2019 - 2020

Dr. Fazıl Küçük Faculty of Medicine, EMU

Year 1

Biostatistics Course

INSTRUCTOR: Assist. Prof. Dr. İlke Akçay

ilke.akcay@emu.edu.tr

Today's Topics

- Describing Data with numbers
- Describing Data with graphics

Measurements/techniques to organize and summarize data

- Graphical Methods
- Numerical Methods
- Numerical Measurements (Descriptive Measures / Descriptive Statistics)

Measurements/techniques to organize and summarize data

- Graphical Methods
  - Bar Charts / Line Graphs
  - Pie Charts
  - Histogram (bar chart of the frequency distribution)
  - Frequency Polygon (line graph of the frequency distribution)
  - Ogive (line graph of the cumulative frequency distribution)
  - Box and Whisker plots (representation of min, max, quartiles, and outliers)

Measurements/techniques to organize and summarize data

- Numerical Methods
  - Ordered array
  - Frequency distribution
  - Stem and Leaf displays/tables

Measurements/techniques to organize and summarize data

- Numerical Measurements (Descriptive Measures / Descriptive Statistics)
  - Measures of Central Tendency / Measures of Location
    - Mean, Median, Mode, Quartiles
  - Measures of Dispersion/Variability
    - Range, Variance, Standard Deviation, Coefficient of Variation, Percentiles and Quartiles

Graphical Methods
Bar Charts

BAR CHART A graph in which the classes are reported on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are proportional to the heights of the bars.

TABLE 2–2 Relative Frequency Table of Vehicles Sold By Type At Whitner Autoplex Last Month

Vehicle Type	Number Sold	Relative Frequency
Domestic	50	0.625
Foreign	30	0.375
Total	80	1.000

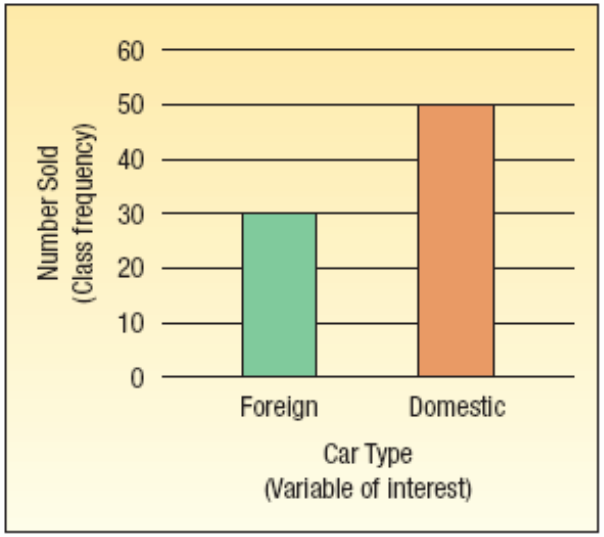


CHART 2–1 Vehicle Sold by Type Last Month At Whitner Autoplex

Graphical Methods
Pie Charts

PIE CHART A chart that shows the proportion or percent that each class represents of the total number of frequencies.

Use of Sales	Amount ($ million)	Percent of Share
Prizes	1,276.0	59
Payments to Education	648.1	30
Bonuses/Commissions	132.8	6
Operating Expenses	97.7	5
Total	2,154.6	100

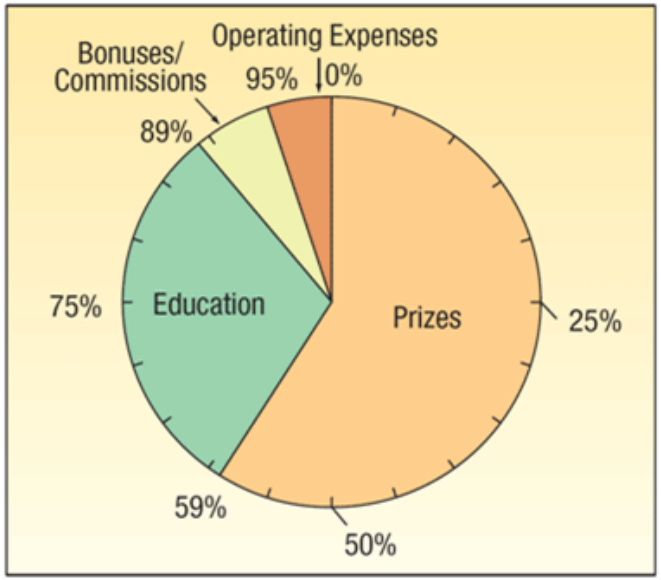


CHART 2–2 Pie Chart of Ohio Lottery Expenses in 2004

It is an informative way of showing how a single variable is divided among various classes or categories. It is particularly useful when there are a number of categories.

Numerical Methods
Ordered Array

Ordered Array

- An ordered array is a listing of the values of a collection (either population or sample) in order of magnitude from the smallest value to the largest value
  - Enables us to quickly determine values of the smallest and the largest measurements, and other facts about the arrayed data
  - If the size of the collection is large, use of a computer tool is suggested

Table 4.1 Hemoglobin Levels of 90 High-Altitude Mine Workers (g/cm^3)

18.5	16.8	23.2	19.4	19.5	20.6	22.0	17.8	16.2
23.3	19.7	21.6	24.2	21.4	20.8	19.7	21.1	23.0
21.7	18.4	22.7	20.9	20.5	16.1	16.9	24.8	12.2
17.4	17.8	19.3	17.3	18.3	17.8	17.1	18.4	19.7
17.8	19.0	19.2	15.5	26.2	19.1	20.9	18.0	21.0
20.2	18.3	19.2	17.2	19.8	19.5	20.0	18.4	15.9
19.9	16.4	18.4	17.8	23.0	19.4	20.3	18.2	13.1
20.3	18.5	24.1	14.3	17.8	19.9	23.5	19.7	19.3
20.6	18.3	20.8	17.6	18.1	19.7	19.1	19.5	23.5
18.5	20.0	22.4	18.8	16.2	15.6	15.5	18.5	19.0

Table 4.2 Ordered Array of Hemoglobin Levels of 90 High-Altitude Mine Workers (g/cm^3)

12.2	16.4	17.8	18.4	19.0	19.5	20.0	20.9	23.0
13.1	16.8	17.8	18.4	19.1	19.5	20.0	20.9	23.0
14.3	16.9	17.8	18.4	19.1	19.7	20.2	21.0	23.2
15.5	17.1	17.8	18.4	19.2	19.7	20.3	21.1	23.3
15.5	17.2	18.0	18.5	19.2	19.7	20.3	21.4	23.5
15.6	17.3	18.1	18.5	19.3	19.7	20.5	21.6	23.5
15.9	17.4	18.2	18.5	19.3	19.7	20.6	21.7	24.1
16.1	17.6	18.3	18.5	19.4	19.8	20.6	22.0	24.2
16.2	17.8	18.3	18.8	19.4	19.9	20.8	22.4	24.8
16.2	17.8	18.3	19.0	19.5	19.9	20.8	22.7	26.2

Numerical Methods
Frequency Distribution

Frequency Distribution

- Although a set of observations can be made by means of an ordered array, **further useful summarization may be achieved by grouping the data.**
- To group a set of observations we select a set of contiguous, nonoverlapping intervals such that each value in the set of observations can be placed in only one of the intervals. These intervals are called class intervals.

Frequency Distribution

- Example: Create a frequency distribution for the given data below:

TABLE 2.2.1 Ordered Array of Ages of Subjects from Table 1.4.1

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

Class Interval	Frequency
30-39	11
40-49	46
50-59	70
60-69	45
70-79	16
80-89	1
Total	189

Class Interval	Frequency
30-35	3
36-41	10
42-47	30
48-53	49
54-59	35
60-65	32
66-71	21
72-77	5
78-83	4
Total	189

Frequency Distribution

- Midpoint of class interval:
  - The sum of the upper and lower limits of the class interval is divided by 2
  - In ex. (30+39)/2=34.5 (interval 1)
- The Cumulative Frequency
  - It can be computed by adding successive frequencies
- The Cumulative Relative Frequency
  - It can be computed by adding successive relative frequencies

Class interval	Midpoint	Frequency Freq (f)	Cumulative Frequency	Relative Frequency R.f	Cumulative Relative Frequency
30 – 39	34.5	11	11	0.0582	0.0582
40 – 49	44.5	46	57	0.2434	0.3016
50 – 59	54.5	70	127	0.3704	0.6720
60 – 69	64.5	45	172	0.2381	0.9101
70 – 79	74.5	16	188	0.0847	0.9948
80 – 89	84.5	1	189	0.0053	1
Total		189		1	

Rel.freq= freq/n

Class interval	Midpoint	Frequency Freq (f)	Cumulative Frequency	Relative Frequency R.f	Cumulative Relative Frequency
30 – 39	34.5	11	11	0.0582	0.0582
40 – 49	44.5	46	57	0.2434	0.3016
50 – 59	54.5	70	127	0.3704	0.6720
60 – 69	64.5	45	172	0.2381	0.9101
70 – 79	74.5	16	188	0.0847	0.9948
80 – 89	84.5	1	189	0.0053	1
Total		189		1	

- From the above frequency table, answer the following questions:

1) The number of subjects with age less than 50 years ?
   - 46+11=57
2) The number of subjects with age between 40-69 years ?
   - 46+70+45=161
3) Percentage of subjects with age between 70-79 years ?
   - 8.47%
4) Percentage of subjects with age more than 69 years ?
   - 8.47+0.53=9%

Class interval	Midpoint	Frequency Freq (f)	Cumulative Frequency	Relative Frequency R.f	Cumulative Relative Frequency
30 – 39	34.5	11	11	0.0582	0.0582
40 – 49	44.5	46	57	0.2434	0.3016
50 – 59	54.5	70	127	0.3704	0.6720
60 – 69	64.5	45	172	0.2381	0.9101
70 – 79	74.5	16	188	0.0847	0.9948
80 – 89	84.5	1	189	0.0053	1
Total		189		1	

- From the above frequency table, answer the following questions:

5) The percentage of subjects with age between 40-49 years ?

- 46/189=0.2434 -> 24.34%

6) The percentage of subjects with age less than 60 years ?

- 5.82+24.34+37.04=67.2%

7) Number of intervals (k)?

- k=6

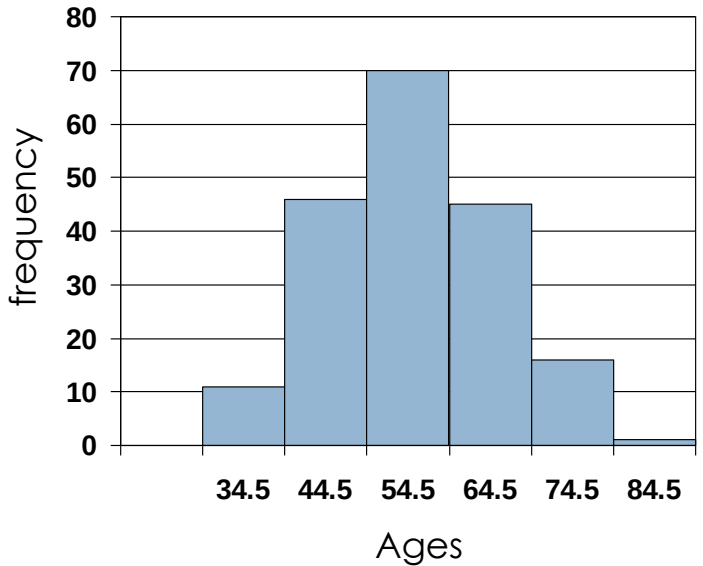
8) The width of the interval (w) ?

- w=10

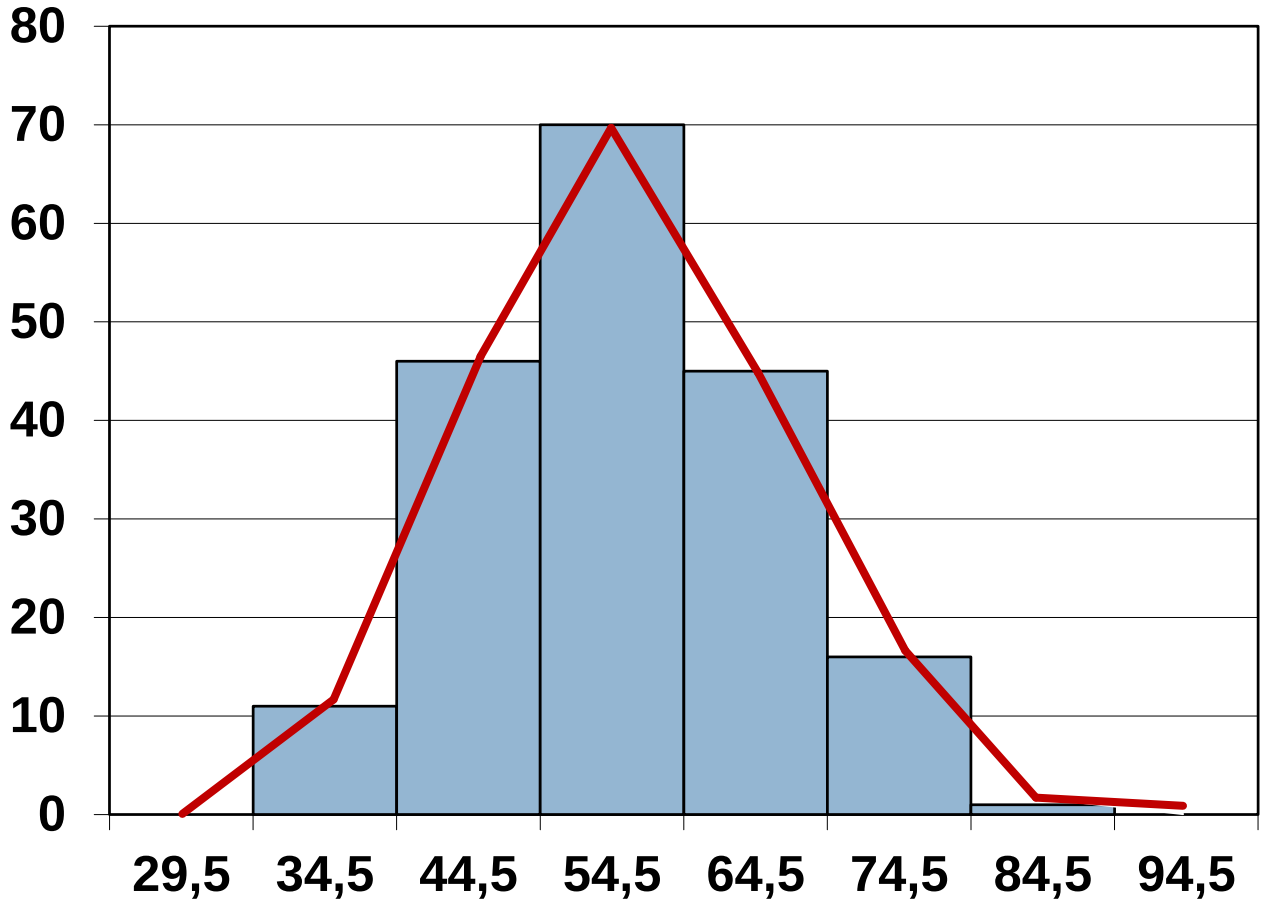
Graphical Methods
The Histogram

- Representing the grouped frequency table using the histogram
  - To draw the histogram, the true classes limits should be used.
  - They can be computed by subtracting 0.5 from the lower limit and adding 0.5 to the upper limit for each interval.

True class limits	Frequency
29.5 – <39.5	11
39.5 – < 49.5	46
49.5 – < 59.5	70
59.5 – < 69.5	45
69.5 – < 79.5	16
79.5 – < 89.5	1
Total	189



Graphical Methods
Frequency Polygon



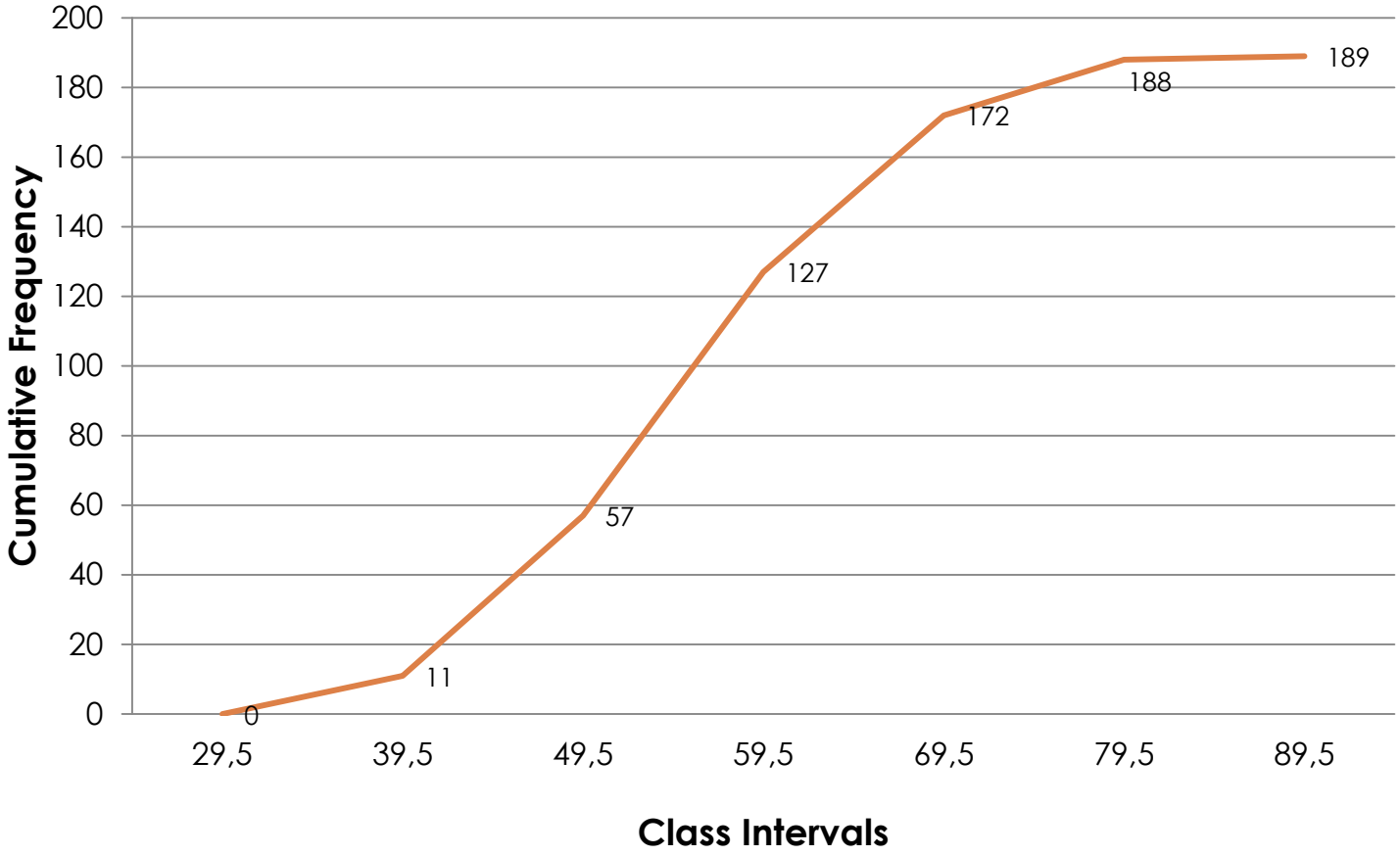
Graphical Methods
Ogive

- An **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution. It shows how many of values of the data are below certain boundary.
- **Steps for constructing an ogive:**

- Draw and label the x (horizontal) and the y (vertical) axes.
- Represent the cumulative frequencies on the y axis and the class boundaries on the x axis.
- Plot the cumulative frequency at each upper class boundary with the height being the corresponding cumulative frequency.
- Connect the points with segments. Connect the first point on the left with the x axis at the level of the lowest lower class boundary.

- **Note:** For the ogive we need the class intervals and the cumulative frequencies

Graphical Methods
Ogive



Numerical Methods
Stem and Leaf Displays

- A stem and leaf plot
  - provides information regarding the range of the dataset
  - shows the location of the highest concentration of measurements
  - reveals the presence or absence of symmetry.

- Preserves information contained in the individual measurements.

- Use the age data of 189 subjects to construct a stem-and-leaf display

TABLE 2.2.1 Ordered Array of Ages of Subjects from Table 1.4.1

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

Stem	Leaf
3	04577888899
4	00223333334444445556666667777778888888999999
5	00000000111122222233333333333333333444444444445556666667777778899999
6	00001111111111222222233444444556666667888999
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).

Stem Unit = 5
Leaf Unit = 1

```
Stem-and-leaf of Age          N = 189
Leaf Unit = 1.0
  2     3  04
 11     3  577888899
 28     4  0022333334444444
 57     4  55566666777777888888899999999
(46)    5  00000000111122222333333333333333344444444444
 86     5  55566666677777788999999
 62     6  00001111111111222222233444444
 32     6  556666667888999
 17     7  0111111123
  7     7  567888
  1     8  2
```

FIGURE 2.3.8 Stem-and-leaf display prepared by MINITAB from the data on subjects' ages shown in Table 2.2.1; class interval width = 5.

Numerical Measurements (Descriptive Measures)

- Descriptive Measure: a single number computed to summarize the data
- A Statistic: a descriptive measure computed from the data of a sample
- A Parameter: a descriptive measure computed from the data of a population
- Types of Descriptive Measures are limited as:
  - Measures of Central tendency
  - Measures of dispersion

Numerical Measurements
Measures of Central Tendency

- A measure of central tendency is a measure which indicates where the middle of the data is.
- The three most commonly used measures of central tendency are:

The **Mean**, the **Median**, and the **Mode**.

Measures of Central Tendency Mean

- **Arithmetic Mean: (the most familiar measure of CT)**
  - It is the average of the data.
- **Population Mean**
  - $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
    - It is usually unknown, then we use the sample mean to estimate or approximate it.
- **Sample Mean**
  - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- **Example:**
  - Here is a random sample of size 10 of ages, where

    $x_1 = 43$, $x_2 = 66$, $x_3 = 61$, $x_4 = 64$, $x_5 = 65$,

    $x_6 = 38$, $x_7 = 59$, $x_8 = 57$, $x_9 = 57$, $x_{10} = 50$

    $\bar{x}$= (43 + 66 + … + 57+50) / 10 = 56

Properties of the Mean

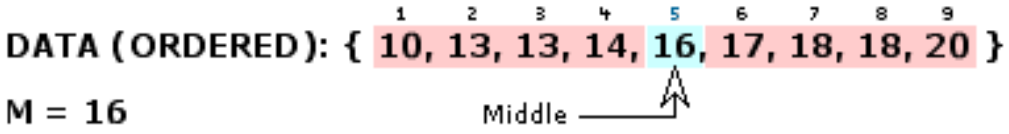
- Uniqueness
  - For a given set of data there is one and only one mean.
- Simplicity
  - It is easy to understand and to compute.
- Affected by extreme values
  - Since all values enter into the computation.
  - Extreme values have an influence on the mean

- **Example:**
  - Assume the values are 115, 110, 119, 117, 121 and 126.
    - The mean = 118.
  - But assume that the values are 75, 75, 80, 80 and 280.
    - The mean = 118, a value that is not representative of the set of data as a whole.

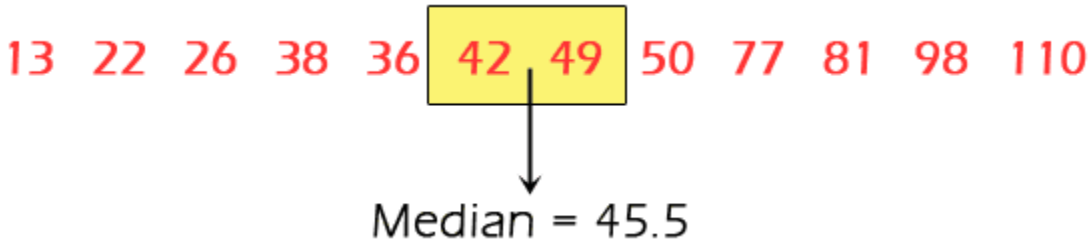
Measures of Central Tendency Median

- **The Median:**
  - It is a measure in the center of the data set
  - When ordering the data, it is the observation that divide the set of observations into two equal parts such that half(50%) of the data are before it and the other are after it.

- If n is odd, the median will be the middle of observations.
  - It will be the $(n+1)/2^{\underline{th}}$ ordered observation.
    - When n = 11, then the median is the 6^{th} observation.



- If n is even, there are two middle observations. The median will be the mean of these two middle observations.
  - It will be the $(n+1)/2^{\underline{th}}$ ordered observation.
    - When n = 12, then the median is the 6.5^{th} observation, which is an observation halfway between the 6^{th} and 7^{th} ordered observation.



- Example
  - Find the median of the ordered data: 38,43,50,57,57,59,61,64,65,66
    - Since n is even, there is no middle value
    - Two middle values are 57 and 59
    - The median = (57+59)/2=58

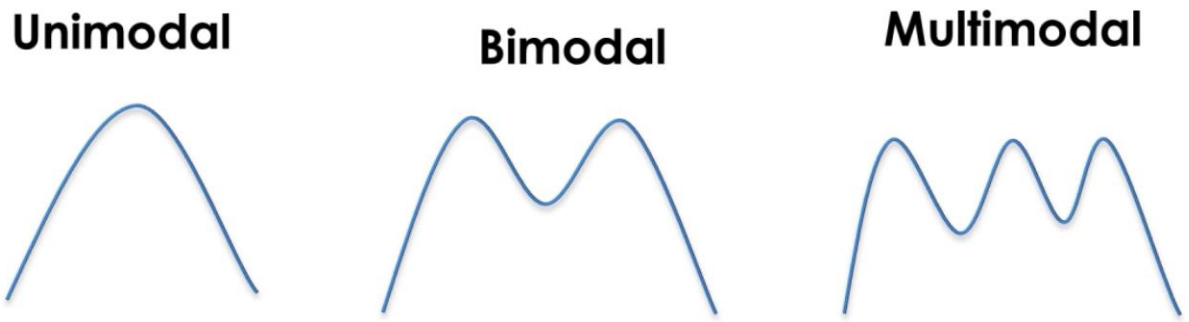
- Warning: To find the median, data has to be ordered

Properties of the Median:

- Uniqueness For a given set of data there is one and only one median.
- Simplicity It is easy to calculate.
- It is not affected by extreme values as is the mean.

Measures of Central Tendency Mode

- **The Mode:**
- It is the value which occurs most frequently.
- If all values are different there is no mode.
- Sometimes, there are more than one mode.
  - If there are two modes -> bimodal distribution
  - If there are more than two modes -> multimodal distribution



Measures of Central Tendency Mode

- Ex1 Consider a lab with 10 employees whose ages are 20, 21, 20, 20, 34, 22, 24, 27, 27, 27.
  - Mode values are **20** and **27**.
- Ex2 Sample with values 10, 21, 33, 53, 54
  - has **no mode**!!!

- **Properties of the Mode:**
  - Sometimes, it is not unique.
  - It may be used for describing qualitative data.

central tendency measures

Mode	→	**Depends only on the frequency of the observations**
Median	→	**Depends only on the relative positions of the observations**
Mean	→	**Calculated by using the values of all the observations**

Skewness

- If the graph (histogram or frequency polygon) of a distribution is asymmetric, the distribution is said to be **skewed**.
- In symmetric distributions, the mode, median and the arithmetic mean are the same.

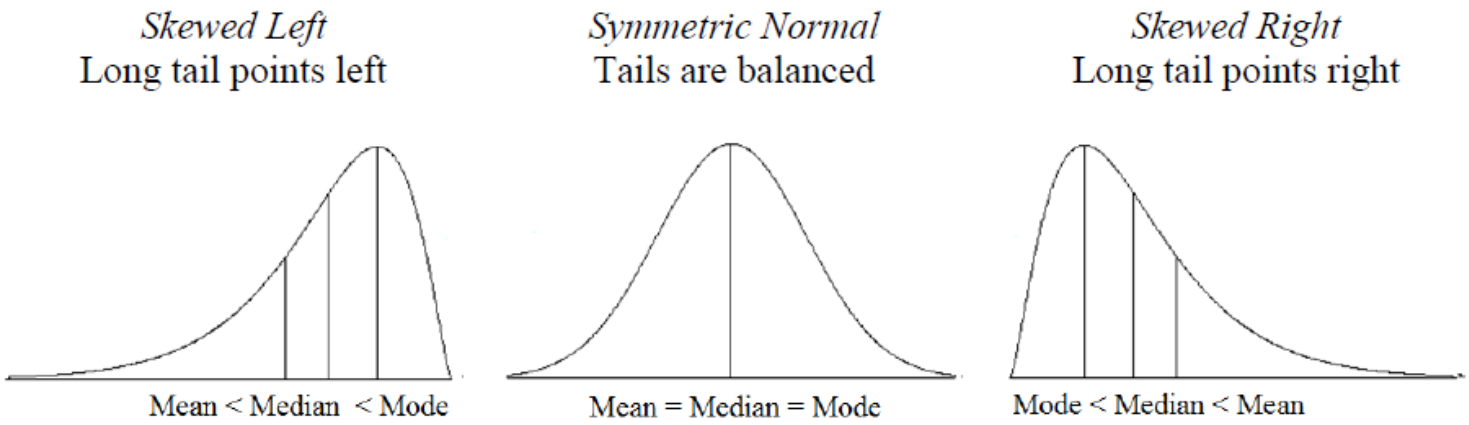


Figure 1. Sketches showing general position of mean, median, and mode in a population.

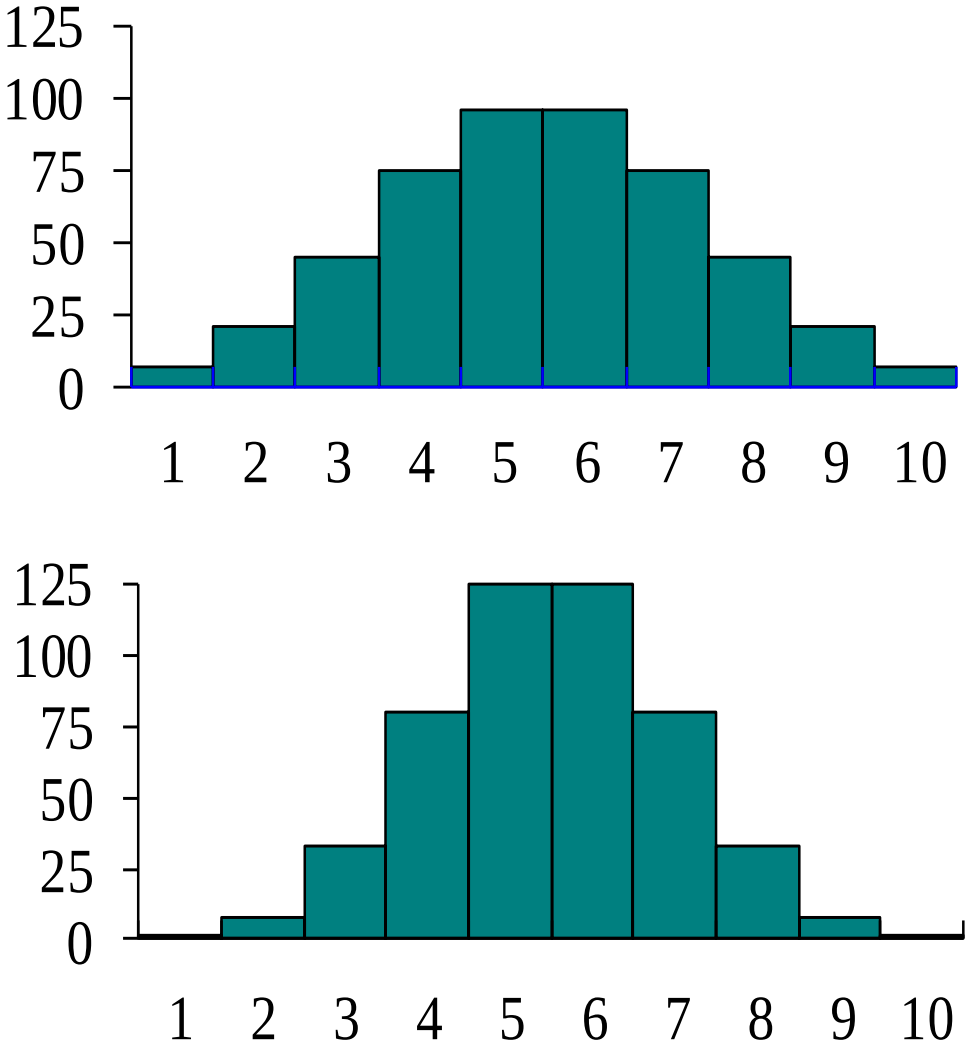
Numerical Measurements
Measures of Dispersion

- A measure of dispersion conveys information regarding the amount of variability present in a set of data.
- Note:
  - If all the values are the same
    - There is no dispersion
  - If all the values are different
    - There is a dispersion
      - If the values close to each other
        - The amount of Dispersion is small.
      - If the values are widely scattered
        - The Dispersion is greater.

- Which of the distributions of scores has the larger dispersion?

- The upper distribution has more dispersion because the scores are more spread out

That is, they are less similar to each other



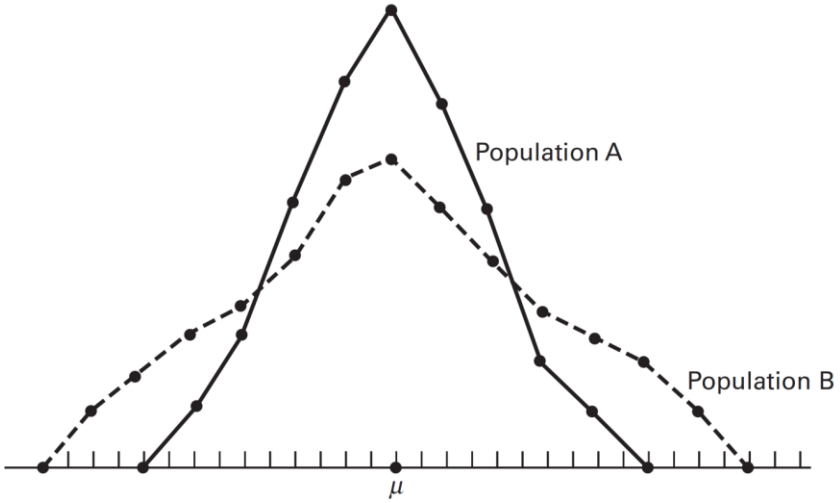


FIGURE 2.5.1 Two frequency distributions with equal means but different amounts of dispersion.

- Measures of Dispersion are:
  - The Range
  - The Variance
  - Standard Deviation
  - The Coefficient of Variation
  - Percentiles and Quartiles

Measures of Dispersion
The Range

- Measures the variation in a frequency distribution. It is defined as the difference between the largest (x_L) and smallest values (x_S).

- **Range (R) = $x_L - x_S$**

Range only defines the difference between two end values. It defines the variation of values in the data set.

↓ but

It does not give information about the distribution of the values between the two end values.

Variance is a better measure of dispersion.

↓ because

It is calculated by using all the values in the data set.

Measures of Dispersion
The Variance

- Variance is the measure of dispersion relative to the scatter of the values about the mean

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

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

What Does the Variance Formula Mean?

- First the mean is subtracted from each of the scores
  - This difference is called a *deviate* or a *deviation score*
  - The deviation tells us how far a given score is from the mean
- Variance is the mean of the squared deviation scores
- The larger the variance is, the more the scores deviate away from the mean
- If the variance is small, then it means the deviation is low

Measures of Dispersion
Standard Deviation

- The standard deviation is the square root of variance

Sample Standard Deviation: $s = \sqrt{s^2}$

Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$

- When the deviate scores are squared in variance, their unit of measure is squared as well
  - E.g. If people's weights are measured in pounds, then the variance of the weights would be expressed in pounds^2 (or squared pounds)
- Since squared units of measure are often awkward to deal with, the square root of variance is often used instead

Example

- Let's say you are given a data set for trees in California (in feet):
- 3,21,98,203,17,9

$$\bar{X} = \frac{351}{6} = 58.5$$

$$S^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{(3-58.5)^2+(21-58.5)^2+\cdots+(9-58.5)^2}{6-1} = \frac{31{,}099.5}{5} = 6{,}219.9$$

$S = \sqrt{6129.9} = \mathbf{78.87}$ is the standard deviation

Measures of Dispersion
The Coefficient of Variation

- **Coefficient of variation (CV)** is used to compare the dispersion in two sets of data
  - it expresses the standard deviation as a percentage of the mean

$$C.V = \frac{S}{\bar{X}}(100)$$

  - $\bar{X}$ is the sample mean and s is the sample std dev

- **Example**
  - Suppose two samples of human males yield the following data:

	Sample1	Sample2
Age	25-year-olds	11year-olds
Mean weight	145 pound	80 pound
Standard deviation	10 pound	10 pound

$$V = \frac{10}{145} \text{ x } 100 = \%6.9 \qquad V = \frac{10}{80} \text{ x } 100 = \%12.5$$

Thus, the variation is much bigger in the sample of 11-year olds than in the sample of 25 years old.