all but the rightmost digit of the number, and the leaf component consists of the rightmost digit. Thus the stem of the number 483 is 48, and the leaf is 3.
2. Write the smallest stem in the data set in the upper left-hand corner of the plot.
3. Write the second stem, which equals the first stem + 1, below the first stem.
4. Continue with step 3 until you reach the largest stem in the data set.
5. Draw a vertical bar to the right of the column of stems.
6. For each number in the data set, find the appropriate stem and write the leaf to the right of the vertical bar.

• Box plot

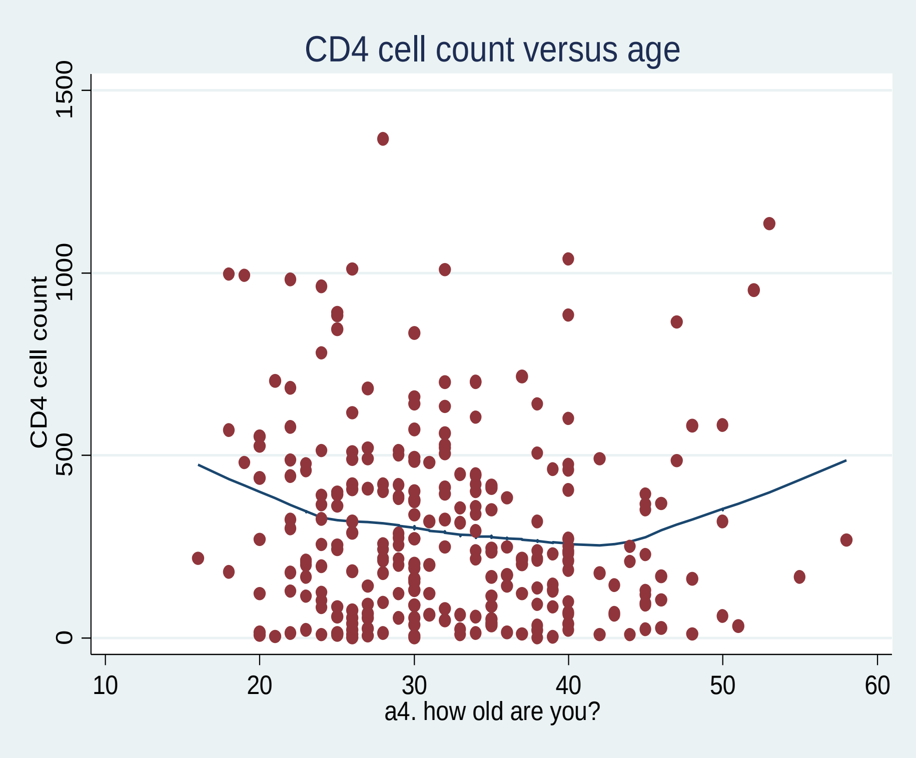


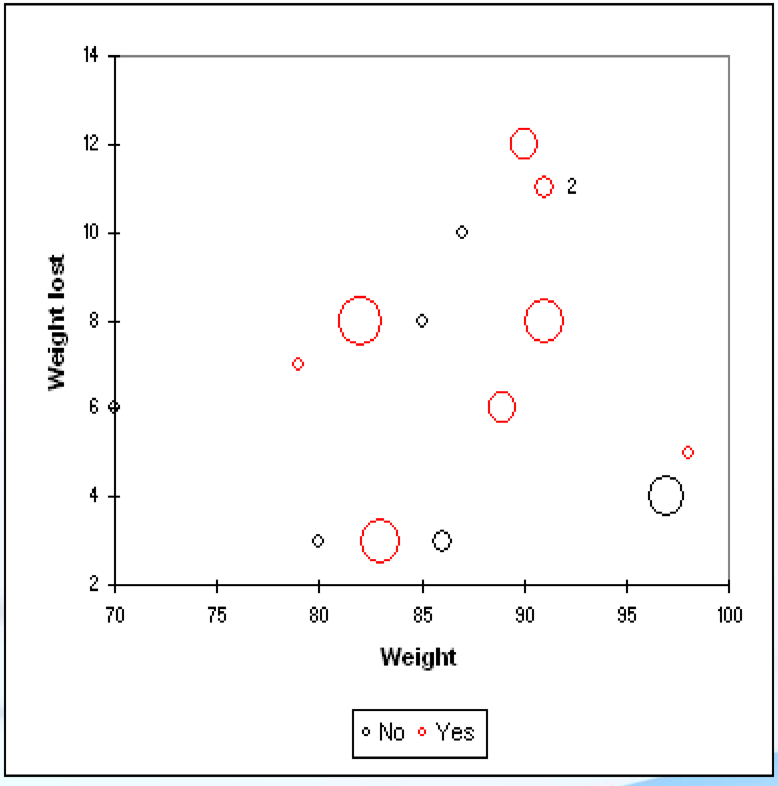
Diastolic Blood Pressure
80
75
70
65
60
Maximum
3rd Quartile
Median
1st Quartile
Minimum



60.00
50.00
40.00
30.00
20.00
Age

- # Scatter plots





General rules for designing graphs

- A graph should have a self-explanatory legend.
- A graph should help reader to understand data.
- Axis labeled, units of measurement indicated.
- Scales important. Start with zero (otherwise // break).

Numerical Methods

1. Measures of location.
2. Measures of dispersion.

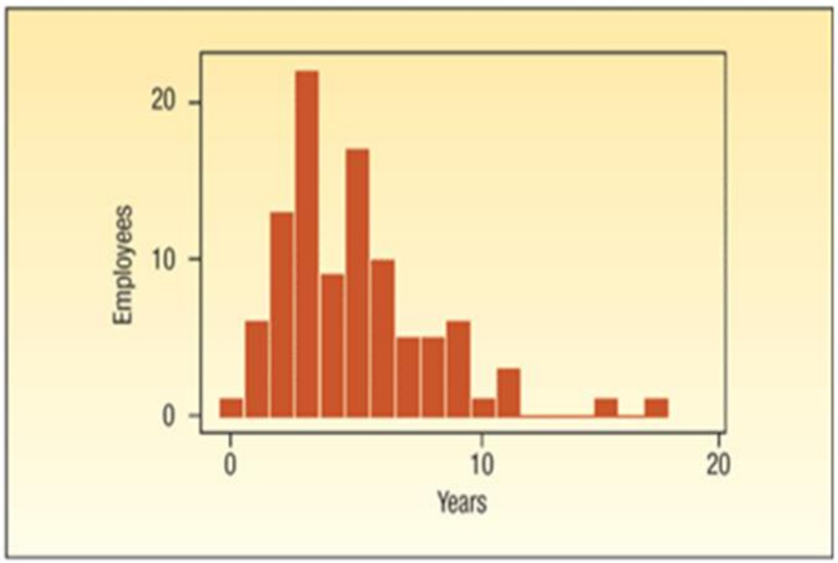


CHART 3–5 Histogram of Years of Employment at Hammond Iron Works, Inc.

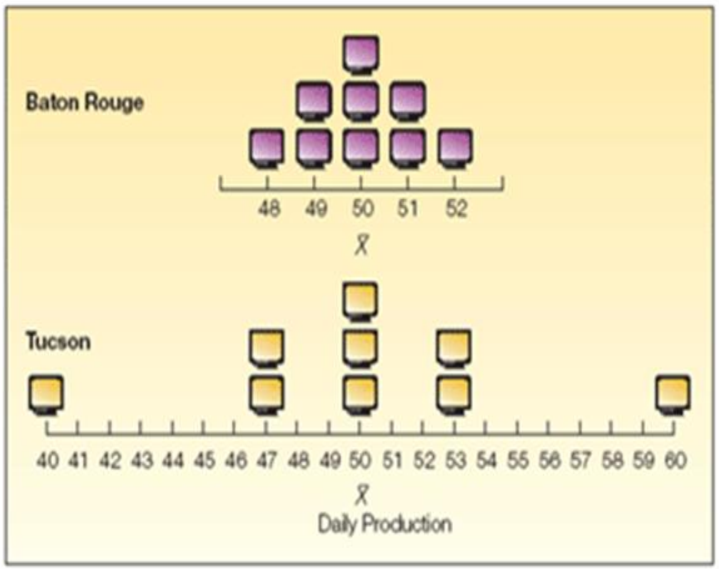


CHART 3–6 Hourly Production of Computer Monitors at the Baton Rouge and Tucson Plants

- You want to know the **average** because that gives you a sense of the **center** of the data, and you might want to know the low score and the high score because they give you a sense of how **spread** out or concentrated the data were.
- Those are the kinds of statistics this section discusses: **measures of central tendency** and **measures of dispersion**.
- Central tendency gets at the typical score on the variable, while dispersion gets at how much variety there is in the scores.

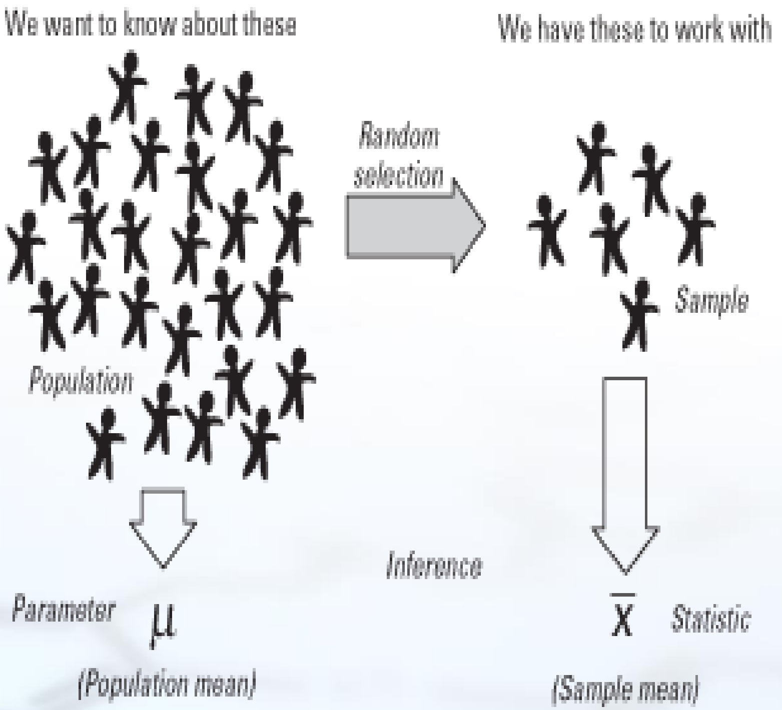
The Statistic and The Parameter

Statistic:

It is a descriptive measure computed from the data of a **sample**.

Parameter:

It is a descriptive measure computed from the data of a **population**.



Measures of location

Measures of central tendency – where is the center of the data?

1. **Mean (Average) - the preferred measure for interval data.**
2. **Median – the preferred measure for ordinal data.**
3. **Mode - the preferred measure for nominal data.**

The Arithmetic Mean

This is the most popular and useful measure of central location

$$\text{Mean} = \frac{\text{Sum of the observations}}{\text{Number of observations}}$$

Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example

The following data consists of white blood counts taken on admission of all patients entering a small hospital on a given day.

7, 35, 5, 9, 8, 3, 10, 12, 8

Compute the mean (average) blood count.

Mean = 97/ 9 = 10.78

The Median

The Median of a set of observations is the value that falls in the **middle** when the observations are arranged in order of magnitude.

Odd number of observations

$$\frac{n+1}{2}$$

Even number of observations

$$\frac{n}{2}, \frac{n}{2}+1$$

Example

Compute the **median** blood count.

- Order data (from the smallest to the largest):

  3, 5, 7, 8, 8, 9, 10, 12, 35

  Median = 8

- If you have even number:

  3, 5, 7, 8, 8, 9, 10, 12, 20, 35

  Median = (8+9)/2 = 8.5

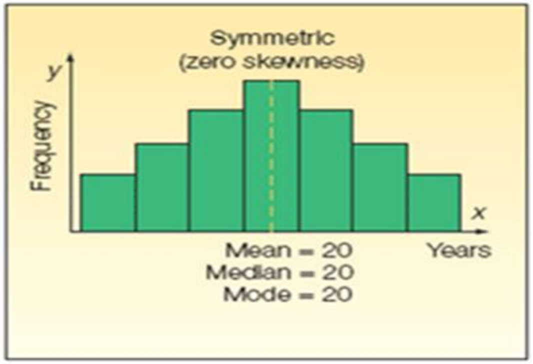
The Mode

The Mode of a set of observations is the value that **occurs most frequently**.

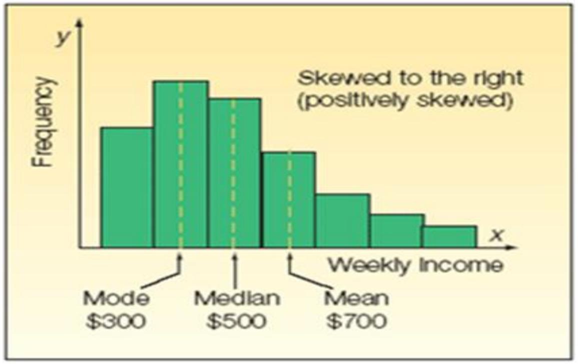
- Set of data may have one mode (or modal class), or two or more modes, or no mode!

What is the mode of the blood count?

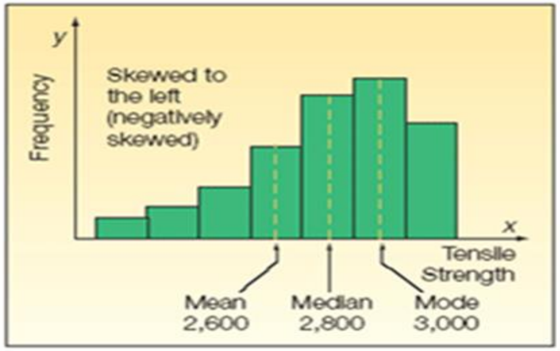
Relationship among Mean, Median, and Mode



zero skewness
mode = median = mean



positive skewness
mode < median < mean



negative skewness
mode > median > mean

Measures of dispersion

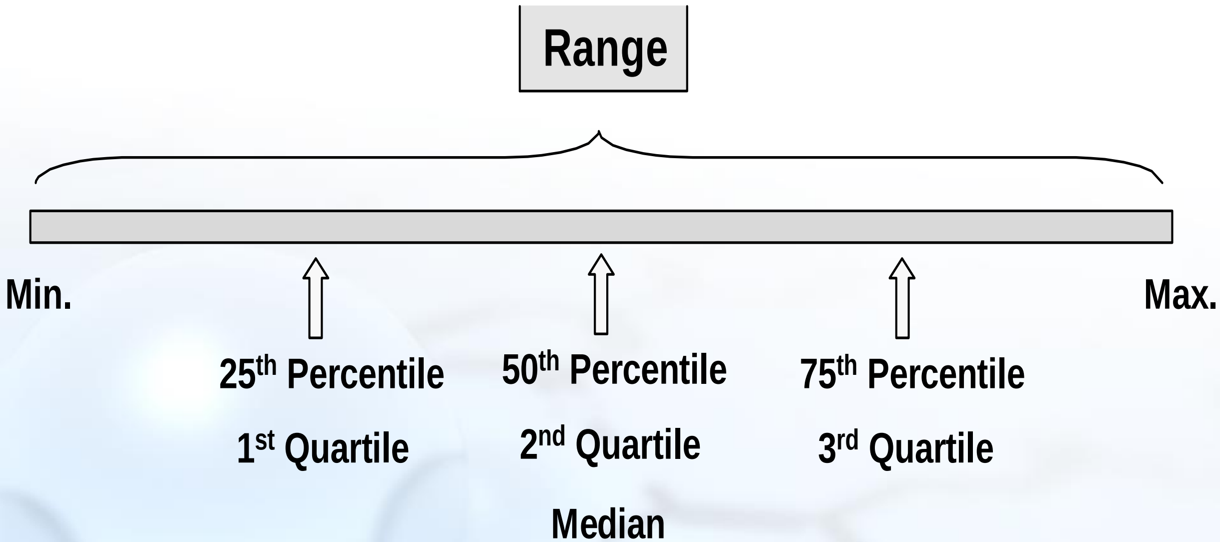
- Measures of central location fail to tell the whole story about the distribution.
- A question of interest still remains unanswered

How much are the observations spread out around the mean value?

1. **Range**
2. **Interquartile Range**
3. **Variance and Standard Deviation**

The Range

Range = Largest value - Smallest value



For example the range of the blood count is given by:

Rang = 35 – 3 = 32

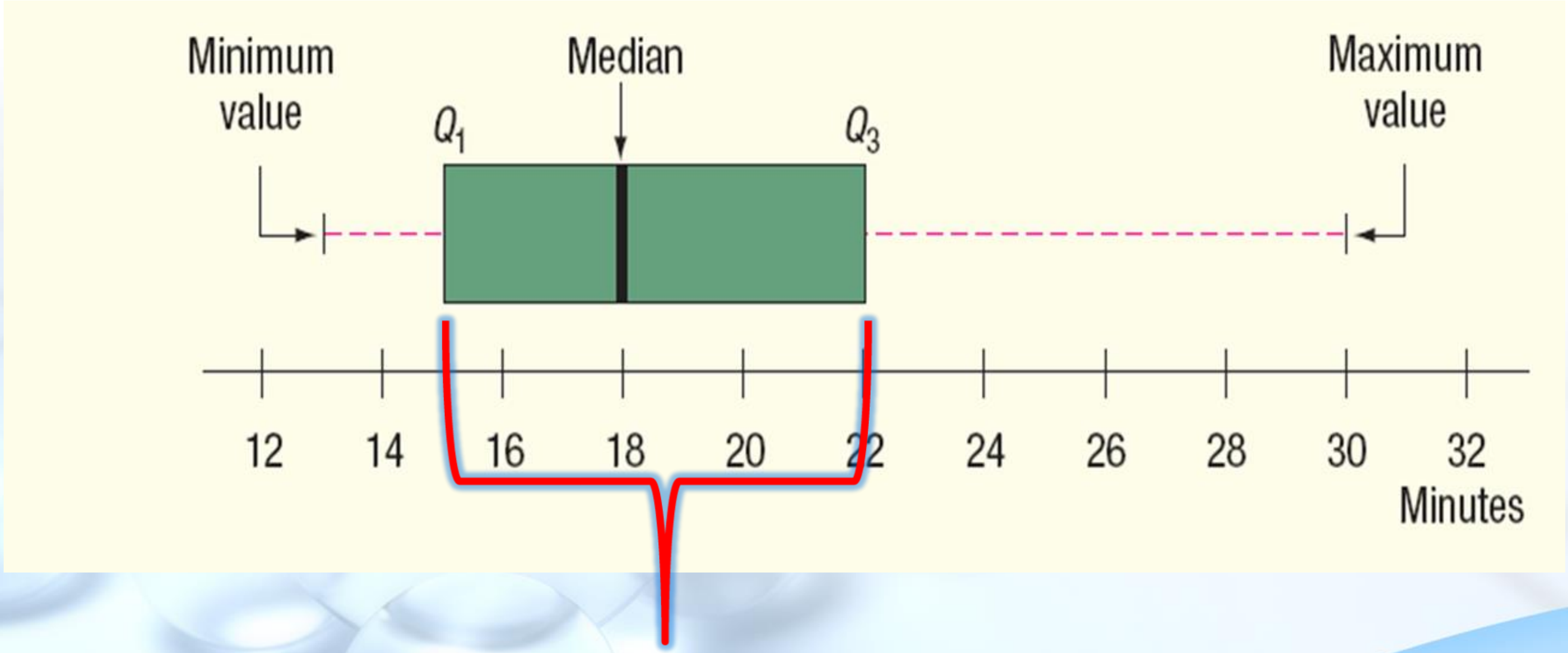
Quartiles and Percentiles

Let $\mathbf{L_p}$ refer to the location of a desired percentile. So if we wanted to find the **25_{th}** percentile we would use L_{25} and if we wanted the median, the **50_{th}** percentile, then **L_{50}.**

LOCATION OF A PERCENTILE

$$L_p = (n + 1)\frac{P}{100}$$

Boxplot Example



IQR = Q3 – Q1

The Variance and Standard Deviation

It measure dispersion relative to the scatter of the values a bout there mean.

- **Sample Variance (S^2) :**

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

The variance of white blood counts is given by:

$S^2 = 89.454$

- **Population Variance (σ^2)**

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

- **The Standard Deviation**
  - For the sample $S = \sqrt{S^2}$
  - For population $\sigma = \sqrt{\sigma^2}$

Why do we need both '**central tendency**' and '**dispersion**' to describe a numerical variable?

Example (age)

A		B	
11, 12, 13, 14, 15, 16, 17, 18, 19	Mean = 15.0 SD = 2.7	7, 9, 11, 13, 15, 17, 19, 21, 23	Mean = 15.0 SD = 5.5

The Coefficient of Variation

- For the same relative spread around a mean, the variance will be larger for a larger mean.
- Can be used to compare variability across measurements that are on a different scale.

$$CV = \frac{s}{\bar{x}} * 100\%$$