

UNIT-2

DESCRIBING DATA

Syllabus: UNIT II

Frequency distributions–Outliers–relative frequency distributions–cumulative frequency distributions–frequency distributions for nominal data–interpreting distributions–graphs–averages–mode–median–mean–averages for qualitative and ranked data – describing variability–range–variance–standard deviation–degrees of freedom–inter quartile range–variability for qualitative and ranked data.

Frequency Distributions

- A frequency distribution is a collection of observations produced by sorting observations into classes and showing their frequency (f) of occurrence in each class.
- A frequency distribution helps us to detect any pattern in the data (assuming a pattern exists) by superimposing some order on the inevitable variability among observations.
- The advantage of using frequency distributions is that they present raw data in an organized, easy-to-read format. The most frequently occurring scores are easily identified, as are score ranges, lower and upper limits, cases that are not common, outliers, and total number of observations between any given scores.
- Frequency distribution shows whether the observations are high or low and also whether they are concentrated in one area or spread out across the entire scale.
- Different Types of Frequency distributions:
 - Ungrouped frequency distribution.
 - Grouped frequency distribution.
 - Relative frequency distribution.
 - Cumulative frequency distribution

Frequency Distribution for Ungrouped Data

- A frequency distribution produced whenever observations are sorted into classes of single values is referred to as a frequency distribution for ungrouped data.
- Frequency distributions for ungrouped data are much more informative when the number of possible values is less than about 20.
- Example:

3	7	2	7	8
3	1	4	10	3
2	5	3	5	8
9	7	6	3	7
8	9	7	3	6

Frequency distribution for ungrouped data:

RATING	TALLY*	f
10	/	1
9	//	2
8	///	3
7	////	5
6	//	2
5	//	2
4	/	1
3	//// /	6
2	//	2
1	/	1
	Total	25

**Tally column usually is omitted from the finished table.*

Frequency Distribution for Grouped Data

→ A frequency distribution produced whenever observations are sorted into classes of more than one value is referred to as a frequency distribution for grouped data.

→ Example:

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

Frequency distribution for grouped data:

IQ	TALLY*	f
120–124	/	1
115–119		0
110–114	//	2
105–109	///	3
100–104	////	4
95–99	//// /	6
90–94	//// //	7
85–89	////	4
80–84	///	3
75–79	///	3
70–74	/	1
65–69	/	1
	Total	35

**Tally column usually is omitted from the finished table.*

Relative Frequency Distributions

→ Relative frequency distributions show the frequency of each class as a part or fraction of the total frequency for the entire distribution.

- This type of distribution allows us to focus on the relative concentration of observations among different classes within the same distribution.
- This type of distribution is especially helpful when we must compare two or more distributions based on different total numbers of observations.
- The conversion to relative frequencies allows a direct comparison of the shapes of these two distributions without having to adjust for the radically different total numbers of observations.
- To convert a frequency distribution into a relative frequency distribution, divide the frequency for each class by the total frequency for the entire distribution.
- Example:

Table 2.5 RELATIVE FREQUENCY DISTRIBUTION		
WEIGHT	<i>f</i>	RELATIVE <i>f</i>
240–249	1	.02
230–239	0	.00
220–229	3	.06
210–219	0	.00
200–209	2	.04
190–199	4	.08
180–189	3	.06
170–179	7	.13
160–169	12	.23
150–159	17	.32
140–149	1	.02
130–139	3	.06
Total	53	1.02*

* The sum does not equal 1.00 because of rounding-off errors.

Cumulative Frequency Distributions

- Cumulative frequency distributions show the total number of observations in each class and in all lower-ranked classes.
- This type of distribution can be used effectively with sets of scores, such as test scores for intellectual or academic aptitude, when relative standing within the distribution assumes primary importance. Under these circumstances, cumulative frequencies are usually converted, in turn, to cumulative percentages.
- Cumulative percentages are often referred to as percentile ranks.
- To convert a frequency distribution into a cumulative frequency distribution, add to the frequency of each class the sum of the frequencies of all classes ranked below it.

→ Example:

Table 2.6 CUMULATIVE FREQUENCY DISTRIBUTION			
WEIGHT	<i>f</i>	CUMULATIVE <i>f</i>	CUMULATIVE PERCENT
240–249	1	53	100
230–239	0	52	98
220–229	3	52	98
210–219	0	49	92
200–209	2	49	92
190–199	4	47	89
180–189	3	43	81
170–179	7	40	75
160–169	12	33	62
150–159	17	21	40
140–149	1	4	8
130–139	<u>3</u>	3	6
Total	53		

Constructing Frequency Distributions

→ For producing a well-constructed frequency distribution, three rules are essential and should not be violated.

1. Each observation should be included in one, and only one, class.
2. List all classes, even those with zero frequencies.
3. All classes should have equal intervals.

→ Step-by-step procedure for constructing Frequency Distributions:

1. Find the range, that is, the difference between the largest and smallest observations.
2. Find the class interval required to span the range by dividing the range by the desired number of classes (ordinarily 10).
3. Round off to the nearest convenient value.
4. Determine where the lowest class should begin. (Ordinarily, this number should be a multiple of the class interval.)
5. Determine where the lowest class should end by adding the class interval to the lower boundary and then subtracting one unit of measurement.
6. Working upward, list as many equivalent classes as are required to include the largest observation.
7. Indicate with a tally the class in which each observation falls.
8. Replace the tally count for each class with a number-the frequency (*f*) -and show the total of all frequencies.
9. Supply headings for both columns and a title for the table.

→ Example:

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

Range = $123 - 69 = 54$

Class width or interval = $\frac{123 - 69}{10} = \frac{54}{10} = 5.4$

Round off to a convenient number, such as 5.

IQ	TALLY*	f
120-124	/	1
115-119		0
110-114	//	2
105-109	///	3
100-104	////	4
95-99	//// /	6
90-94	//// //	7
85-89	////	4
80-84	///	3
75-79	///	3
70-74	/	1
65-69	/	1
	Total	35

**Tally column usually is omitted from the finished table.*

Frequency Distributions for Nominal Data

- When, among a set of observations, any single observation is a word, letter, or numerical code, the data are nominal.
- Frequency distributions for qualitative data are easy to construct. Simply determine the frequency with which observations occupy each class, and report these frequencies.
- Example: Below frequency distribution reveals that Yes replies are approximately twice as prevalent as No replies.

Table 2.7 FACEBOOK PROFILE SURVEY	
Response	f
Yes	56
No	27
Total	83

- They also can be converted into relative frequency distributions and, if the data can be ordered because of ordinal measurement, into percentile ranks.

Interpreting Distributions

- When inspecting a distribution for the first time, we have train to look at the entire table, not just the distribution. Read the title, column headings, and any footnotes.
- After these preliminaries, inspect the content of the frequency distribution.
- When interpreting distributions, including distributions constructed by someone else, keep an open mind.

Outliers

- A very extreme score that requires special attention because of its potential impact on a summary of the data is called outlier.
- Example: A GPA of 0.06, an IQ of 170, summer wages of \$62,000

Dealing with Outliers

Check for Accuracy:

- Whenever an outlier encounter attempt to verify its accuracy.
- Example: For instance, whether GPA of 3.06 recorded erroneously as 0.06?
- If the outlier survives an accuracy check, it should be treated as a legitimate score.

Might Exclude from Summaries:

- Choose to segregate an outlier from any summary of the data.
- For example, we might relegate it to a footnote instead of using excessively wide class intervals in order to include it in a frequency distribution. Or we might use various numerical summaries, such as the median and inter quartile range

Might Enhance Understanding:

- A valid outlier can be viewed as the product of special circumstances; it can help to understand the data.
- For example, we might understand better why crime rates differ among communities by studying the special circumstances that produce a community with an extremely low (or high) crime rate, or why learning rates differ among third graders by studying a third grader who learns very rapidly (or very slowly).

Graphs

(Describing Data using Graphs)

- Data can be described clearly and concisely with the aid of a well constructed frequency distribution.
- Data can often be described even more vividly, by converting frequency distributions into graphs.

→ Most common types of graphs:

- Graphs for Quantitative Data
 - Histograms
 - Frequency Polygon
 - Stem and Leaf Displays
- Graphs for Qualitative Data
 - Bar graph

Histogram

→ A bar-type graph for quantitative data. The common boundaries between adjacent bars emphasize the continuity of the data, as with continuous variables.

→ Important features of histograms.

- Equal units along the horizontal axis (the X axis, or abscissa) reflect the various class intervals of the frequency distribution.
- Equal units along the vertical axis (the Y axis, or ordinate) reflect increases in frequency.
- The intersection of the two axes defines the origin at which both numerical scales equal 0.
- Numerical scales always increase from left to right along the horizontal axis and from bottom to top along the vertical axis.
- The body of the histogram consists of a series of bars whose heights reflect the frequencies for the various classes.

→ Example:

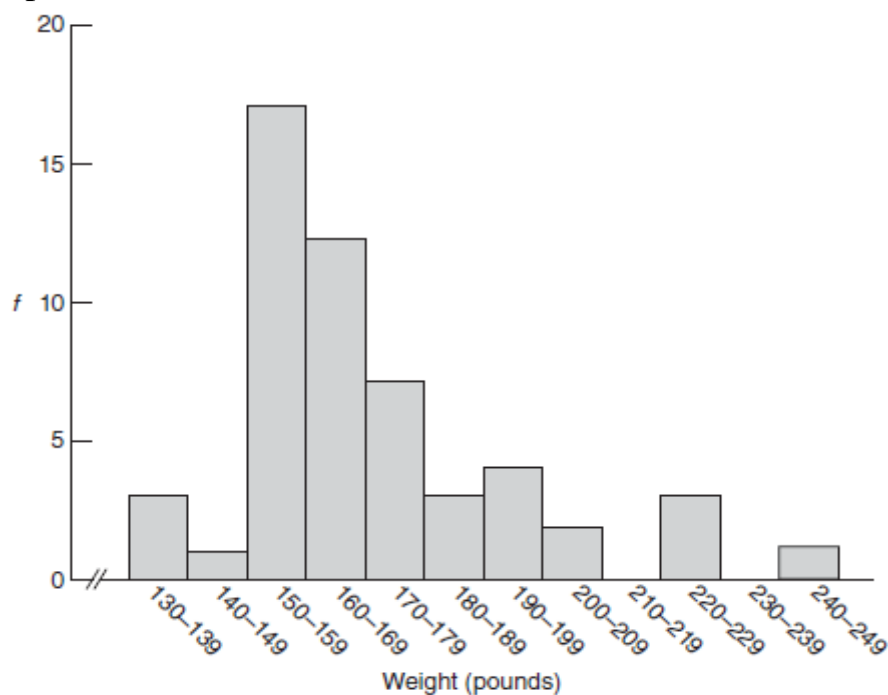


FIGURE 2.1
Histogram.

Frequency Polygon

- An important variation on a histogram is the **frequency polygon**, or line graph.
- Frequency polygons are particularly useful when two or more frequency distributions or relative frequency distributions are to be included in the same graph.
- Frequency polygons can be constructed directly from frequency distributions. It can also be constructed from histogram.
- The step-by-step transformation of a histogram into a frequency polygon:
 - **A:** This panel shows the histogram for the weight distribution.
 - **B:** Place dots at the midpoints of each bar top or, in the absence of bar tops, at midpoints for classes on the horizontal axis, and connect them with straight lines.
 - **C:** Anchor the frequency polygon to the horizontal axis. First, extend the upper tail to the midpoint of the first unoccupied class on the upper flank of the histogram. Then extend the lower tail to the midpoint of the first unoccupied class on the lower flank of the histogram. Now all of the area under the frequency polygon is enclosed completely.
 - **D:** Finally, erase all of the histogram bars, leaving only the frequency polygon.

→ Example:

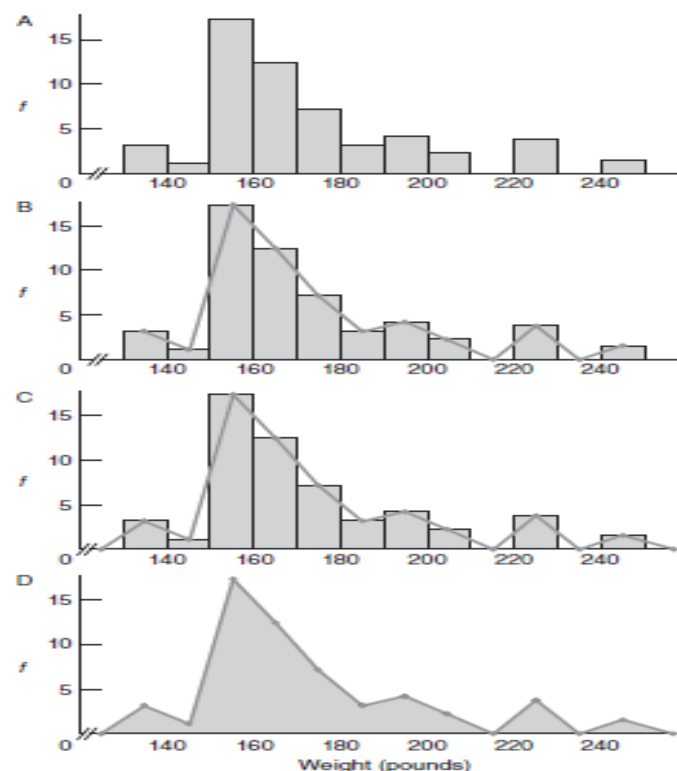


FIGURE 2.2

Transition from histogram to frequency polygon.

Stem and Leaf Displays

- Stem and leaf displays are ideal for summarizing distributions, such as that for weight data, without destroying the identities of individual observations.
- Stem and Leaf display is a device for sorting quantitative data on the basis of leading and trailing digits.
- Stem and leaf displays represent statistical bargains. Just a few minutes of work produces a description of data that is both clear and complete.
- Even though rarely appearing in published reports, stem and leaf displays often serve as the first step toward organizing data.
- A good stem and leaf display
 - shows the first digits of the number (thousands, hundreds or tens) as the *stem* and shows the last digit (ones) as the *leaf*.
 - usually uses whole numbers. Anything that has a decimal point is rounded to the nearest whole number. For example, test results, speeds, heights, weights, etc.
 - looks like a bar graph when it is turned on its side.
 - shows how the data are spread—that is, highest number, lowest number, most common number and outliers
- To construct the stem and leaf display
 - On the left hand side of the page, write down the thousands, hundreds or tens (all digits but the last one). These will be your stems.
 - Draw a line to the right of these stems.
 - On the other side of the line, write down the ones (the last digit of a number). These will be your leaves.
- Example 1: A teacher asked 10 of her students how many books they had read in the last 12 months. Their answers were as follows: 12, 23, 19, 6, 10, 7, 15, 25, 21, 12. Prepare a stem and leaf display for these data.

Books read in a year by 10 students	
Stem	Leaf
0	6 7
1	2 9 0 5 2
2	3 5 1

- Example2: Construct a stem and leaf display for the following IQ scores obtained from a group of four-year-old children.

120	98	118	117	99	111
126	85	88	124	104	113
108	141	123	137	78	96
102	132	109	106	143	»

Solution:

7	8				
8	5	8			
9	8	9	6		
10	8	2	9	6	4
11	8	7	1	3	
12	0	6	3	4	
13	2	7			
14	1	3			

Typical Shapes

- Whether expressed as a histogram, a frequency polygon, or a stem and leaf display, an important characteristic of a frequency distribution is its shape.
- Typical shapes for smoothed frequency polygons:
 - Normal
 - Bimodal
 - Positively Skewed
 - Negatively Skewed

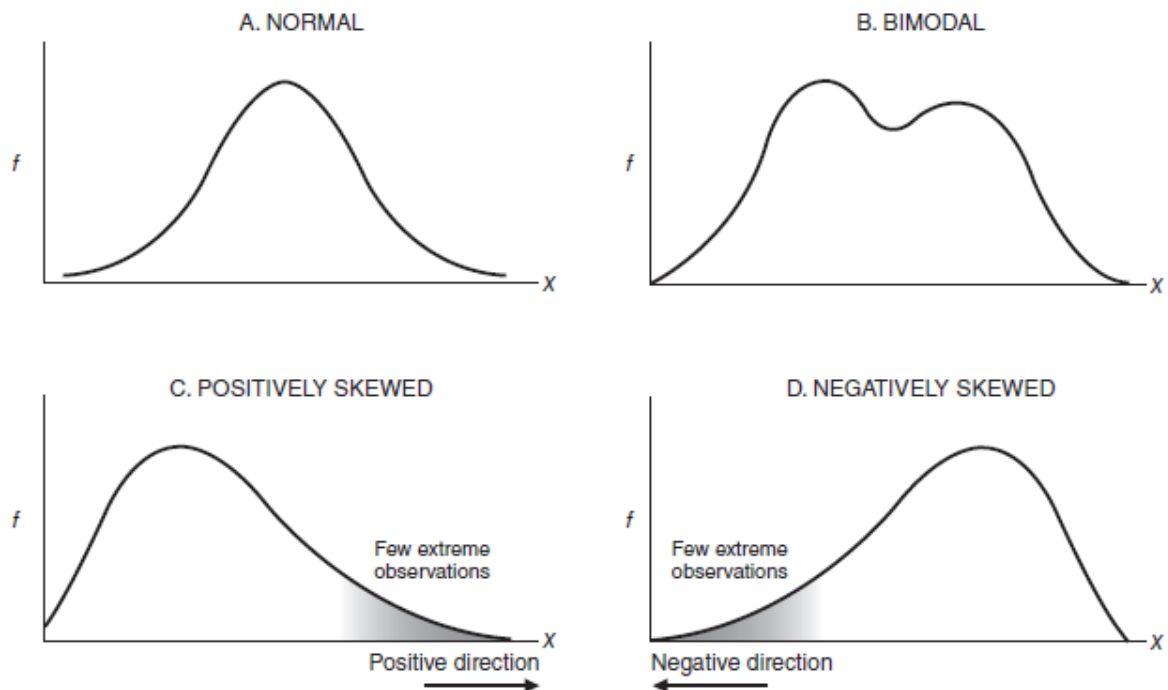


FIGURE 2.3
Typical shapes.

Normal

- The familiar bell-shaped silhouette of the normal curve can be superimposed on many frequency distributions, including those for uninterrupted gestation periods of human fetuses, scores on standardized tests, and even the popping times of individual kernels in a batch of popcorn.

Bimodal

- It reflects the coexistence of two different types of observations in the same distribution.
- For instance, the distribution of the ages of residents in a neighborhood consisting largely of either new parents or their infants has a bimodal shape.

Positively Skewed

- A lopsided distribution caused by a few extreme observations in the positive direction (to the right of the majority of Observations), is a positively skewed distribution.
- The distribution of incomes among U.S. families has a pronounced positive skew, with most family incomes under \$200,000 and relatively few family incomes spanning a wide range of values above \$200,000.

Negatively Skewed

- A lopsided distribution caused by a few extreme observations in the negative direction (to the left of the majority of observations), is a **negatively skewed distribution**.
- The distribution of ages at retirement among U.S. job holders has a pronounced negative skew, with most retirement ages at 60 years or older and relatively few retirement ages spanning the wide range of ages younger than 60.

Bar graphs: A Graph for Qualitative (Nominal) Data

- Bar graphs are often used with qualitative data and sometimes with discrete quantitative data.
- They resemble histograms except that gaps separate adjacent bars in bar graphs.

Example 1:

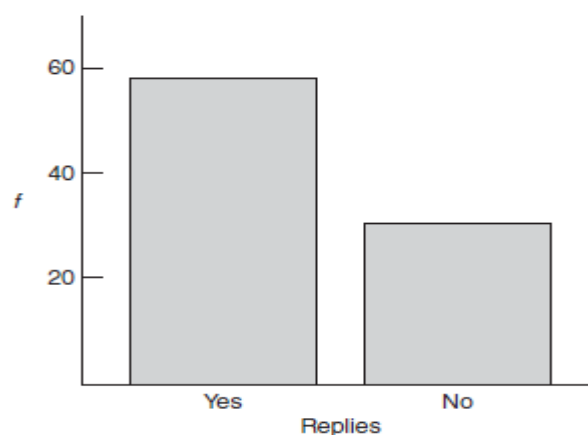


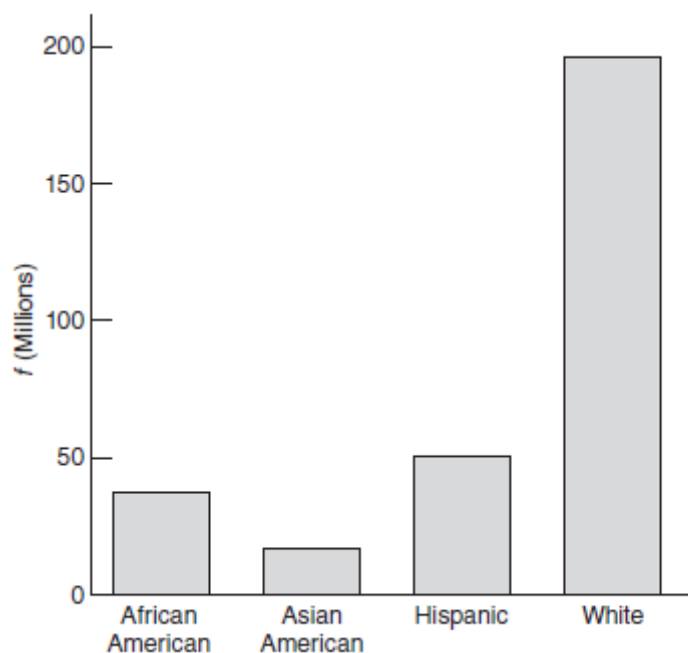
FIGURE 2.4
Bar graph.

A glance at this graph confirms that Yes replies occur approximately twice as often as No replies.

→ Example2: construct a bar graph for the data shown in the following table:

RACE/ETHNICITY OF U.S. POPULATION, 2010 (IN MILLIONS)	
Race/Ethnicity	<i>f</i>
African American	37.7
Asian American*	17.2
Hispanic	50.5
White	<u>196.8</u>
Total**	<u>302.2</u>

Solution:



Interpreting graphs

→ When interpreting graphs, beware of various unscrupulous techniques, such as using bizarre combinations of axes to either exaggerate or suppress a particular data pattern.

Describing Data with Averages

- Averages consist of numbers (or words) about which the data are, in some sense, centred. They are often referred to as **measures of central tendency**
- A **measure of center** is a single number used to describe a set of numeric data. It describes a typical value from the data set.
- Several types of average yield numbers or words that attempt to describe, most generally, the middle or typical value for a distribution.

- Three different measures of central tendency are:
 - Mode
 - Median
 - Mean.
- Each of these has its special uses, but the mean is the most important average in both descriptive and inferential statistics.

Mode

- The mode equals the value of the most frequently occurring or typical score.
- It is easy to assign a value to the mode. If the data are organized. However, if the data are not organized, some counting may be required.
- The mode is readily understood as the most prevalent or typical value.
- Distributions can have more than one mode (or no mode at all).
- Distributions with two obvious peaks, even though they are not exactly the same height, are referred to as **bimodal**.
- Distributions with more than two peaks are referred to as **multimodal**.
- The presence of more than one mode might reflect important differences among subsets of data. For instance, the distribution of weights for both male and female statistics students would most likely be bimodal, reflecting the combination of two separate weight distributions—a heavier one for males and a lighter one for females.
- Example1: Determine the mode for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.
Answer: mode = 63
- Example1: The owner of a new car conducts six gas mileage tests and obtains the following results, expressed in miles per gallon: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9. Find the mode for these data.
Answer: mode = 27.4

Median

- The median reflects the middle value when observations are ordered from least to most.
- The median splits a set of ordered observations into two equal parts, the upper and lower halves.
- In other words, the median has a percentile rank of 50, since observations with equal or smaller values constitute 50 percent of the entire distribution.
- To find the median, scores always must be ordered from least to most (or vice versa). This task is straightforward with small sets of data but becomes increasingly cumbersome with larger sets of data that must be ordered manually.

- When the total number of scores is odd, there is a single middle-ranked score, and the value of the median equals the value of this score. When the total number of scores is even, the value of the median equals a value midway between the values of the two middlemost scores.
- In either case, the value of the median always reflects the *value* of middle-ranked scores, not the *position* of these scores among the set of ordered scores
- Example 1: Find the median for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.
Solution: median = 63
- Example 2: Find the median for the following gas mileage tests: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9.
Solution: median = 27.15 (halfway between 26.9 and 27.4)

Mean

- The mean is the most common average.
- The mean is found by adding all scores and then dividing by the number of scores.
- That is

$$\text{Mean} = \frac{\text{sum of all scores}}{\text{number of scores}}$$

- There is no requirement that presidential terms be ranked before calculating the mean.
- Even when large sets of unorganized data are involved, the calculation of the mean is usually straightforward, particularly with the aid of a calculator or computer.
- The mean serves as the balance point for its frequency distribution.
- Mean cannot be used with qualitative data.
- Example 1: Find the mean for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.

Solution:

$$\text{mean} = \frac{672}{11} = 61.09$$

- Example 2: Find the mean for the following gas mileage tests: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9.

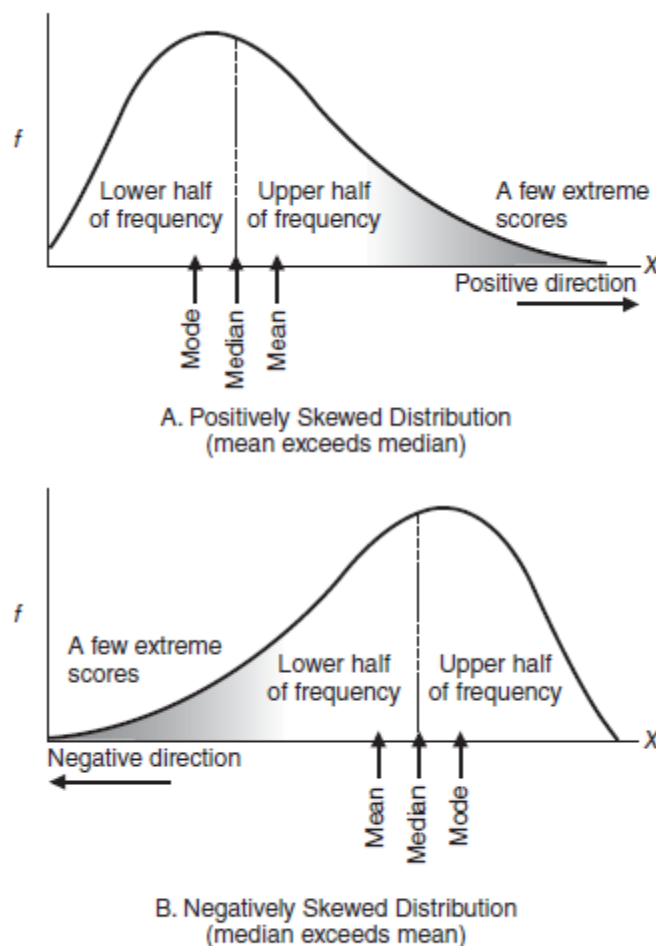
Solution:

$$\text{mean} = \frac{163.3}{6} = 27.22$$

Which Average?

- When a distribution of scores is not too skewed, the values of the mode, median, and mean are similar, and any of them can be used to describe the central tendency of the distribution.

- When extreme scores cause a distribution to be skewed, **the** values of the three averages can differ appreciably.
- Unlike the mode and median, the mean is very sensitive to extreme scores, or outliers.
- Ideally, when a distribution is skewed, report both the mean and the median. Appreciable differences between the values of the mean and median signal the presence of a skewed distribution.
- If the mean exceeds the median, the underlying distribution is positively skewed because of one or more scores with relatively large values.
- On the other hand, if the median exceeds the mean, the underlying distribution is negatively skewed because of one or more scores with relatively small values.
- In the long run, however, the *mean is the single most preferred average for quantitative data.*
- Following summarizes the relationship between the various averages and the two types of skewed distributions (shown as smoothed curves).

**FIGURE 3.2**

Mode, median, and mean in positively and negatively skewed distributions.

Averages for Qualitative and Ranked Data**Mode Always Appropriate for Qualitative Data**

- For quantitative data, in principle, all three averages can be used.
- *The mode always can be used with qualitative data.*

Median Sometimes Appropriate for Qualitative Data

- The median can be used whenever it is possible to order qualitative data from least to most because the level of measurement is ordinal.
- It's easiest to determine the median class for ordered qualitative data by using relative frequencies

Mean cannot be used with qualitative data.**Averages for Ranked Data**

- When the data consist of a series of ranks, with its ordinal level of measurement, the median rank always can be obtained. It's simply the middlemost or average of the two middlemost ranks.
- The mean and modal ranks tend not to be very informative and will not be discussed.

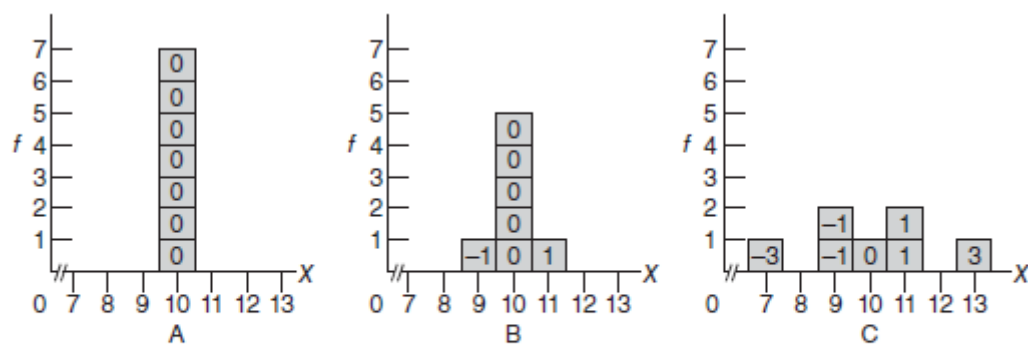
Describing Variability

- Measures of the amount by which scores are dispersed or scattered in a distribution are referred as **measures of variability**.
- Measures of variability reflect the amount by which observations are dispersed or scattered in a distribution.
- These measures provide an idea of the *dispersion of the data*. That is, how are the data spread out?
- In simple terms, if the scores in a distribution are all the same, then there is no variability.
- If there are small differences between scores, then the variability is small, and if there are large differences between scores, then the variability is large.
- Variability provides a quantitative measure of the degree to which scores in a distribution are spread out or clustered together.
- These measures assume a key role in the analysis of research results.
- Different measures of variability:
 - Range
 - Inter Quartile Range(IQR)
 - Variance
 - Standard Deviation.
 - Degrees of Freedom (*Df*)

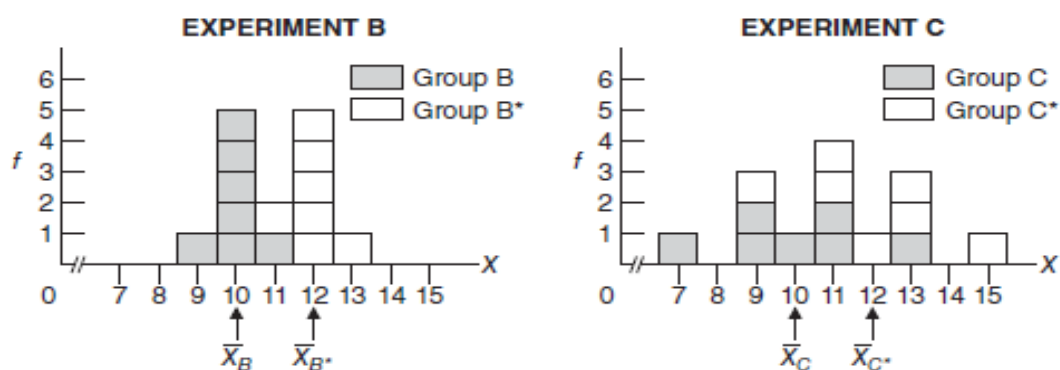
Importance of variability:

- Variability assumes a key role in an analysis of research results.
- Exact measures of variability not only aid communication but also are essential tools in statistics.

- In general, a good measure of variability serves two purposes:
- Variability describes the distribution. Specifically, it tells whether the scores are clustered close together or are spread out over a large distance.
 - Variability measures how well an individual score (or group of scores) represents the entire distribution. This aspect of variability is very important for inferential statistics where relatively small samples are used to answer questions about populations.
- Variabilities within groups assume a key role in inferential statistics.
- The relatively smaller variabilities within groups in experiment translate into *more statistical stability* for the observed mean difference *when it is viewed as just one outcome among many possible outcomes for repeat experiments*.
- On the other hand, the relatively larger variabilities within groups translate into *less statistical stability* for the observed mean difference when it is viewed as just one outcome among many possible outcomes for repeat experiments.

**FIGURE 4.1**

Three distributions with the same mean (10) but different amounts of variability. Numbers in the boxes indicate distances from the mean.

**FIGURE 4.2**

Two experiments with the same mean difference but dissimilar variabilities.

Range

- The **range** is the difference between the largest and smallest scores
 - The range is a handy measure of variability that can readily be calculated and understood.
 - Example: Determine the values of the ranges for the following sets of data.
 - (a) Retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63
 - (b) Residence changes: 1, 3, 4, 1, 0, 2, 5, 8, 0, 2, 3, 4, 7, 11, 0, 2, 3, 4
- Solution:** (a) Range=25 (b) Range= 11

Shortcomings of Range

- The **range** has several shortcomings.
 - First, since its value depends on only two scores—the largest and the smallest—it fails to use the information provided by the remaining scores.
 - Furthermore, the value of the range tends to increase with increases in the total number of scores.
- For this reason, the range is considered to be a crude and unreliable measure of variability.

Variance

- Although both the range and its most important spinoff, the interquartile range, serve as valid measures of variability, neither is among the statistician's preferred measures of variability.
- Those roles are reserved for the variance and particularly for its square root, the standard deviation, because these measures serve as key components for other important statistical measures.
- Accordingly, the variance and standard deviation occupy the same exalted position among measures of variability as does the mean among measures of central tendency.
- Although a measure of variability, the variance also qualifies as a type of mean, that is, as the balance point for some distribution. To qualify as a type of mean, the values of all scores must be added and then divided by the total number of scores. In the case of the variance, each original score is re-expressed as a distance or deviation from the mean by subtracting the mean.
- Variance is defined as the mean of all squared deviation scores. i.e., adding the consistently positive values of all squared deviation scores and then dividing by the total number of scores to produce *the mean of all squared deviation scores, also known as the variance*.

$$\text{Variance} = \text{mean squared deviation} = \frac{\text{sum of squared deviations}}{\text{number of scores}}$$

Standard Deviation

- Standard deviation describes variability in the original units of measurement
- The standard deviation is defined as the square root of the mean of all squared deviations from the mean, that is

$$\text{standard deviation} = \sqrt{\text{variance}}$$

- We think of standard deviation as a rough measure of the average (or standard) amount by which scores deviate on either side of their mean.
- Strictly speaking, the standard deviation usually exceeds the mean deviation or, more accurately, the mean absolute deviation.
- Nevertheless, it is reasonable to describe the standard deviation as the average amount by which scores deviate on either side of their mean.
- For most frequency distributions, a majority (often as many as 68 percent) of all scores are within one standard deviation on either side of the mean.
- For most frequency distributions, a small minority (often as small as 5 percent) of all scores deviate more than two standard deviations on either side of the mean.
- **Difference between the standard deviation the mean:** There's an important difference between the standard deviation and its indispensable co-measure, the mean. The mean is a measure of position, but the standard deviation is a measure of distance (on either side of the mean of the distribution).
- **Value of Standard Deviation Cannot Be Negative:** Standard deviation distances always originate from the mean and are expressed as positive deviations above the mean or negative deviations below the mean. Note, however, that although the actual value of the standard deviation can be zero or a positive number, it can never be a negative number because any negative deviation disappears when squared.
- As with the mean, statisticians distinguish between population and sample for both the variance and the standard deviation, depending on whether the data are viewed as a complete set (population) or as a subset (sample).
- **Sum of Squares (SS):** Calculating the standard deviation requires that we obtain first a value for the variance. However, calculating the variance requires, in turn, that we obtain the sum of the squared deviation scores. The sum of squared deviation scores or more simply the **sum of squares**, symbolized by SS.
- There are two formulas for the sum of squares: the **definition formula**, which is easier to understand and remember, and the **computation formula**, which usually is more efficient.

- The definition formula is cumbersome when, the mean equals some complex number, such as 169.51, or the number of scores is large. In these cases, use the more efficient computation formula
- Sum of Squares (SS) for Population (Definition Formula):

$$SS = \sum (X - \mu)^2$$

where SS represents the sum of squares, \sum directs us to sum over the expression to its right, and $(X - \mu)^2$ denotes each of the squared deviation scores.

Table 4.1 CALCULATION OF POPULATION STANDARD DEVIATION σ (DEFINITION FORMULA)		
A. COMPUTATION SEQUENCE		
Assign a value to N 1 representing the number of X scores		
Sum all X scores 2		
Obtain the mean of these scores 3		
Subtract the mean from each X score to obtain a deviation score 4		
Square each deviation score 5		
Sum all squared deviation scores to obtain the sum of squares 6		
Substitute numbers into the formula to obtain population variance, σ^2 7		
Take the square root of σ^2 to obtain the population standard deviation, σ 8		
B. DATA AND COMPUTATIONS		
X	4 $X - \mu$	5 $(X - \mu)^2$
13	3	9
10	0	0
11	1	1
7	-3	9
9	-1	1
11	1	1
9	-1	1
1 $N = 7$ 2 $\sum X = 70$ 6 $SS = \sum (X - \mu)^2 = 22$		
3 $\mu = \frac{70}{7} = 10$		
7 $\sigma^2 = \frac{SS}{N} = \frac{22}{7} = 3.14$ 8 $\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{22}{7}} = \sqrt{3.14} = 1.77$		

- Sum of Squares (SS) for Population (Computation Formula):

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

where $\sum X^2$, the sum of the squared X scores, is obtained by *first squaring each X score and then summing all squared X scores*; $(\sum X)^2$, the square of sum of all X scores, is obtained by *first adding all X scores and then squaring the sum of all X scores*; and N is the population size.

Table 4.2 CALCULATION OF POPULATION STANDARD DEVIATION (σ) (COMPUTATION FORMULA)	
A. COMPUTATIONAL SEQUENCE	
Assign a value to N representing the number of X scores 1	
Sum all X scores 2	
Square the sum of all X scores 3	
Square each X score 4	
Sum all squared X scores 5	
Substitute numbers into the formula to obtain the sum of squares, SS 6	
Substitute numbers into the formula to obtain the population variance, σ^2 7	
Take the square root of σ^2 to obtain the population standard deviation, σ 8	
B. DATA AND COMPUTATIONS	
	4
X	X^2
13	169
10	100
11	121
7	49
9	81
11	121
9	81
1 $N = 7$	2 $\sum X = 70$ 5 $\sum X^2 = 722$
	3 $(\sum X)^2 = 4900$
6 $SS = \sum X^2 - \frac{(\sum X)^2}{N} = 722 - \frac{4900}{7} = 722 - 700 = 22$	
7 $\sigma^2 = \frac{SS}{N} = \frac{22}{7} = 3.14$ 8 $\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{22}{7}} = \sqrt{3.14} = 1.77$	

→ Sum of Squares (SS) for Sample (Definition Formula):

$$SS = \sum (X - \bar{X})^2$$

Table 4.3 CALCULATION OF SAMPLE STANDARD DEVIATION (S) (DEFINITION FORMULA)		
A. COMPUTATION SEQUENCE Assign a value to n 1 representing the number of X scores Sum all X scores 2 Obtain the mean of these scores 3 Subtract the mean from each X score to obtain a deviation score 4 Square each deviation score 5 Sum all squared deviation scores to obtain the sum of squares 6 Substitute numbers into the formula to obtain the sample variance, s^2 7 Take the square root of s^2 to obtain the sample standard deviation, s 8		
B. DATA AND COMPUTATIONS		
	X	$X - \bar{X}$ 4 $(X - \bar{X})^2$ 5
	7	4
	3	0
	1	-2
	0	-3
	4	1
1 $n = 5$	2 $\Sigma X = 15$	6 $SS = \Sigma(X - \bar{X})^2 = 30$
	3 $\bar{X} = \frac{15}{5} = 3$	
7 $s^2 = \frac{SS}{n-1} = \frac{30}{4} = 7.50$	8 $s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{30}{4}} = \sqrt{7.50} = 2.74$	

→ Sum of Squares (SS) for Sample (Computation Formula):

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

Table 4.4 CALCULATION OF SAMPLE STANDARD DEVIATION (S) (COMPUTATION FORMULA)		
A. COMPUTATIONAL SEQUENCE Assign a value to n representing the number of X scores 1 Sum all X scores 2 Square the sum of all X scores 3 Square each X score 4 Sum all squared X scores 5 Substitute numbers into the formula to obtain the sum of squares, SS 6 Substitute numbers into the formula to obtain the sample variance, s^2 7 Take the square root of s^2 to obtain the sample standard deviation, s 8		
B. DATA AND COMPUTATIONS		
	X	X^2 4
	7	49
	3	9
	1	1
	0	0
	4	16
1 $n = 5$	2 $\Sigma X = 15$	5 $\Sigma X^2 = 75$
	3 $(\Sigma X)^2 = 225$	
6 $SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 75 - \frac{225}{5} = 75 - 45 = 30$		
7 $s^2 = \frac{SS}{n-1} = \frac{30}{4} = 7.50$	8 $s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{30}{4}} = \sqrt{7.50} = 2.74$	

→ **Problem:** Using the computation formula for the sum of squares, calculate the population standard deviation for the scores in (a) and the sample standard deviation for the scores in (b).

(a) 1, 3, 7, 2, 0, 4, 7, 3 (b) 10, 8, 5, 0, 1, 1, 7, 9, 2

Solution:

$$(a) \sigma = \sqrt{\frac{137 - \frac{729}{8}}{8}} = \sqrt{5.73} = 2.39$$

$$(b) s = \sqrt{\frac{325 - \frac{1849}{9}}{9-1}} = \sqrt{14.95} = 3.87$$

Degrees of Freedom (Df)

- Degrees of freedom (df) refers to the number of values that are free to vary, given one or more mathematical restrictions, in a sample being used to estimate a population characteristic.
- The notion of degrees of freedom is used extensively in inferential statistics.
- In any event, however, degrees of freedom (*df*) always indicate the number of values that are free to vary, given one or more mathematical restrictions, in a set of values used to estimate some unknown population characteristic.
- When n deviations about the sample mean are used to estimate variability in the population, only $n - 1$ are free to vary. As a result, there are only $n - 1$ degrees of freedom, that is, $df = n - 1$. One df is lost because of the zero-sum restriction.
- **Degrees of freedom**, often represented by df , is the number of independent pieces of information used to calculate a statistic. It's calculated as the sample size minus the number of restrictions.
- Degrees of freedom refer to the number of values free to vary in the sample, not in the population.
- All observations are assumed to be equal in quality. Degrees of freedom are introduced because of mathematical restrictions when sample observations are used to estimate a population characteristic.
- One degree of freedom is lost because, when expressed as a deviation from the sample mean, the final deviation in the sample fails to supply information about population variability.
- Degrees of freedom make sense only if we wish to estimate some unknown characteristic of a population.

- **Example:** Consider a data sample consisting of five positive integers. The values of the five integers must have an average of six. If four of the items within the data set are {3, 8, 5, and 4}, the fifth number must be 10. Because the first four numbers can be chosen at random, the degree of freedom is 4.
- **Problem:** As a first step toward modifying his study habits, Phil keeps daily records of his study time.
- (a) During the first two weeks, Phil's mean study time equals 20 hours per week. If he studied 22 hours during the first week, how many hours did he study during the second week?
- (b) During the first four weeks, Phil's mean study time equals 21 hours. If he studied 22, 18, and 21 hours during the first, second, and third weeks, respectively, how many hours did he study during the fourth week?
- (c) If the information in (a) and (b) is to be used to estimate some unknown population characteristic, the notion of degrees of freedom can be introduced. How many degrees of freedom are associated with (a) and (b)?
- (d) Describe the mathematical restriction that causes a loss of degrees of freedom in (a) and (b).

Solution:

- (a) 18 hours
- (b) 23 hours
- (c) $df = 1$ in (a) and $df = 3$ in (b)
- (d) When all observations are expressed as deviations from their mean, the sum of all deviations must equal zero.

Inter Quartile Range (IQR)



- The most important spinoff of the range, the **inter quartile range (IQR)**, is simply the range for the middle 50 percent of the scores.
- More specifically, the IQR equals the distance between the third quartile (or 75th percentile) and the first quartile (or 25th percentile), that is, after the highest quarter (or top 25 percent) and the lowest quarter (or bottom 25 percent) have been trimmed from the original set of scores.
- Since most distributions are spread more widely in their extremities than their middle, the IQR tends to be less than half the size of the range.
- A key property of the IQR is its resistance to the distorting effect of extreme scores, or outliers.
- If we are concerned about possible distortions caused by extreme scores, or outliers, we can use the IQR as the measure of variability, along with the median (or second quartile) as the measure of central tendency.

Table 4.6
CALCULATION OF THE IQR

A. INSTRUCTIONS

- 1 Order scores from least to most.
- 2 To determine how far to penetrate the set of ordered scores, begin at either end, then add 1 to the total number of scores and divide by 4. If necessary, round the result to the nearest whole number.
- 3 Beginning with the largest score, count the requisite number of steps (calculated in step 2) into the ordered scores to find the location of the third quartile.
- 4 The third quartile equals the value of the score at this location.
- 5 Beginning with the smallest score, again count the requisite number of steps into the ordered scores to find the location of the first quartile.
- 6 The first quartile equals the value of the score at this location.
- 7 The IQR equals the third quartile minus the first quartile.

B. EXAMPLE

- 1 7, 9, 9, 10, 11, 11, 13
- 2 $(7 + 1)/4 = 2$
- 3 7, 9, 9, 10, 11, 11, 13

- 4 third quartile = 11
- 5 7, 9, 9, 10, 11, 11, 13

- 6 first quartile = 9
- 7 $IQR = 11 - 9 = 2$

→ **Problem:** Determine the values of the range and the IQR for the following sets of data.

(a) Retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63

(b) Residence changes: 1, 3, 4, 1, 0, 2, 5, 8, 0, 2, 3, 4, 7, 11, 0, 2, 3, 4

Solution:

(a) range = 25; $IQR = 65 - 60 = 5$

(b) range = 11; $IQR = 4 - 1 = 3$

Variability for Qualitative and Ranked Data

Qualitative Data

→ Measures of variability are virtually nonexistent for qualitative or nominal data.

→ It is probably adequate to note merely whether scores are evenly divided among the various classes (maximum variability), unevenly divided among the various classes (intermediate variability), or concentrated mostly in one class (minimum variability).

- For example, if the ethnic composition of the residents of a city is about evenly divided among several groups, the variability with respect to ethnic groups is maximum; there is considerable heterogeneity.
- At the other extreme, if almost all the residents are concentrated in a single ethnic group, the variability will be minimum; there is little heterogeneity.
- If the ethnic composition falls between these two extremes—because of an uneven division among several large ethnic groups—the variability will be intermediate, as is true of many U.S. cities and counties.

Ordered Qualitative and Ranked Data

- If qualitative data can be ordered because measurement is ordinal (or if the data are ranked), then it's appropriate to describe variability by identifying extreme scores (or ranks).
- For instance, the active membership of an officers' club might include no one with a rank below first lieutenant or above brigadier general.

Tutorial Questions:

1. What is frequency distribution? What are uses of frequency distributions? Outline the different types of frequency distributions with examples.
2. Analyze how graphs can be used to represent qualitative and quantitative data.
3. What are the outliers in the data? How to deal with outliers.
4. What is steam and leaf display? Outline the procedure to construct steam and leaf display. Construct a stem and leaf display for the following IQ scores obtained from a group of four-year-old children.

120	98	118	117	99	111
126	85	88	124	104	113
108	141	123	137	78	96
102	132	109	106	143»	

5. Describe different types of typical frequency distributions shapes with neat diagrams.
6. What is a measure of central tendency? Explain three different measures of central tendency with their uses in describing data.
7. What is a measure of variability? Illustrate the importance of variability. Describe different measures of variability, with their uses in describing data.
8. Specify an important difference between the standard deviation and the mean. Why can't the value of the standard deviation ever be negative?

9. Write step-by-step procedure to Construct Frequency Distributions.
The IQ scores for a group of 35 high school dropouts are as follows:

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

Construct a frequency distribution for grouped data and then convert to a relative frequency distribution, to a cumulative frequency Distribution, to a cumulative percent frequency distribution.

10. Write short note on the following:
- Frequency Distributions for Qualitative Data
 - Graphs for Qualitative data
 - Averages for Qualitative data
 - Variability for Qualitative data.