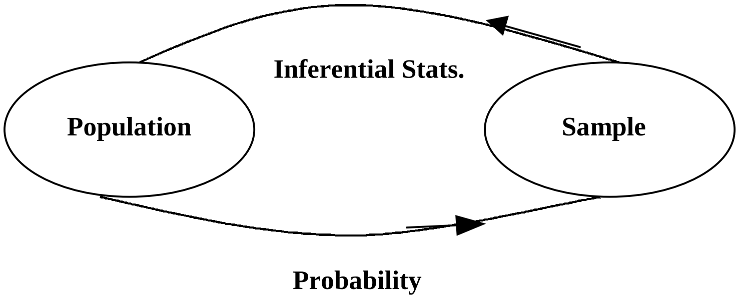
DESCRIPTIVE STATISTICS

Statistics – deals with the theories and methods used in the collection, organization, interpretation and presentation of data.
Data – raw material used in statistical investigation

Origins of Modern Statistics

1. government (political science)
2. games of chance

Descriptive Statistics – covers summarizing or describing data without attempting to infer anything that goes beyond the data.
Inferential Statistics – methods used when the data is a sample and the objective is to go beyond the sample and draw conclusions about the population.



Relationship between probability and inferential statistics

Nature of Statistical Data

1. **Nominal data** – most limited type of measurement; merely used to differentiate classes or categories for purely classification or identification purposes.
2. **Ordinal data** – do not only classify but also order the classes. It is expressed in ranks and is possible if different degrees of a property are present.
3. **Interval data** – has the attributes of ordinal data plus the ability to differentiate between any two classes in terms of degrees of differences.

   Note: The zero point of the interval scale is arbitrary.

4. **Ratio data** – differs from the interval data only in one aspect: it has a true zero which indicates the total absence of the property being measured.

Exercises:

1. The paid attendance at a small college's home football game was

   12305 10984 6850 11733 10641

   Which of these conclusions can be obtained from these figures by purely descriptive methods and which require generalizations?

a) The attendance at the third home game was low because it rained.

b) Among the games, the paid attendance was highest at the first game.

c) The paid attendance increased from the third home game to the fourth home game because the college's football team had been wining.

d) The paid attendance exceeded 11000 at two of the five games.

2. Identify the nature of the following statistical data:

a) SSS number
b) 1st, 2nd and 3rd place in a lantern contest
c) physics test scores
d) IQ
e) number of live births in a month
f) socio-economic status: 1-high, 2-low, 3-average
g) religious affiliations: 1-Catholic, 2- non-Catholic
h) metric measurements
i) Mohs' scale of hardness

PRESENTATION OF DATA

A. SUMMARIZING DATA

DISPLAYING NUMERICAL VALUES

Sorting a large set of numbers into increasing or decreasing order is a difficult task.

Stem-and-leaf diagram – a technique used in sorting numbers; devised by Prof. John Tukey of Princeton University in the late 1960s.

Procedures:

1. Select one or more leading digits in stem values. The trailing digit or digits become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate units for display of stem and leaves.

Outliers – data that are much smaller than or much larger than the bulk of the data.

Drill:

3. The following are the weights in pounds of 20 applicants for jobs with a city's fire department:

225	182	194	210	205	172	181	198	164	176
180	193	178	193	208	186	183	170	186	188

Construct a stem-and-leaf display with stem labels 16, 17, 18, 19, 20, 21 and 22.

B. FREQUENCY DISTRIBUTIONS

In the field of statistics, one mechanism for reducing and summarizing data is the frequency distribution.

Frequency distribution – a table conveyed by grouping data into a number of classes, intervals or categories.

Raw data – data in its ungrouped or original form.

Types of Frequency Distributions:

1. **Numerical or quantitative distribution** – when data are grouped in size.

No. of Subscribers	No. of Newspapers
Less than 1000	244
1000 – 3499	157
3500 – 9999	96
10000 – 19999	37
20000 - 49999	24
50000 or more	6
Total	564

2. **Categorical or qualitative distribution** – when the data are grouped into non-numerical categories.

Rank	Language	Speakers (millions)
1	Mandarin	930
2	English	463
3	Hindi	400
4	Spanish	371
5	Russian	291
6	Arabic	214
7	Bengali	192

Principal languages of the world, 1993

The construction of a frequency distribution consists essentially of three steps:

1) Choosing the classes (intervals or categories)
2) Sorting the data into these classes.
3) Counting the number of items in each class.

For step 1: Choice of class is essentially arbitrary but the following rules are usually observed:

a) We seldom used fewer than six or more than 15 classes; the exact number we use in a given situation depends largely on how many observations there are.
b) We make sure that each item goes into one and only one class.
c) Whenever feasible, all classes should cover equal ranges of values.

Open classes – classes of the "less than", "or less", "more than" or "more." Generally these should be avoided because they make it impossible to calculate certain fields of interest.

Drill 4: Construct a frequency distribution of the following amounts of sulfur oxides (in tons) emitted by an industrial plant in 80 days:

15.8	26.4	17.3	11.2	23.9	24.8	18.7	13.9	9.0	13.2
22.7	9.8	6.2	14.7	17.5	26.1	12.8	28.6	17.6	23.7
26.8	22.7	18.0	20.5	11.0	20.9	15.5	19.4	16.7	10.7
19.1	15.2	22.9	26.6	20.4	21.4	19.2	21.6	16.9	19.0
18.5	23.0	24.6	20.1	16.2	18.0	7.7	13.5	23.5	14.5
14.4	29.6	19.4	17.0	20.8	24.3	22.5	24.6	18.4	18.1
8.3	21.9	12.3	22.3	13.3	11.8	19.3	20.0	25.7	31.8
25.9	10.5	15.9	27.5	18.1	17.9	9.4	24.1	20.1	28.5

Answer:

Tons of Sulfur Oxides	Frequency
5.0 –8.9	3
9.0 –12.9	10
13.0 –16.9	14
17.0 – 20.9	25
21.0 – 24.9	17
25.0 – 28.9	9
29.0 – 32.9	2

Total ***80***

Grouping terminology

Classes – categories for grouping data

Frequency – the number of pieces of data in a class

Class limits:

Upper class limit – the largest value that can go in a class

Lower class limit – the smallest value that can go in a class

Class mark – the midpoint of a class

Class width or class interval – the difference between the lower class limit of the given class and the lower class limit of the next higher class

Class boundaries:

Lower class boundary – the midpoint between the lower class limit of the class and the upper class limit of the next lower class

Upper class boundary - the midpoint between the upper class limit of the class and the lower class limit of the next higher class

Modifications to frequency distributions

1. **Percentage distribution** – done by dividing each class frequency by the total number of items grouped and multiplying by 100%
2. **Cumulative distribution** – converting the frequency distribution to a "less than" or "more than" distribution. To construct a cumulative distribution, simply add the class frequencies starting either at the top or bottom of the distribution.

Drill 5 : Convert the distribution of sulfur oxides emission data into a percentage dist. and a less than cumulative distribution.

Answer: *Percentage distribution*

Tons of Sulfur Oxides	Percentage
5.0 – 8.9	3.75
9.0 – 12.9	12.50
13.0 – 16.9	17.50
17.0 – 20.9	31.25
21.0 – 24.9	21.25
25.0 – 28.9	11.25
29.0 – 32.9	2.50
Total	100.00

Less than cumulative distribution

Tons of Sulfur Oxides	Cumulative Frequency
Less than 5.0	*0*
Less than 9.0	*3*
Less than 13.0	*13*
Less than 17.0	*27*
Less than 21.0	*52*
Less than 25.0	*69*
Less than 29.0	*78*
Less than 33.0	*80*

C. GRAPHICAL DEPICTION OF DATA

Presentation of statistical data by graphic depictions can often communicate information in a manner that is effective, efficient and meaningful to the receiver.

1. **Histogram** – a mere bar diagram where the bars are adjacent and the base extends from the lower limit of a class to its upper limit. The horizontal scale represents the measurements that are grouped. The vertical scale represents the height of the class frequencies.

   Note: Histograms cannot be drawn for distributions with open classes.
2. **Bar chart** – similar to histogram but there is no pretense of having a continuous horizontal scale.
3. **Line diagram** – simplest graph to construct and is primarily used for estimates, forecast and for interpolation.
4. **Frequency polygon** – a graph in which the line segments “connecting the dots” depict a frequency distribution. Here, class frequencies are plotted at the class marks and the successive points are connected by straight lines. Classes with zero frequencies are added at both ends of the distribution to tie down the graph to the horizontal scale.
5. **Ogive** – also called the cumulative frequency polygon. In an ogive, cumulative frequencies are plotted at the class boundaries instead of the class marks.

D. Pictorial Descriptions

1. **Pie chart** – a circle divided into sectors which are proportional in size to the corresponding frequencies or percentages.
2. **Pictogram** – immediately suggests the nature of the data shown; combines the attention of getting the quality of a picture and accuracy of a bar chart.
3. **Statistical map** – used when quantitative data have to be shown by geographical location.

DESCRIPTIVE MEASURES

E. Measures of Central tendency

- statistical measures which describe the center or the middle of the data.

Population – set of data consisting of all conceivable possible (or hypothetically possible) observations of a given phenomenon.
Sample – set of data consisting of only a part of these observations.

Measures of Central tendency for ungrouped data

1. **Mode** – most frequently occurring value in a set of data
   Advantages:
   1. It requires no calculation, only counting.
   2. The mode can be determined even for nominal data.

   Disadvantage: It may not exist.

2. **Median** - the middle value in an ordered array of numbers.
   Array – an ordering of the numbers in magnitude from smallest to largest.
   If we denote *n* as the number of pieces of data, then the median is at position (n + 1) / 2 in the ordered list.

The median is unaffected by the magnitude of extreme values. For this reason, the median is often the best measure of location to use in the analysis of variables such as house cost, income and age.

A disadvantage of the median is that not all the information from the numbers is used.

3. **Mean** (or ***arithmetic mean***) – synonymous to average and computed by summing all numbers and dividing by the number of items.

Sample Mean, $\overline{x} = \frac{\sum x}{n}$ Population Mean, $\mu = \frac{\sum x}{N}$

The description of a population is known as a **parameter** and the description of a sample is known as a **statistic**. Parameters are denoted by Greek letters.

From $\overline{x} = \frac{\sum x}{n}$, it follows that $\sum x = \overline{x} * n$, and therefore, no single x-value can exceed $n * \overline{x}$.

Drill problems:

6. *Climbing* magazine gave the weights (in ounces) of 15 popular climbing shoes. The data are as follows: 20,14, 20, 13, 15, 20, 22, 16, 16, 18, 16, 14, 12, 16 and 18. Calculate the a) mean weight; b) median weight; c) modal weight.
7. A bridge is designed to carry a maximum load of 150,000 lbs. Is the bridge overloaded if it is carrying 18 vehicles having a mean weight of 4630 lbs?

Properties of the Mean:

1. It can be calculated from any set of numerical data, so it always exists.
2. A set of data has one and only one mean, so it is always unique.
3. It leads itself to further statistical treatment.
4. Reliable, since the mean do not usually fluctuate widely as other statistical measures.

Weighted Mean

To give quantities being averaged their proper degree of importance, it is often necessary to assign them (relative importance) *weights*, and then calculate a weighted mean.

Weighted Mean $\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_1 + w_2 + ... + w_n} = \frac{\Sigma w \bullet x}{\Sigma w}$

where $\Sigma w \bullet x$ = the sum of the products obtained by multiplying each x by the corresponding weight.

Grand Mean of combined data, $\bar{\bar{x}}$

A special application of the formula for the weighted mean arises when we must find the overall mean, or grand mean, of k sets of data having the means $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ and consisting of n_1, n_2, . . . n_k observations. The result is given by:

$$\bar{\bar{x}} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + ... n_k \bar{x}_k}{n_1 + n_2 + ... n_k} = \frac{\Sigma\, n * \bar{x}}{\Sigma\, n}$$

Drill Problems:

8. The average quarterly salaries of elementary school teachers in three cities are

$38,300 $44,500 $41,000

Given that there are 720, 660 and 520 elementary school teachers in these cities, find their average quarterly salary. *Ans: $41,192.63*

9. A sample survey of students in a junior high school yields the following data on the mean daily time (in hours) spent in doing homework:

Grade	No. of students in the sample	Mean Homework Hours
7	80	0.9
8	70	0.7
9	65	1.0

Find the mean for all 215 students in the sample. *Ans: 0.865 hour*

Considerations in the Use of a Measure of Central Tendency

1. Use the mean when
   a. greatest reliability is desired
   b. there is a need for further statistical computation
   c. distribution is symmetrical about the center
   d. data are ratio or interval
2. Use the median when
   a. distribution is badly skewed
   b. we are interested in determining whether cases fall within the upper half or lower half of the distribution
   c. data are ordinal or ranked
3. Use the mode when
   a. the quickest estimate of central value is wanted
   b. a rough estimate of central value is wanted
   c. data are nominal or categorical by nature

Some Special Averages

1. **Geometric mean** – The geometric mean of a given set of positive values $x_1, x_2, \ldots, x_n$ is the n^{th} root of the product of the n values

$$\mathbf{GM} = \sqrt[n]{x_1 * x_2 * \ldots * x_n}$$

Generally, the geometric mean is used in finding the average of ratios, estimating the average rate of change in a set of variates constituting a time series and in computing an average for a series of values which are in geometric progression. The geometric mean is used when relative changes among the variates are more significant than absolute changes.

2. **Quadratic mean –** also called the root-mean square (RMS); computed using the formula

$$\mathbf{RMS} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \sqrt{\frac{\sum x^2}{n}}$$

This type of average finds application in the physical sciences.

3. **Harmonic mean** – defined as the reciprocal of the arithmetic mean of the reciprocals of the given values

$$\textbf{HM} = \frac{N}{\sum \frac{1}{x}}$$

When any of the variates is either zero or negative, the harmonic mean cannot be computed.

A common application of the harmonic mean is in connection with "work limit" tests. In such tests, the score is the amount of time required to finish a fixed quantity of work, or the scores may be in terms of units of work accomplished in a fixed time, then such are called as "time limit" tests.

Relationship among the special averages: AM > GM > HM

Drill problems:

10. Find the geometric mean of 1, 2, 8 and 16 Ans: 4

11 a. If an investor buys $18,000 worth of company' stock at $45 a share and then buys $18,000 worth at $36 a share, find the average price that the investor had paid per share. Ans: $40

b. If a bakery buys $36 worth of an ingredient at 60¢ per pound, $36 worth at 72¢ per pound and $36 worth at 90¢ per pound, what is the average cost per pound? Ans: 72¢

Percentiles – are measures of location that divide a group of data into 100 parts. There are 99 percentiles, because it takes 99 dividers to separate a group of data into 100 parts.

The n^{th} percentile is the value such that at most *n* percent of the data are below that value and at most (*100 – n*) percent are above that value.

Percentiles are stair-step values and they are practically used in reporting test results.

Steps in determining the location of a percentile

1. Organize the numbers into an ascending-order array.
2. Calculate the percentile location by:

$$i = \frac{p}{100}(n) \quad \text{where}$$

p = percentile of interest n = number in the data set

i = percentile location

3. Determine the location by either (a) or (b)

a. If i is a whole number, the pth percentile is the average of the value at the ith location and the value at the (i + 1)st location.

b. If i not a whole number, the pth percentile value is located at the whole number part of (i + 1).

Note: A percentile may or may not be one of the data values.

Quartiles – are measures of location that divide a group of data into four subgroups. There are three quartiles, denoted by Q_1, Q_2 and Q_3.

The first quartile separates the lowest one-fourth of the data from the upper three-fourths and is equal to the 25th percentile (Q_1=P_{25}). The second quartile separates the second quarter of the data from the third quarter. Q_2 is located at the 50th percentile and equals the median of the data, (Q_2=P_{50}=$\tilde{x}$). The third quartile divides the first 3 quarters of the data from the last quarter and is equal to the 75th percentile, ($Q_{3=}P_{75}$).

Deciles – partition the ranked data into ten groups which are about 10% of the data in each group.

Drill Problem 12: The library records of a large university show that 22 senior philosophy majors checked out the following number of books during the academic year: 62, 65, 73, 69, 40, 82, 72, 50, 79, 66, 88, 103, 35, 68, 51, 54, 48, 38, 42, 52, 75, 72. Compute for the (a) Q_3; (b) Q_2; (c) D_4; (d) P_{80}.

Measures of central tendency for grouped data

1. **Mean**

$$\bar{x} = \frac{\sum f * M}{N}$$

where: f = class frequency N = number of observations
M = midpoint of each class

2. **Median**

$$\tilde{x} = L + (\frac{\frac{N}{2} - F_b}{f_m})(i)$$

where: L = lower class boundary where the median lies
F_b = sum of all frequencies below L
f_m = frequency of the class containing the median
N = number of observations
i = class interval

3. **Mode**

The mode for grouped data is the class midpoint of the modal class. The modal class is the class interval with the greatest frequency.

4. **Percentiles**

$$P_p = L + (\frac{pN - f_c}{f})(i)$$

where: P_p = desired percentile
L = lower class boundary of class containing P_p
N = number of observations
P = proportion of the class exceeded by P_p; e.g. 0.30 for the P_{30}
f_c = cumulative frequency below the class containing the P_p
f = frequency of the class containing the P_p
i = class interval size

Drill Problems:

13. Given the distribution of the ages of the members of a labor union:

Age	Frequency (f)	M	f * M	<cf
15-19	18			
20-24	42			
25-29	78		2106	138
30-34	115		3680	253
35-39	178		6586	431
40-44	107		4494	538
45-49	88		4136	626
50-54	52		2704	678
55-59	30		1710	708
60-64	11		682	719
	719		$\sum f*M = 27328$	

Find the (a) mean; (b) median; (c) mode; (d) second quintile; (e) third quintile.

14. Compute the mean, median and mode of the sampled data:

Class Interval	Frequency
10 - under 15	6
15 – under 20	22
20 – under 25	35
25 - under 30	29
30 – under 35	16
35 – under 40	8
40 – under 45	4

F. Measures of dispersion (or measures of variability)

Measures of location do not yield sufficient information to describe the data set. What is needed is a second dimension of measurement, a measure of the variability or dispersion of the data.

Measures of dispersion for ungrouped data

1. **Range** – the difference between the largest value of a data and the smallest value.

An advantage of the range is its ease of computation. One important use of the range is in quality assurance, where the range is used to construct control charts.

A disadvantage of the range is that it is affected by extreme values.

2. **Interquartile range** – the range of values between the first and third quartile. This is useful in situations where data users are more interested in values toward the middle and less interested in the extremes.

$$IQR = Q_3 - Q_1$$

Semi-interquartile range or Quartile deviation = $\frac{1}{2}(Q_3 - Q_1)$

Suppose a company has started a production line to build computers. During the first five weeks of operation, the output is 5, 9, 16, 17 and 18 computers. One way to look at the spread of the data is to subtract the mean from each data value. This would yield *deviations from the mean*.

For any given set of data, ***the sum of all deviations from the arithmetic mean is always zero.***

This property requires considerations of alternative ways to obtain measures of variability.

3. **Mean absolute deviation (MAD)** – the average of the absolute deviations from the mean

$$MAD = \frac{\sum|x-\mu|}{N}$$

Because it is computed by using absolute values the MAD is useful in statistics than other measures of dispersion. However, in the field of forecasting, it is occasionally used as a measure of error.

4. **Variance** – the average of the squared deviations from the arithmetic mean

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

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

Ordinarily, the purpose of calculating a sample statistic is to estimate the corresponding population parameter.

Estimates which have the desirable property that their values will on the average equal the quantity that they are supposed to estimate are said to be ***unbiased***; otherwise, they are ***biased***.

Because the variance is computed from squared deviations, the final result is expressed in terms of squared units of measurement.

5. **Standard deviation** – the square root of the variance

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

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

6. **Coefficient of variation** – ratio of the standard deviation to the mean expressed as a percentage

$V = \frac{\sigma}{\mu}(100)$ $V = \frac{s}{\bar{x}}(100)$ - for samples

Drill Problems:

15. The number of misdirected luggage cases reported for six consecutive weeks at a small airport were 13, 8, 15, 11, 3 and 10. From these samples, find the (a) standard deviation; (b) variance.

16. One patient's blood pressure was measured daily for several weeks. These measurements had mean 188 with standard deviation 14.2 A second patient was also measured daily, obtaining an average 136 with standard deviation 8.6. Which patient's blood pressure is relatively more variable?
Ans: for patient A = 7.55%, for B = 6.32%

Applications of the standard deviation

A. **Empirical rule**

The empirical rule is a guideline that states the approximate percentage of the values that are within a given number of standard deviations of the mean of a set of data if the data are normally distributed. The empirical rule is used only for three numbers of standard deviations: 1σ, 2σ and 3σ.

It says that if a set of data is normally distributed or bell-shaped, *approximately 68% of the values will lie within one standard deviation of the mean; about 95% will lie within two standard deviations of the mean; about 99.7% will lie within three standard deviations of the mean.*

B. **Chebyshev's theorem**

The Russian mathematician Pafnuty Lvovich Chebyshev developed a rule that applies to all distributions regardless of shape.

It states that for any set of data (population or sample) and any constant ***k*** greater than 1, the proportion (percentage) of the data that must lie within ***k*** standard deviations on either side of the mean is **at least 1 -** $\frac{1}{k^2}$**.**

C. **z-score**

A z-score represents the number of standard deviations a value (x) is above or below the mean. Using z-score allows translation of a value's raw distance from the mean into units of standard deviations.

$$z = \frac{x-\mu}{\sigma} \quad \text{or} \quad z = \frac{x-\bar{x}}{s} \text{ (for samples)}$$

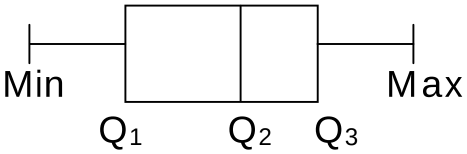
If a z-score is negative, the raw value (x) is below the mean. If the z-score is positive, the raw value (x) is above the mean.

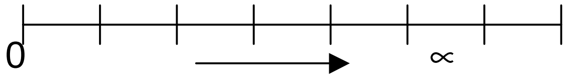
Five-number summary

The five-number summary of a data set consists of the minimum, maximum and quartiles written in increasing order: Min, Q_1, Q_2, Q_3 and Max.

Box-and-whisker diagram (boxplot) - a graphical representation of the center and variation of the data set based on the five-number summary.

Basic representation:





Scale of data to be represented

Drill Problems:

17. The following are the number of accidents that occurred in July 1990 in a certain town at 18 intersections without left-turn arrows:

8 29 31 14 35 28 12 18 22
13 6 32 2 10 26 22 32 25

Draw a boxplot for the accident data.

18. Suppose that the average fuel usage rate for automobiles in the U.S. is 16.85 mi/gal, with a standard deviation of 4.7 mi/gal. Suppose further that the distribution of fuel rates is unknown. Within what fuel rate limits would at least 85% of the values be? Ans: 5.1 – 28.6 mi/gal

19. From problem 18, if the distribution of fuel rates is bell-shaped, between what fuel rate limits would (a) 68%, (b) 95%, (c) 99.7% of the cars be? 12.15-21.55; 7.45-26.25; 2.75-30.95

20. Shown below are the top 12 car rental companies in the US ranked by number of locations. Determine the z-value of Hertz in this group. Determine the z-value for Thrifty. What do these value tell us? Data represents the population.

Company	Number of locations
Hertz	5,400
Avis	4,800
National	4,500
Sears	3,097
Enterprise	1,700
Dollar	1,087
Thrifty	1,039
U-Save	500
Agency	475
Carey	393
Payless	146
Alamo	140

Source: Business travel News, 1996 Business Almanac

Ans: Hertz = 1.83; Thrifty = -0.48

Measures of dispersion for grouped data

$$\sigma^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{N}}{N} \qquad s^2 = \frac{\sum fM^2 - \frac{(\sum fM)^2}{n}}{n-1}$$

For the standard deviation, the square root of the variance is obtained.

Drill Problem 21: Given the frequency distribution for the IQs of 112 children attending a kindergarten school in California, find the sample standard deviation.

Ans: 16.38

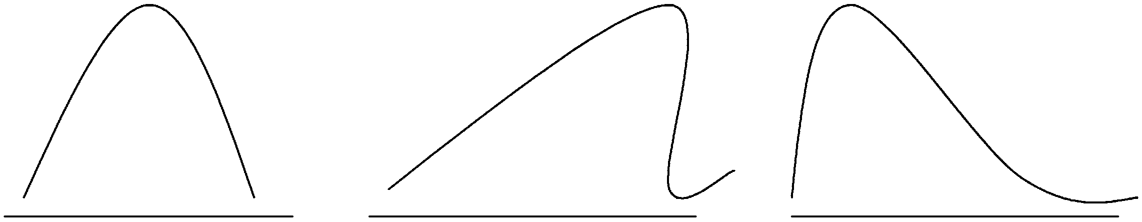
IQ	f	M	M^2	fM	fM^2
60-69	1				
70-79	5				
80-89	13	84.5	7140.25	1098.5	92823.25
90-99	22	94.5	8930.25	2079	196465.5
100-109	28	104.5	10920.25	2926	305767
110-119	23	114.5	13110.25	2633.5	301535.75
120-129	14	124.5	15500.25	1743	217003.5
130-139	3	134.5	18090.25	403.5	54270.75
140-149	2	144.5	20880.25	289	41760.5
150-159	1	154.5	23870.25	154.5	23870.25
	N= 112			$\Sigma fM = 11764$	1265408

G. Measures of shape

Measures of shape are tools which can be used to describe the shape of a distribution of data

Skewness - is exhibited when a distribution is asymmetrical or lacks symmetry.

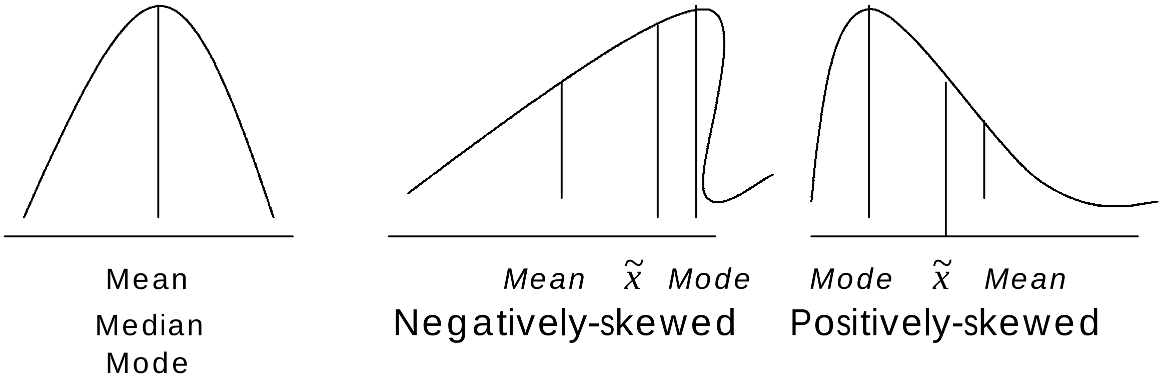
Bell-shaped distribution- a distribution symmetric about the center.



Symmetric Negatively-skewed Positively-skewed

The skewed portion is the long, thin part of the curve.

Relationship of the mean, median and mode on shape of distributions



1. **Pearsonian coefficient of skewness**

$$SK = \frac{3(mean - median)}{std.dev'n}$$

For symmetric distributions, SK = 0. If SK is positive, distribution is positively skewed. If SK is negative, the distribution is negatively skewed. The greater the magnitude of distribution, the more skewed is the distribution.

In general, its value must lie between –3 and +3.

2. **Kurtosis** - degree of peakedness and flatness of a distribution

$$Ku = \frac{1}{2}(P_{90} - P_{10})$$

If Ku = 3, the distribution is normal or *mesokurtic*.

If Ku < 3, the distribution is *platykurtic* or less peaked than the normal curve

If Ku > 3, the distribution is *leptokurtic* or more peaked than the normal curve.

Drill Problems:

22. The following are the number of false alarms which a fire department recorded during 17 weeks: 8, 3, 12, 5, 6, 12, 6, 3, 4, 11, 8, 7, 5, 6, 8, 8, 4. Discuss the symmetry or skewness of these data. Ans: slight positive skewness

23. The distribution of the number of mistakes made by 200 students taking German in a multiple-choice quiz on vocabulary is as follows:

No. of mistakes	f	M	M^2	<cf	fM	fM^2
6-10	12					
11-15	73					
16-20	52	18	324	137	936	16848
21-25	39	23	529	176	897	20631
26-30	24	28	784	200	672	18816
	200				$\Sigma fM=3550$	$\Sigma fM^{2=}69400$