Basic Biostatistics

Dr. Ei Ei Zar Nyi

Assistant Director

Central Epidemiology Unit

Statistics

- Statistics is a field of study concerned with

(1) collection, organization, summarization and analysis of data and

(2) the drawing of inferences about a body of data when only a part of the data is observed.

Biostatistics (Biomedical statistics)

- When the data analyzed are derived from the biological sciences and medicine, we use the term biostatistics.

Synonym = Medical statistics

Uses

Biostatistics is necessary for

- To measure the status of health and disease in a community.
- Provide the basic not only to monitor the health status of the community but also for the scientific advancement of medicine.
- For the collection, analysis, and interpretation of scientific data gathered from clinical, laboratory or field investigation.
- Clear thinking and sound understanding of statistical methods is fundamental for the research project.

Descriptive Statistics

Descriptive Statistics

- Descriptive Statistics are Used by Researchers to Report on Populations and Samples.

- Descriptive Statistics are a means of organizing and summarizing observations, which provide us with an overview of the general features of the set of data.

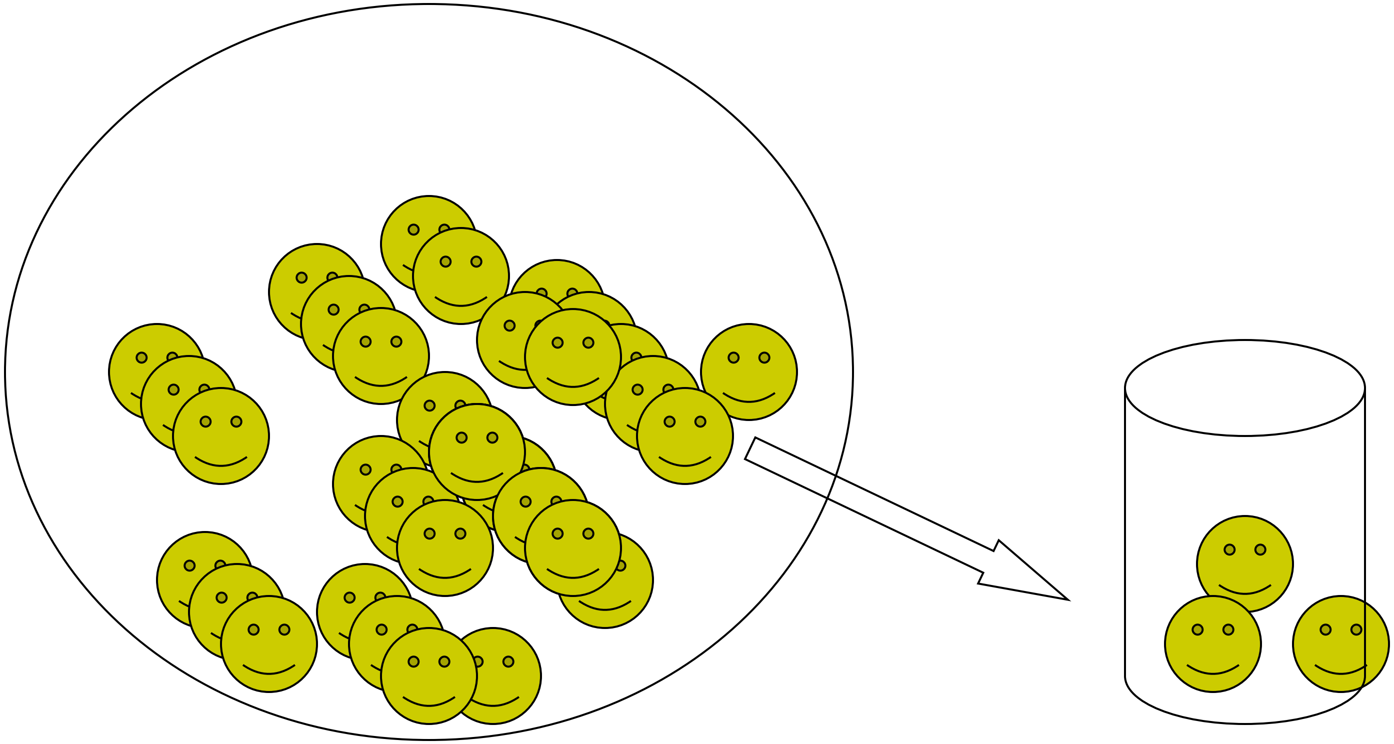
- **Raw data**:

  Measurements which have not been organized, summarized, or otherwise manipulated

- **Descriptive measures**:

  Single numbers calculated from organized and summarized data to describe these data. eg. Percentage, average

Sample vs. Population



	Population	**Sample**
Data	Parameter	statistic
Sample size	N	n
Mean	μ	x
Variance	σ^2	s^2
SD	σ	s

Descriptive Statistics

An Illustration:

Which Group is Smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.

Descriptive Statistics

Which group is smarter now?

Class A--Average IQ	Class B--Average IQ
110.54	110.23

They’ re roughly the same!

With a summary descriptive statistic, it is much easier to answer our question.

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs

- Summarize Data
  - Central Tendency
  - Variation

Descriptive Statistics

Types of descriptive statistics: (Data Presentation)

- Organizing Data
  - Tables
    - Simple table
    - Frequency Distribution table
    - Contingency table
    - Correlation table
  - Graphs
    - Bar Chart
    - Pie chart
    - Histogram
    - Frequency Polygon
    - Line diagram
    - Stem and Leaf Plot
    - Box Plots

Simple Table

Table (1) Population of some states in country X

State	Population
State A	5,000
State B	70,000
State C	30,000
State D	150,000

Source: Census of country X, 2000

Frequency Distribution table

Table (2) Age distribution of study population

Age group	Frequency	Percentage
0-4 years	15	30
5-9 years	20	40
10-14 years	5	10
≥15 years	10	20
Total	50	100

Contingency table

Table (3) Association between Sex and smoking status among study population

Sex	Smoking +	Smoking -	Total
Male	80	6	86
Female	70	4	74
Total	**150**	**10**	**160**

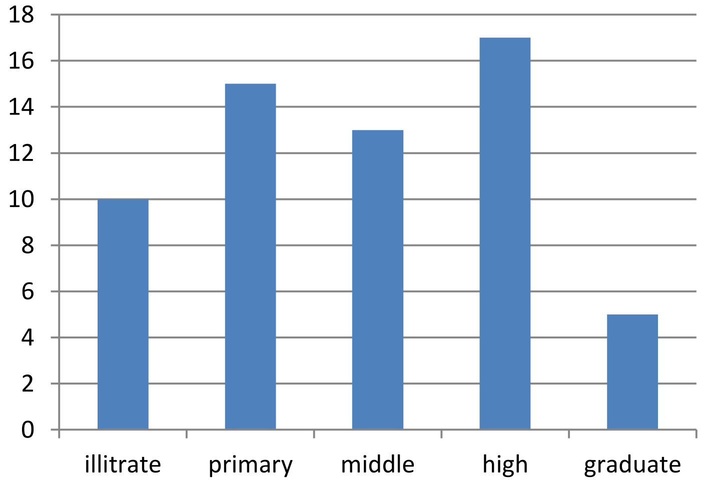
Correlation table

Age	Weight
1 month	6 lbs
2 months	10 lbs
3 months	14 lbs

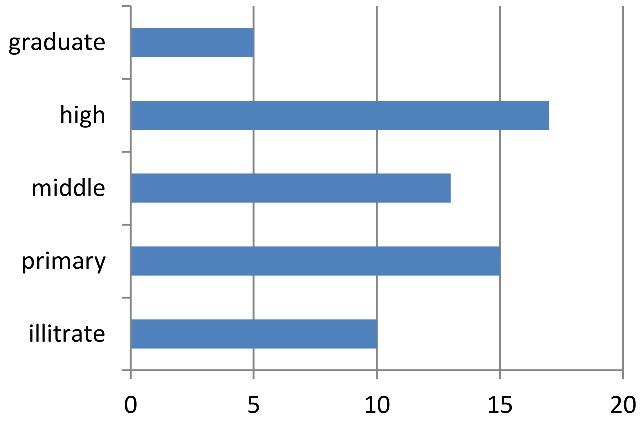
Bar chart

- Consist a set of vertical or horizontal bars
- Same width
- Height of each bar represent the frequency of each specific category
- Equal space between bars
- Purpose of the use of bar chat is to compare the categories of the same variable

Simple vertical bar chart

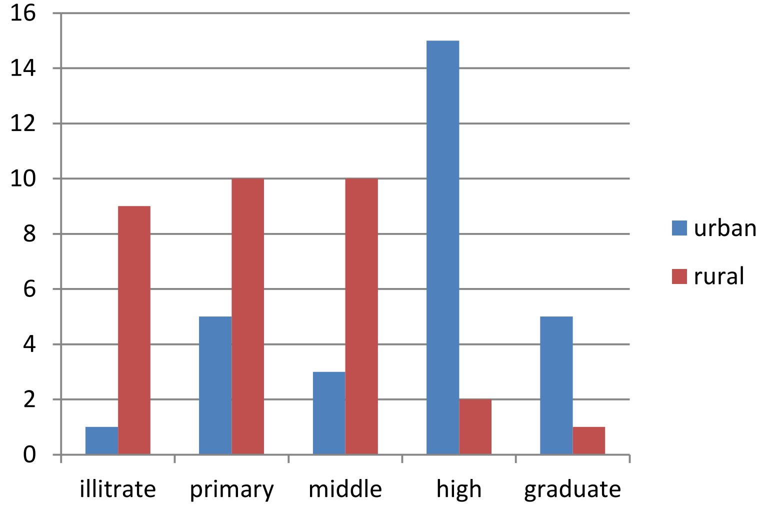


Simple horizontal bar chart

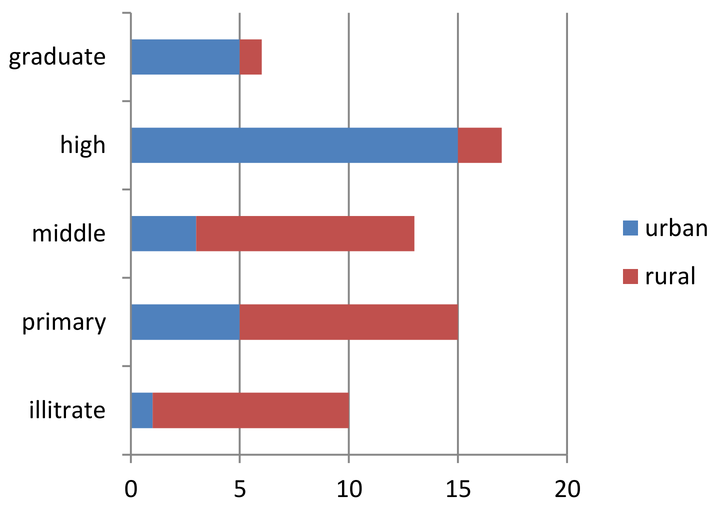


Bar chart

Multiple bar chart

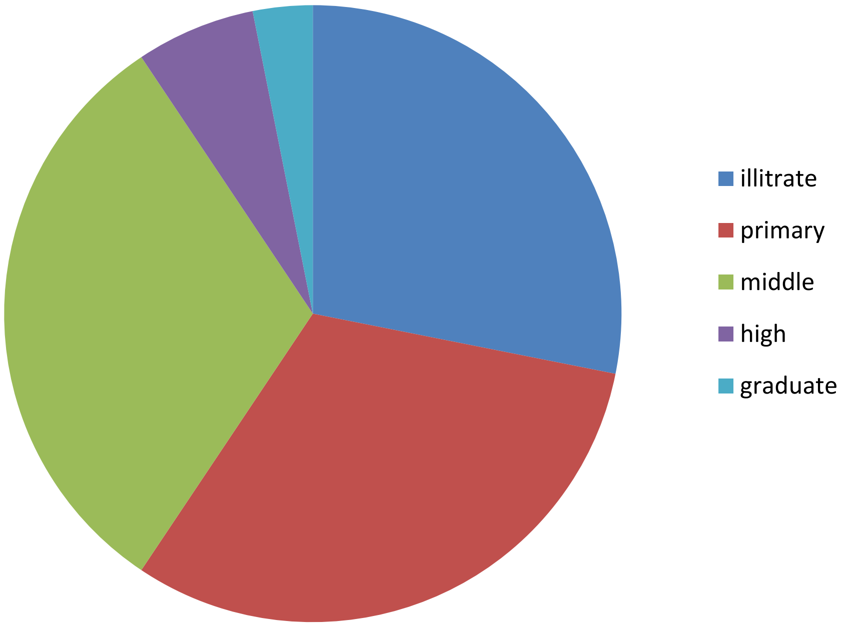


Component bar chart



Pie chart

- A circle containing 360 degrees
- Pie chart is the best adapted for illustrating the problem of hoe, the whole is sub-divided into segments
- Segments can be colored or shaded differently for greater clarity



Histogram

- A special type of bar graph showing frequency distribution
- It consists of a set of columns with no space between each of them
- The area under each column represents the frequency of each class
- If the data have been grouped into unevenly spaced intervals, a histogram is the most suitable kind of diagram

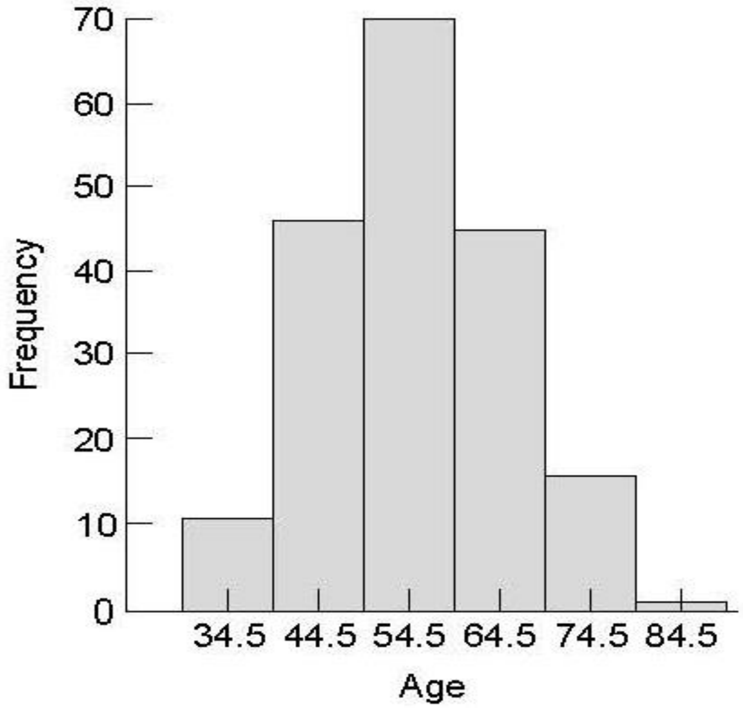


FIGURE 2.3.2 Histogram of ages of 189 subjects from Table 2.3.1.

Frequency Polygon

- A special kind of line graph connecting midpoints at the tops of bars or cells of histogram
- Total area under the frequency polygon is equal to that of histogram

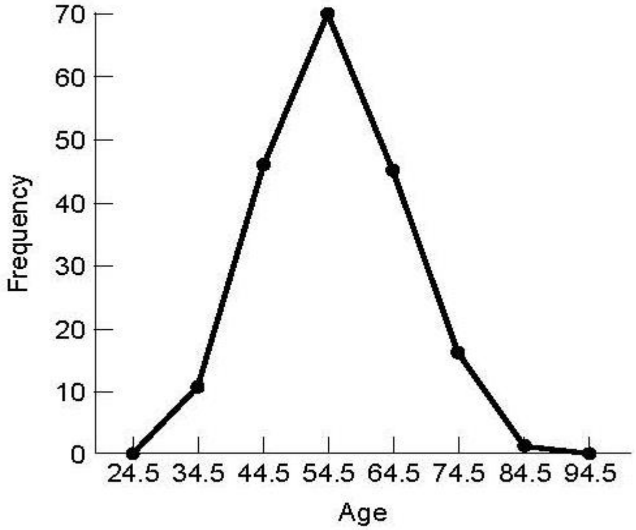


FIGURE 2.3.4 Frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

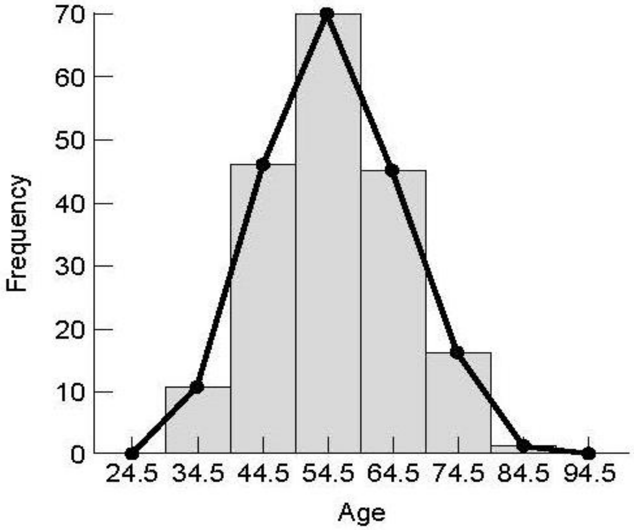
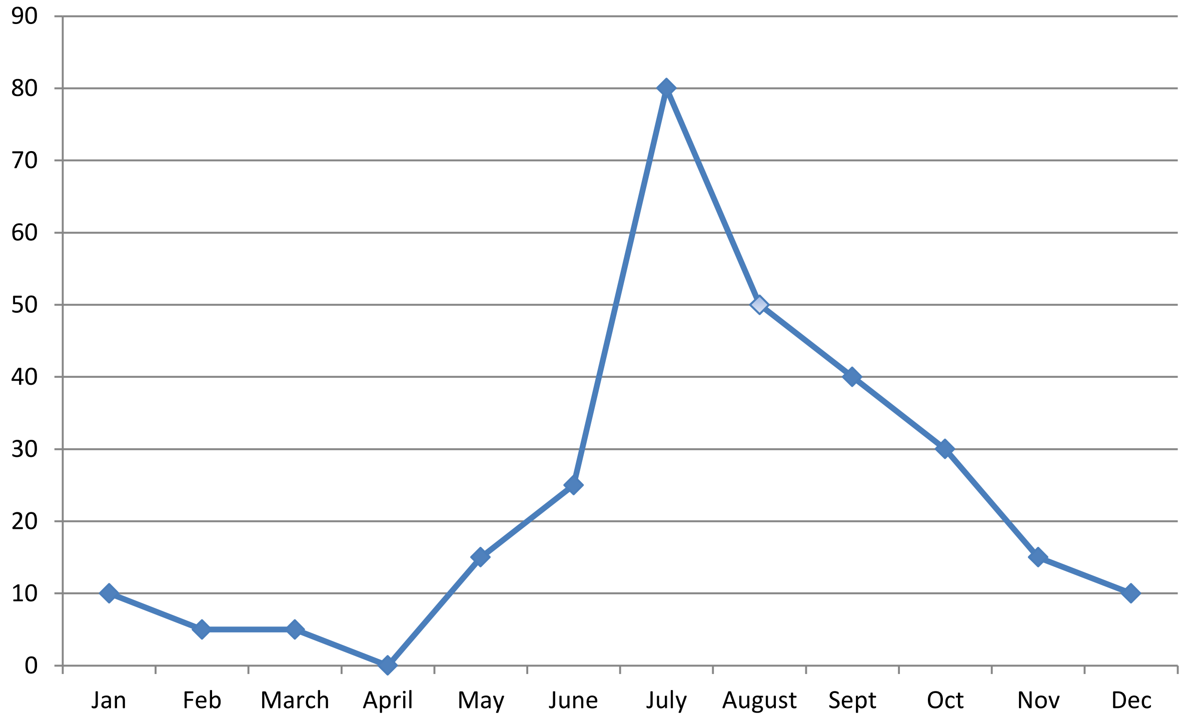


FIGURE 2.3.5 Histogram and frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

Line Diagram

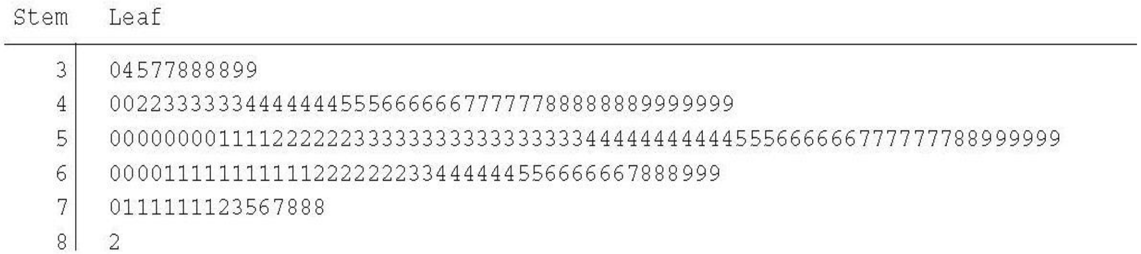
- Most commonly used for showing changes of values with the passage of time



DHF incidence during 2007, in X hospital

Stem and leaf plot

- It resembles with the histogram and has the same purpose (range of data set, location of highest concentration of measurements, presence or absence of symmetry)



Stem	Leaf
3	04577888899
4	0022333333444444455566666677777788888889999999
5	0000000011112222223333333333333333344444444444555666666777777788999999
6	000011111111111222222233444444556666667888999
7	0111111123567888
8	2

FIGURE 2.3.6 Stem-and-leaf display of ages of 189 subjects shown in Table 2.2.1 (stem unit = 10, leaf unit = 1).

Box plot

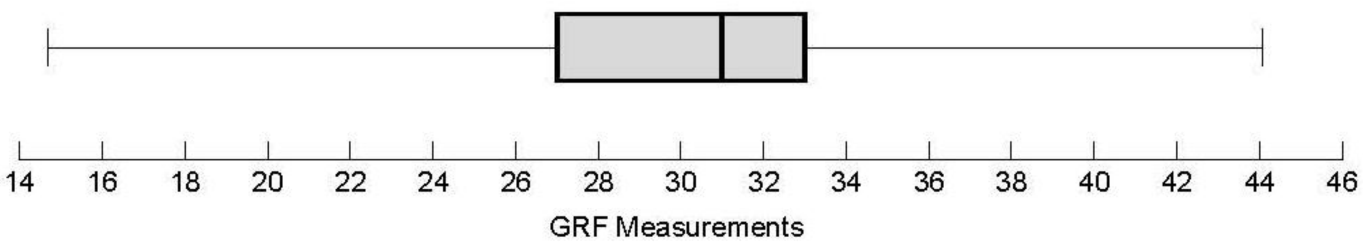


FIGURE 2.5.5 Box-and-whisker plot for Example 2.5.5.

Descriptive Statistics

- Summarizing Data
  - Central Tendency (or Groups' "Middle Values")
    - Mean
    - Median
    - Mode
  - Variation (or Summary of Differences Within Groups)
    - Range
    - Standard Deviation
    - Variance
    - Coefficient of variation

Measures of Central Tendency

- **Statistic** : A descriptive measure computed from the data of a sample

- **Parameter** : A descriptive measure computed from the data of a population

- Most commonly used measures of central tendency:

  Mean, Median, Mode

Mean

- also called 'Average'
- obtained by adding all the values in a population or a sample
- dividing by the number of values that are added

Formula of the mean : For a finite population: $\mu = \sum x_i / N$

: For a sample : $x = \sum x_i / n$

Eg. Mean age (year) of the following 9 subjects

56, 54, 61, 60, 54, 44, 49, 50, 63

$$x = \sum xi / n$$

$$= 56+54+61+60+54+44+49+50+63 / 9$$

$$= 54.55 \text{ year}$$

Properties of the mean

- Uniqueness
- Simplicity
- Being influenced by extreme values

Exercises Mean

Class A--IQs of 13 Students

102

128

131

98

140

93

110

115

109

89

106

119

97

Σ x = *1437*

x bar= $\frac{\Sigma\ x}{n} = \frac{1437}{13} = 110.54$

Class B--IQs of 13 Students

127

131

96

80

93

120

109

162

103

111

109

87

105

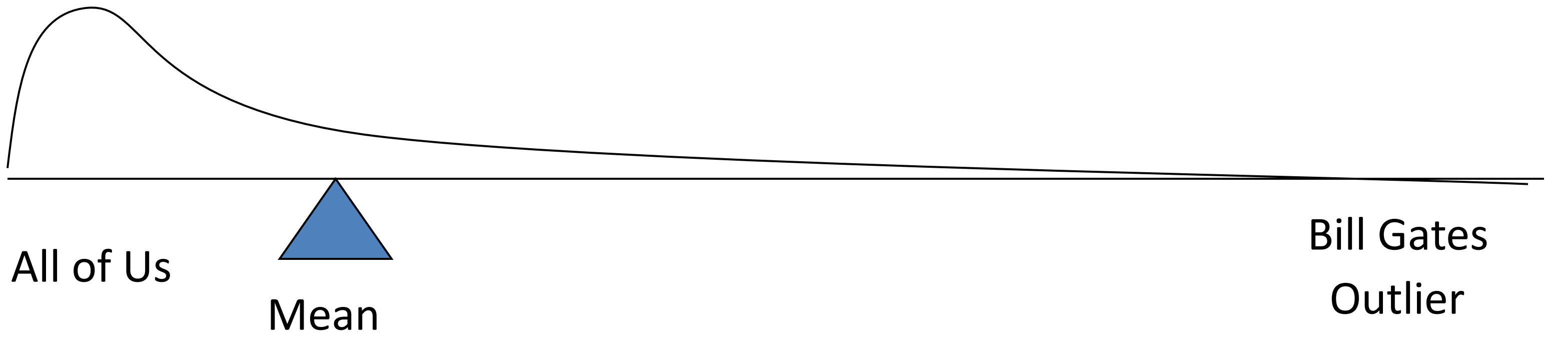
Σ x = *1433*

x bar = $\frac{\Sigma x}{n} = \frac{1433}{13} = 110.23$

Mean

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency or common experience

Income in the U.S.



Median

- The middle value of the data set which is arrayed from the lowest to the highest.
- 50% < Median > 50%
- For the series of odd numbers, median is the middle value.
- For even numbers, median is the average of two middle values.

Formula: **(n + 1) / 2 th value**

Properties of Median

- Uniqueness
- Simplicity
- Median can avoid the effect of skewed distribution

eg. Median age (year) of the following 9 subjects

56, 54, 61, 60, 54, 44, 49, 50, 63

Ordered array → 44, 49, 50, 54, 54, 56, 60, 61, 63

(n + 1) / 2th value → (9 + 1) /2 = 10/ 2 = 5th value

5th value is 54, so median is 54

Median

Class A--IQs of 13 Students

89

93

97

98

102

106

109

110

115

119

128

131

140

Median = 109

(six cases above, six below)

Median

If the first student were to drop out of Class A, there would be a new median:

~~89~~

93

97

98

102

106

109

110

115

119

128

131

140

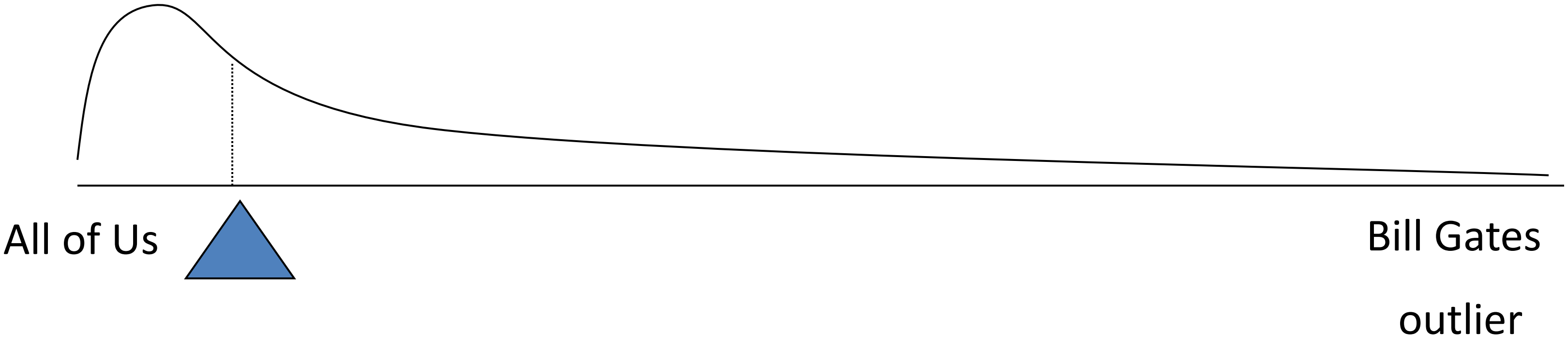
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

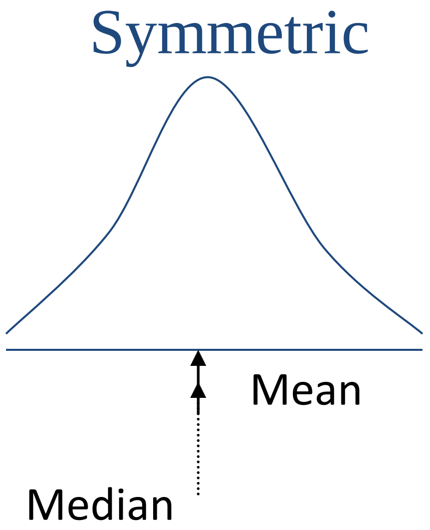
Median

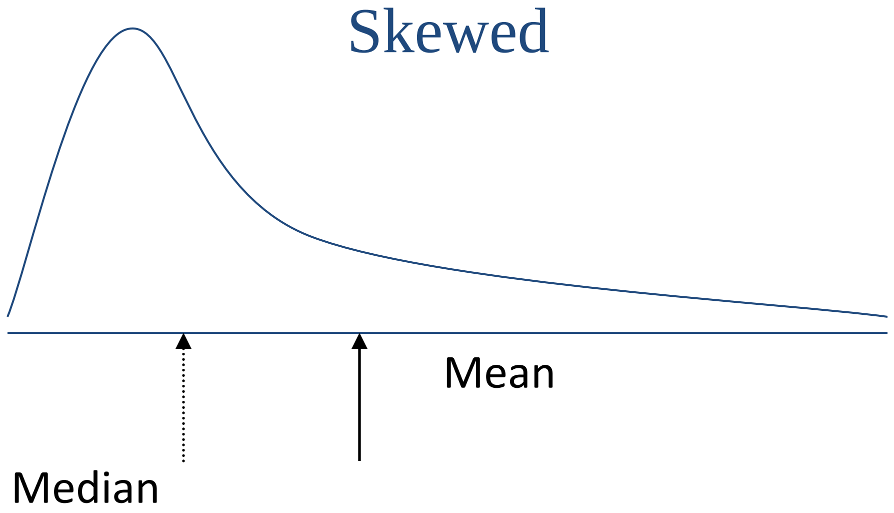
1. The median is unaffected by outliers, making it a better measure of central tendency, better describing the “typical person” than the mean when data are skewed.



Median

2. If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
3. In skewed data, the mean lies further toward the skew than the median.





Mode

- Value most frequently occurring in a set of data
- More than one mode present
- Can be used for the categorical data.

eg. Modal age (year) of the following 9 subjects

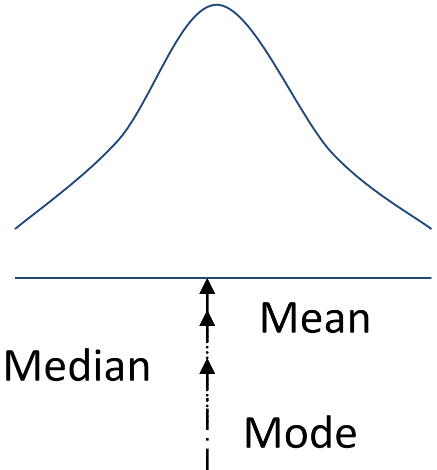
56, **54**, 61, 60, **54**, 44, 49, 50, 63

54 is modal age

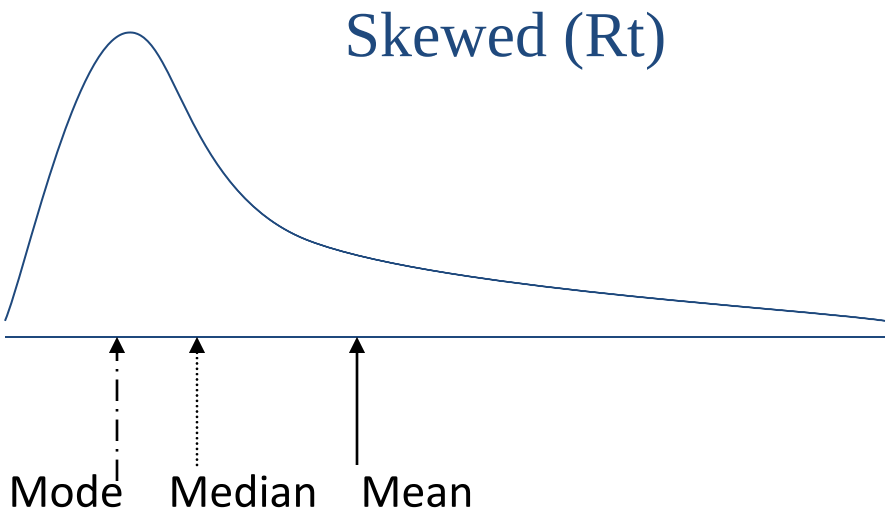
Mode

1. It may give you the most likely experience rather than the "typical" or "central" experience.
2. In symmetric distributions, the mean, median, and mode are the same.
3. In skewed data, the mean and median lie further toward the skew than the mode.

Symmetric



Skewed (Rt)



Measures of dispersion

- Dispersion: synonyms → variation, spread, scatter

- **Range**: The difference between the largest and smallest value in a set of data and poor measure of dispersion.

$$\mathbf{R = x_L - x_S}$$

eg. The range of ages (year) of the following 9 subjects

56, 54, 61, 60, 54, **44**, 49, 50, **63**

$$R = \mathbf{x_L} \;\mathbf{-}\; x_S = 63 – 44 = 19$$

Range

- The spread, or the distance, between the lowest and highest values of a variable.
- To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students

102

128

131

98

140

93

110

115

109

89

106

119

97

Class A Range = 140 - 89 = 51

Class B--IQs of 13 Students

127

131

96

80

93

120

109

162

103

111

109

87

105

Class B Range = 162 - 80 = 82

Standard Deviation

- **Standard deviation**: s for sample

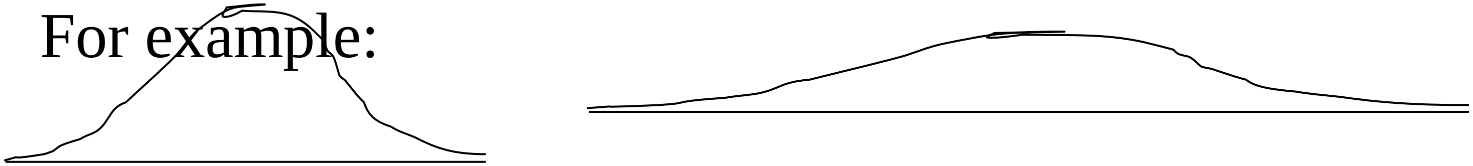
  : σ for population
- It measures how each observation in the data set differs from the mean
- The square root of the variance reveals the average deviation of the observations from the mean.

$$\text{s.d.} = \sqrt{\text{variance}}$$

Standard Deviation

1. The larger s.d. the greater amounts of variation around the mean.

   For example:

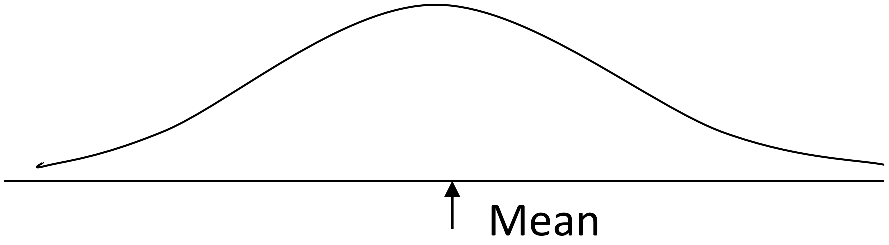


2. s.d. = 0 only when all values are the same (only when you have a constant and not a “variable”)
3. Like the mean, the s.d. will be inflated by an outlier case value.

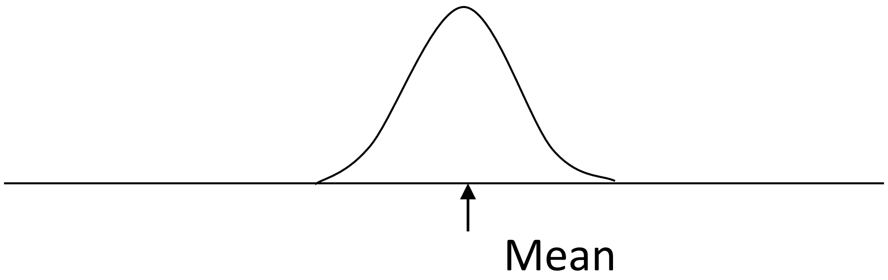
Variance

- An average measure of squared deviation of observations from the mean.

- The larger the variance, the further the individual cases are from the mean.



- The smaller the variance, the closer the individual scores are to the mean.



Variance

- Variance is a number that at first seems complex to calculate.
- Calculating variance starts with a “deviation.”
- A deviation is the distance away from the mean of a case’ s score.

$$\text{variance } (s^2) = \frac{\Sigma\,(x - \bar{x})^2}{n-1}$$

The coefficient of variation

- To compare the dispersion in two sets of data.
- Express the standard deviation as a percentage of mean.
- Useful in comparing the relative variability of different kinds of characteristics or with different unit.

$$\text{CV} = \frac{s}{x} * 100 = (\ )\ \%$$

Descriptive Statistics

Summarizing Data:

- ✓ Central Tendency (or Groups’ “Middle Values”)
  - Mean
  - Median
  - Mode

- ✓ Variation (or Summary of Differences Within Groups)
  - Range
  - Standard Deviation
  - Variance
  - Coefficient of variation

– …Wait! There’s more

Normal Distribution

- Symmetrical distribution of data
- Normal curve or Gaussian distribution
- The shape of curve depends on mean and SD

Properties of normal distribution

- Symmetrical, belled shape
- Usually not touch to the base line
- Mean, Median, Mode are the same
- Area under the curve, ± 1 SD = 68.26%
  ± 2 SD = 95.46%
  ± 3 SD = 99.74%

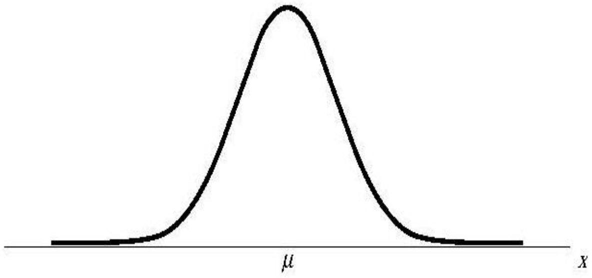
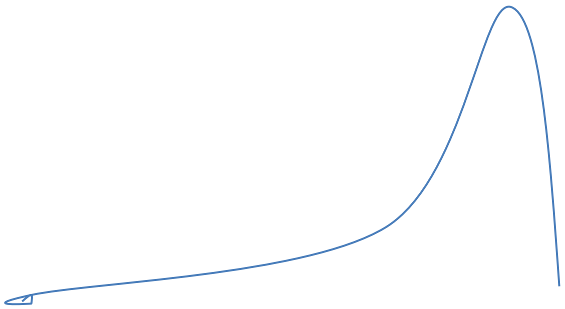


FIGURE 4.6.1 Graph of a normal distribution.

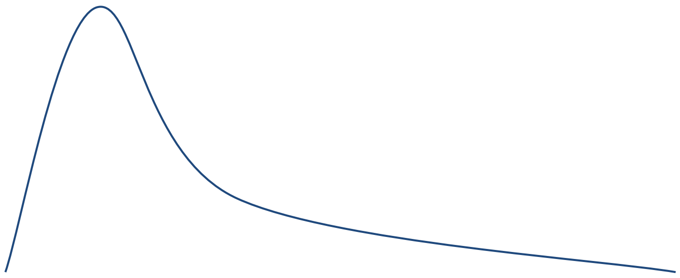
Skew Distribution

- If a graph (histogram or frequency polygon) of distribution is asymmetric, the distribution is said to be skewed.
- Right or positively skewed : if the graph extends further to the right, long tail to the right.
- Left or negatively skewed : if the graph extends further to the left, long tail to the left.

Skewed (Lt)



Skewed (Rt)

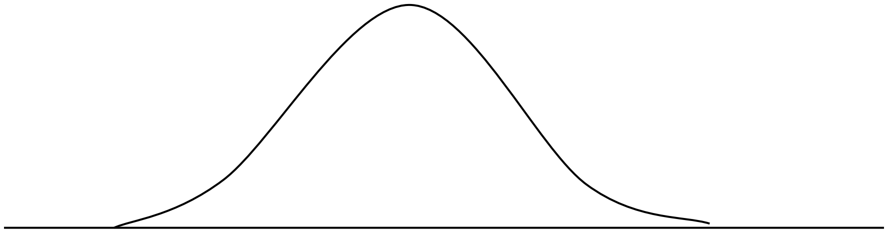


Kurtosis

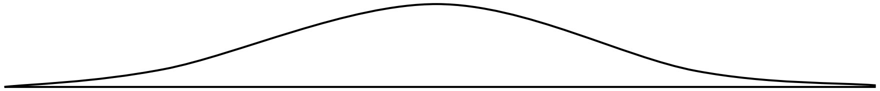
- Is the measure of the degree to which a distribution is peaked or flat in comparison to normal distribution whose graph is characterized by bell-shaped appearance.

- **Mesokurtic** : Kurtosis measure = 0
- Leptokurtic : Kurtosis measure > 0
- Platykurtic : Kurtosis measure < 0

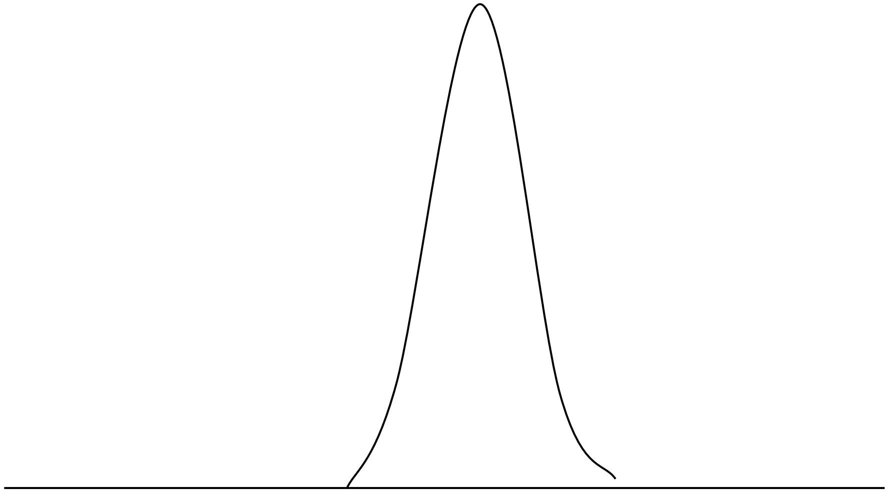
Curve Name



Mesokurtic (Normal)



Platykurtic



Leptokurtic

Descriptive Statistics

- Now you are qualified use descriptive statistics!
- Questions?

Thank you!