

201110 chapter 1

Assoc.Prof Phisanu Chiawkhun

Dept. of Statistics, CMU.

Introduction

- Frequently, we see the following types of statements in everyday life :
- The national Association of Realtors in USA reported that the median selling price for a house was \$222,600 (The Wall Street Journal,2007)
- A Jupiter Media survey found 31% of adult males watch television 10 or more hours a week. For adult women it was 26% (The Wall Street Journal,2004)
- The Dow Jones Industrial closed at 13,265 (Barron's May,2007)

What is the Statistics?

- **Statistics** is defined as the art and science of collecting , analyzing , presenting and interpreting data. Particularly in business and economics engineering information technology the information provided by **collecting , analyzing , presenting and interpreting data** gives manager and decision makers informed and better decisions.

Application of Statistics

- In Today's The most successful managers and decision makers understand the information and know how to use it effectively. Some examples
- Finance , Marketing Production Economics.

What is the data ?

- Data are facts and figures collected , analyzed, and summarized for presentation and interpretation
- Data set collected in a particular study
- Variable is a characteristic of interest for the elements

Scale of Measurement

- **1. Nominal scale** the data for a variable consist of labels or names use to identify an attribute of the element.
- **2. Ordial scale** the data exhibit the properties of nominal data and the order or rank of the the data is meaningful.

Scale of Measurement

- **3. Interval scale the data show the properties of ordinal and the interval between values is expressed in term of a fixed unit of measure. Interval data are always numeric. Interval data have no real zero .**

Scale of Measurement

- 4. Ratio scale the data have all properties of interval data and the ratio of two values is meaningful. The scale requires that a zero value be indicate that nothing exists for the variable at the zero point.

Type of Data

- **Quantitative and Qualitative data**
- Quantitative data require numeric values that indicate how much how many. Obtained using either the interval or ratio scale of measurement.
- Qualitative data use either the nominal or ordinal scale of measurement and may be nonnumeric or numeric.

Descriptive Statistics

- The data that are summarized and presented in a form that is easy for the reader to understand. Such summarized data , which may be tabular, graphical, or numerical , are referred to as descriptive statistics.

Example

Exchange	Frequency	Percent frequency
New York Stock exchange	20	80
Nasdaq National Market	5	20
Totals	25	100

Statistical Inference

- Population is the collection of all the elements of interest.
- Sample is a subset of the population.
- Census is a survey to collect the data for the entire population.

Statistical Inference

- **Statistics use the data from sample to make estimates and test hypothesis about the characteristics of a population referred to Statistical Inference.**

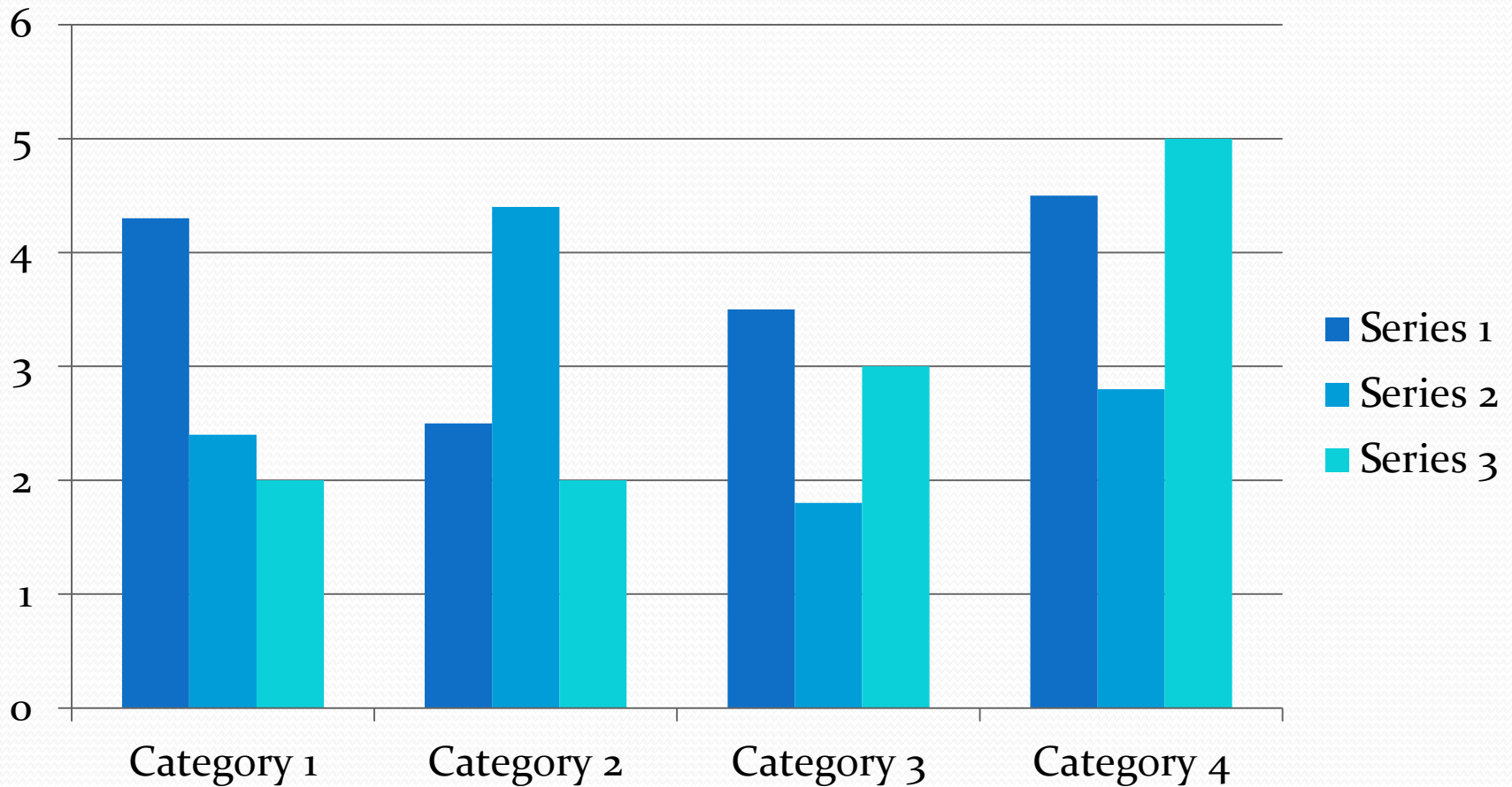
Data summarization

- **Frequency Distribution is a tabular summary of data showing the number (frequency) of items in each of nonoverlapping classes.**
- **Relative Frequency = Frequency of the class/n**
(n = the number of data)

Bar Graphs and Pie Charts

- **Bar graph or bar chart** , is a graphical device for depicting qualitative data summarized in a frequency.

Example

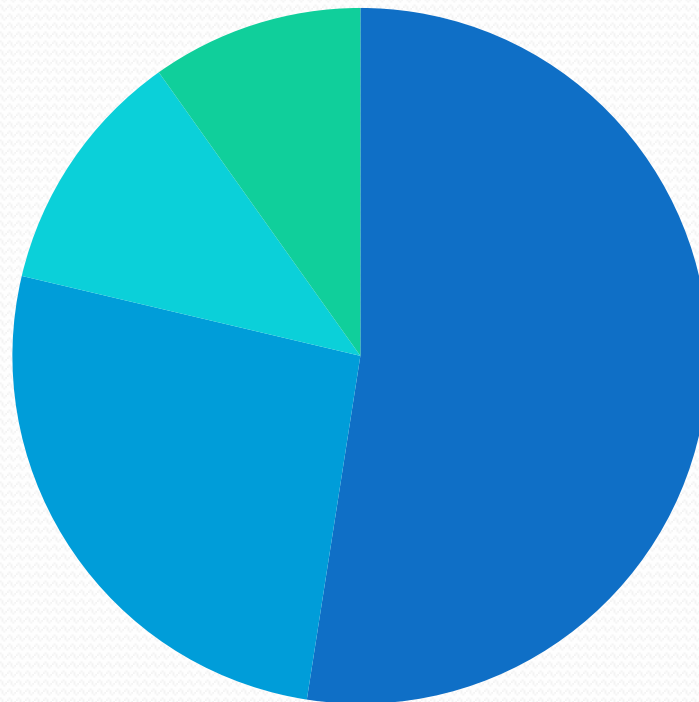


Pie Chart

- The pie chart provides another graphical device for presenting relative frequency and percent frequency distributions for qualitative data.

Example

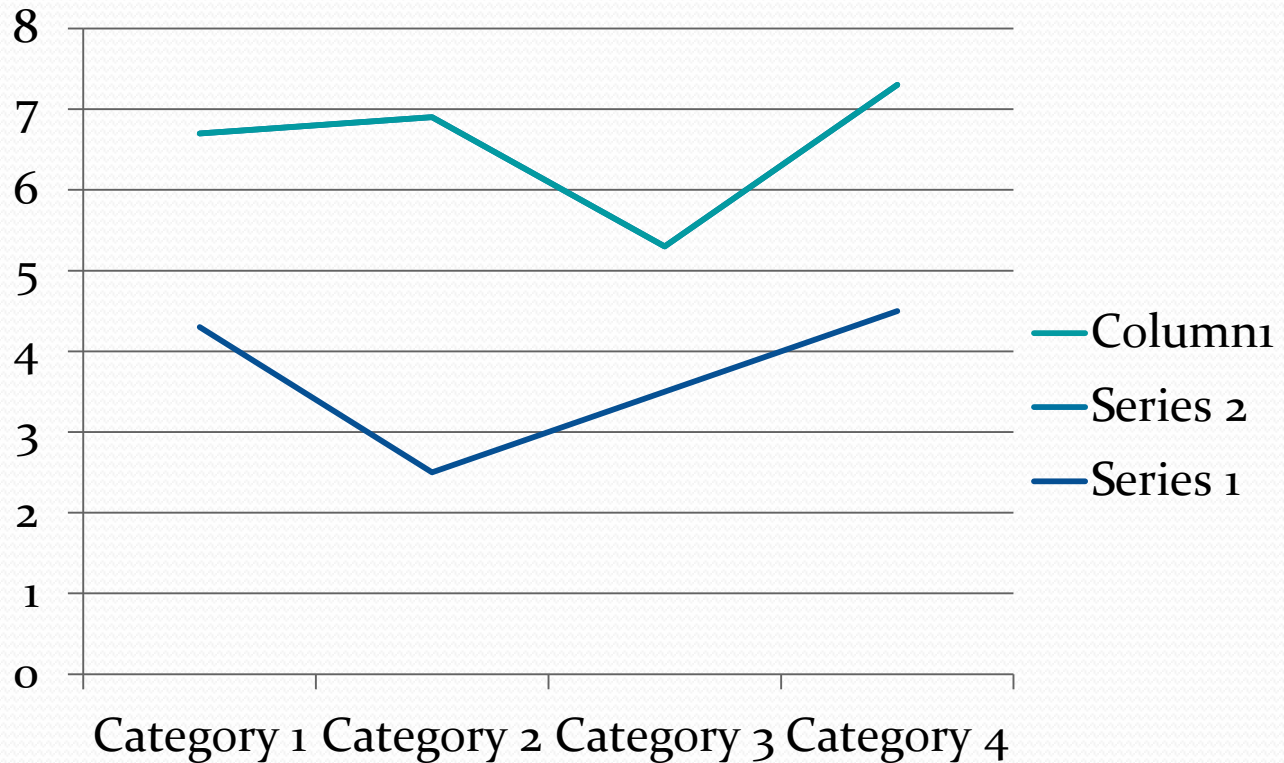
Sales



- 1st Qtr
- 2nd Qtr
- 3rd Qtr
- 4th Qtr

Line Chart

- Line chart shows the movement of data.



Summarizing Quantitative data

- Frequency Distribution is a tabular summary of data

Consider the audit time data (days)

12 14 19 18 15 15 18 17 20 27

22 21 33 28 14 18 16 13 22 23

We would like to construct the frequency distribution of these data

Width of the class = $\text{Max-Min/Number of classes}$

If the number of classes = 5

Width = $33-12/5 = 4.2 (=5)$

Frequency distribution

Audit Time (Days)	Frequency
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1
Total	20

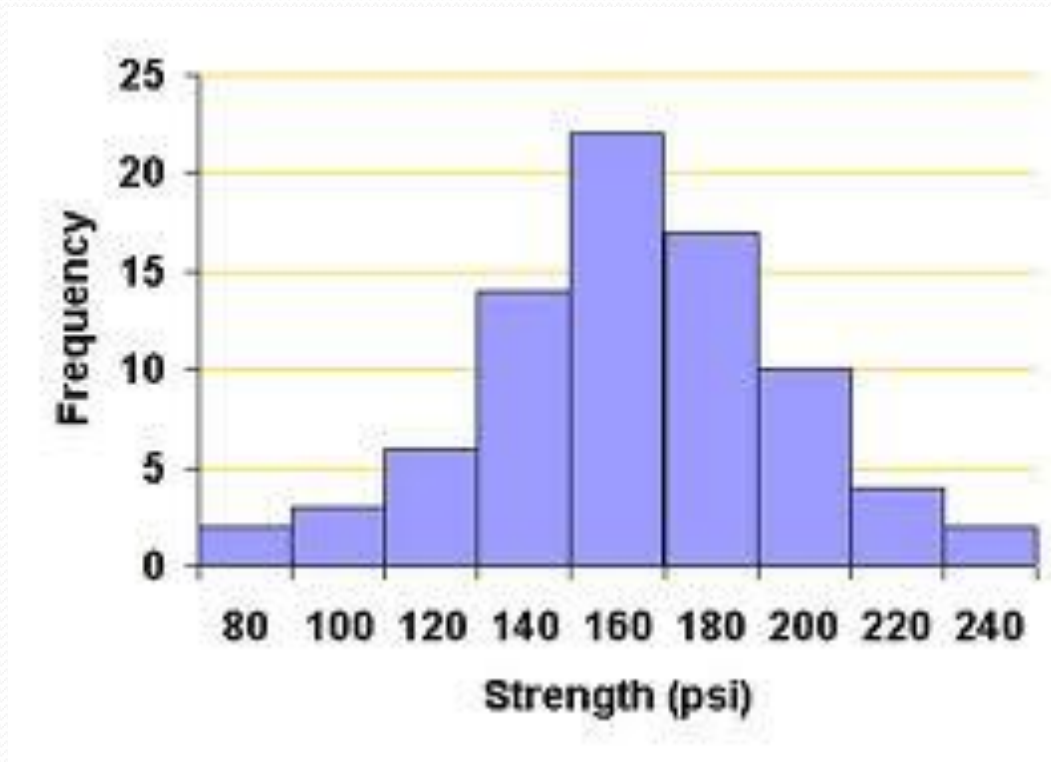
What we learn about frequency distribution ?

- Histogram

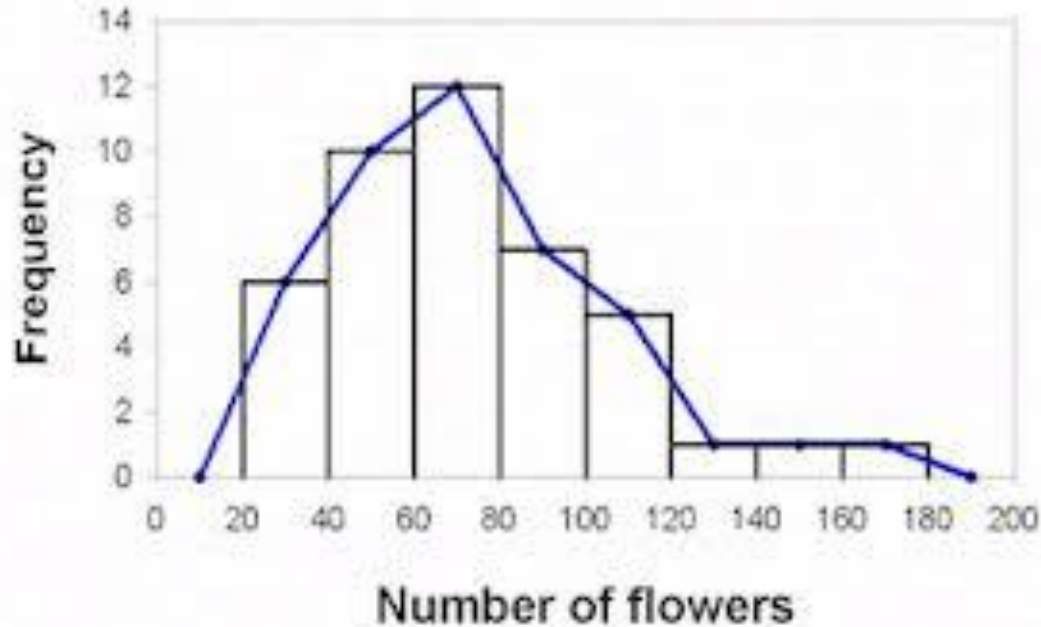
A histogram is constructed by placing the variable of interest on the horizontal axis and the frequency on the vertical axis.

- Frequency polygon is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoint of the classes . The frequencies are represented by the heights of the point

Histogram



Frequency polygon



The Stem-and-Leaf Display

- To develop a stem-and-leaf display, we first arrange the leading digits of each data value to the left of a vertical line. To the right of the vertical line, we record the last digit for each data value.
- The stem-and-leaf plot is a data that used part of the data values as the stem and part of the data value as the leaf to form groups or classes.
- Example data consist of 112 72 69 97 and 107
a stem-and-leaf would be

The Stem-and-Leaf Display

- Although the stem-and-leaf display may appear to offer the same information as a histogram ,it has two primary advantages.
 - 1. The stem-and-leaf display is easier to construct by hand.
 - 2. Within a class interval , the stem-and-leaf display provide more information than the histogram because the stem-and-leaf shows the actual data.

2. Measures of Central Tendency

- A Statistics is a characteristic or measure obtained by using data values from a sample.
- A parameter is a characteristic or measure obtained by using all the data values from a specific population.
- The mean

The mean ,also known as the arithmetic average is found by adding the values of the data dividing by the total number of values.

Mean

- The symbol of mean is \bar{X} or μ

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N}$$

In statistics Greek letters are used to denote parameter

The Roman letters are used to denote statistics

Mean

- If the data showed for grouped data ,we use the midpoints of the classes to calculate the mean

$$\bar{X} = \frac{\sum f_i x_i}{n}$$

x_i = the midpoints of each classes

f_i = the frequencies of each classes

Median

- The median is the midpoint of the data array.
- Step in computing the median of a data array
 - Step 1 Arrange the data in order.
 - Step 2 Select the middle point.
 - Step 3 The midpoint is the median

Mode

- The value that occurs most often in a data set is called the mode
- A data set that has only one value that occurs with the greatest frequency is said to be **unimodal**.
- If A data set that has two value that occurs with the same greatest frequency is said to be **bimodal**.
- If A data set that has more than two value that occurs with the same greatest frequency is said to be **multimodal**.

The weighted mean

- The mean of a data set in which **not all values are equally represented**

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$$

w_i = the weights of x_i

Properties and uses of Central Tendency

- The mean

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics.
4. The mean of the data set is unique.
5. The mean can not be computed for the data in a frequency that has an open-ended class.

Properties and uses of Central Tendency

- The mean

6. The mean is affected by extremely high or low values (**Outliers**)

The median

1. The median is used to find the center or middle value of the data set.
2. The median is used for an open-ended distribution
3. The median is affected less than the mean by extremely high or low values.

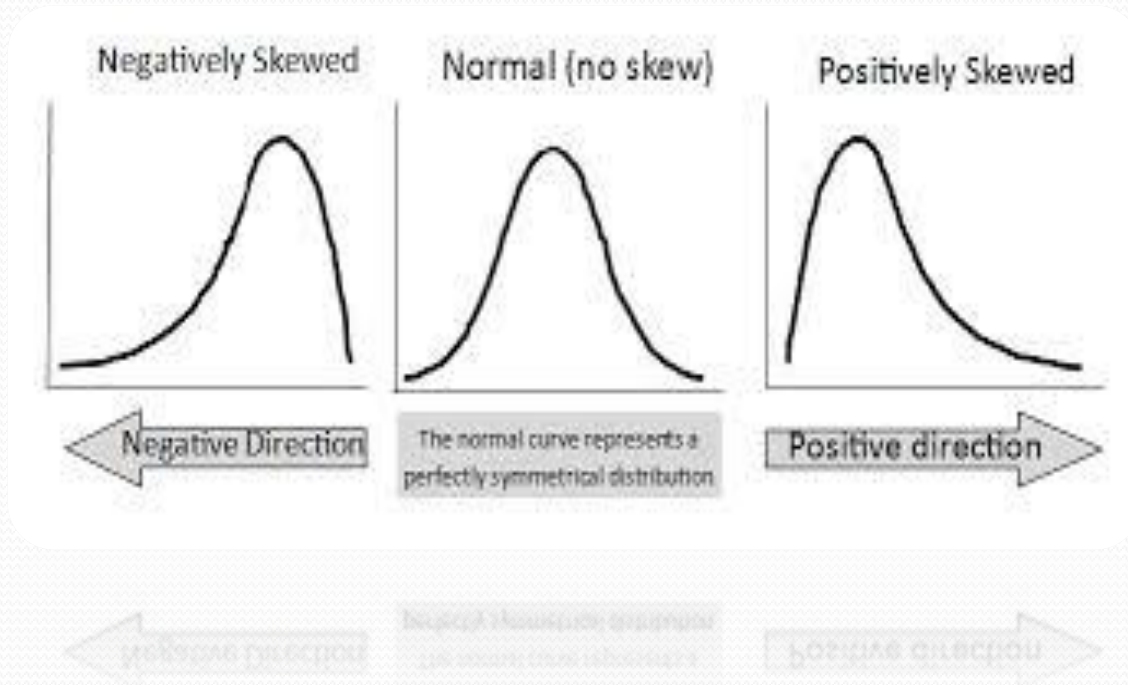
Properties and uses of Central Tendency

- The mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal or categorical, such as religious gender etc.
4. The mode is not always unique.

Distribution shaped

- The three most important shapes are positively skewed, systematic, and negatively skewed.



Measures of Variation

- The main objective of measures of variation is to measure the the spread or variability of a data set.
- Three measures are commonly used : **range, variance and standard variation**

Range

Range is the highest value minus the lowest value.

The symbol R is used for range.

$$R = \text{Max} - \text{Min}$$

The variance

- The variance is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

For sample variance

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

The standard deviation

- The standard deviation is the square root of the variance σ

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

For sample standard deviation

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

Computational Formulas for S² and S

- The shortcut formulas for computing the variance and standard deviation for the data obtained from sample are as follows

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

and standard deviation

$$s = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}$$

Uses of the variance and Standard Deviation

- 1. Variance and standard deviation can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed.
- 2. The variance and standard deviation are used quite often in inferential statistics.

Coefficient of Variation : CV

- Karl Pearson (English Statistician) devised the coefficient of variation to compare the deviations of two different groups such as the height of men and women .

$$CV = \frac{\sigma}{\mu} \quad \text{for populations}$$

$$CV = \frac{s}{\bar{x}} \quad \text{for samples}$$