Chapter 4
Fundamentals of Statistics

Fundamentals of Statistics

- Introduction
- Frequency Distribution
- Measure of Central Tendency
- Measure of Dispersion
- Other Measure
- Concept of a population and a sample
- The normal curve
- Test for Normality
- Computer program

Definition of *statistics*

1. A collection of quantitative data pertaining to any subject or group, especially when the data are systematically gathered and collated.
2. The science that deals with the collection, tabulation, analysis, interpretation, and presentation of quantitative data.

Collection of data

1. Variable are those quality characteristics that are measurable, such as weight measured in grams.
2. Attributes, on the other hand, are those quality characteristics that are classified as either conforming or not conforming to specifications such as a "go/ no go gage."

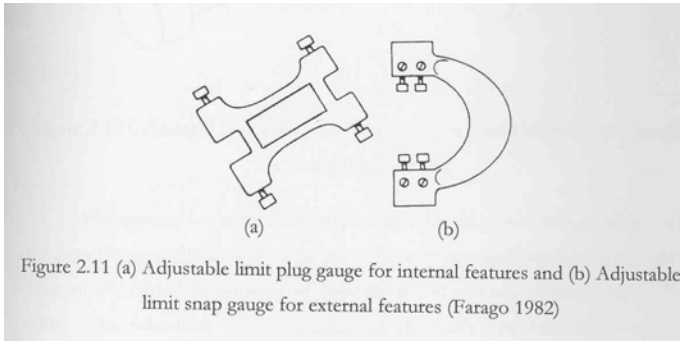


Figure 2.11 (a) Adjustable limit plug gauge for internal features and (b) Adjustable limit snap gauge for external features (Farago 1982)

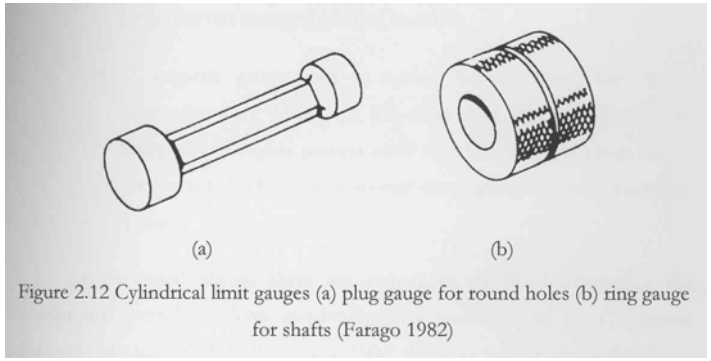


Figure 2.12 Cylindrical limit gauges (a) plug gauge for round holes (b) ring gauge for shafts (Farago 1982)

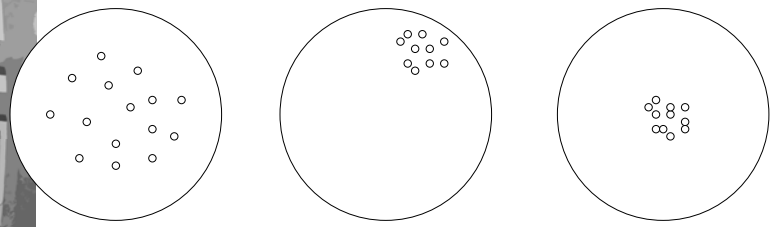
- In collecting data the number of figure is a function of the intended use of the data.

  example
- the life of light bulbs, it is acceptable to record 995.6h. Recording a value of 995.632 is too accurate and unnecessary.
- Specification has lower limit of 9.52 mm and upper limit of 9.58 mm.

0.001 ->0.01

Measuring instruments

may not give a true reading because of problems due to accuracy and precision.



Accurate **Precise** **Accurate and Precise**

Describing the data

TABLE 3-1 **Number of Daily Billing Errors.**

0	1	3	0	1
1	5	4	1	2
1	0	2	0	0
2	1	1	1	2
0	4	1	3	1
1	3	4	0	0
1	3	0	1	2

Two techniques are available to accomplish this summarization of data- *graphical* and *analytical*.

Unorganized data are virtually meaningless.

Graphical

The graphical technique is a plot or picture of a *frequency distribution*, which is a summarization of how the data points occur within each subdivision of observed values or groups of observed values.

Analytical

summarize data by computing a *measure of central tendency* and a *measure of the dispersion*.

Frequency Distribution

Ungrouped Data

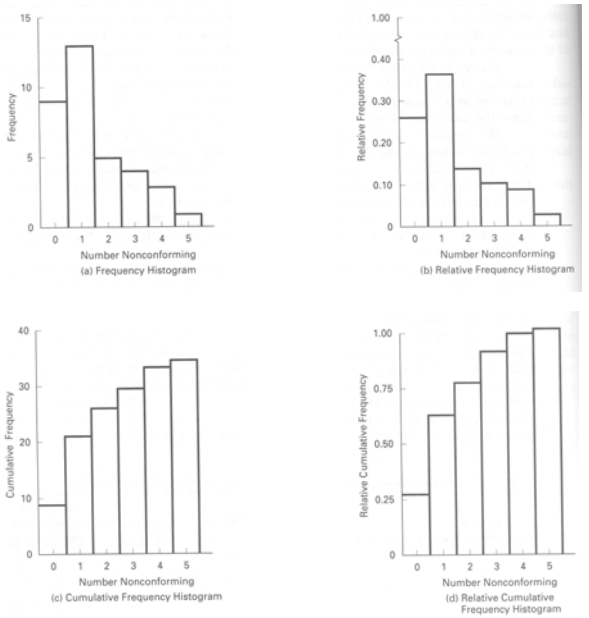
comprise a listing of the observed values.

TABLE 3-2 Tally of Number of Daily Billing Errors.

NUMBER NONCONFORMING	TABULATION	FREQUENCY
0	𝍸 IIII	9
1	𝍸 𝍸 III	13
2	𝍸	5
3	IIII	4
4	III	3
5	I	1

TABLE 3-3 Different Frequency Distributions of Data Given in Table 3-1.

NUMBER NONCONFORMING	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY	RELATIVE CUMULATIVE FREQUENCY
0	9	9 ÷ 35 = 0.26	9	9 ÷ 35 = 0.26
1	13	13 ÷ 35 = 0.37	9 + 13 = 22	22 ÷ 35 = 0.63
2	5	5 ÷ 35 = 0.14	22 + 5 = 27	27 ÷ 35 = 0.77
3	4	4 ÷ 35 = 0.11	27 + 4 = 31	31 ÷ 35 = 0.89
4	3	3 ÷ 35 = 0.09	31 + 3 = 34	34 ÷ 35 = 0.97
5	1	1 ÷ 35 = 0.03	34 + 1 = 35	35 ÷ 35 = 1.00
Total	35	1.00		



Grouped Data

represent a lumping together of the observe values.

TABLE 3-4 Steel Shaft Weight (kilograms).

2.559	2.556	2.566	2.546	2.561
2.570	2.546	2.565	2.543	2.538
2.560	2.560	2.545	2.551	2.568
2.546	2.555	2.551	2.554	2.574
2.568	2.572	2.550	2.556	2.551
2.561	2.560	2.564	2.567	2.560
2.551	2.562	2.542	2.549	2.561
2.556	2.550	2.561	2.558	2.556
2.559	2.557	2.532	2.575	2.551
2.550	2.559	2.565	2.552	2.560
2.534	2.547	2.569	2.559	2.549
2.544	2.550	2.552	2.536	2.570
2.564	2.553	2.558	2.538	2.564
2.552	2.543	2.562	2.571	2.553
2.539	2.569	2.552	2.536	2.537
2.532	2.552	2.575 (H)	2.545	2.551
2.547	2.537	2.547	2.533	2.538
2.571	2.545	2.545	2.556	2.543
2.551	2.569	2.559	2.534	2.561
2.567	2.572	2.558	2.542	2.574
2.570	2.542	2.552	2.551	2.553
2.546	2.531 (L)	2.563	2.554	2.544

1. Collect data and construct a tally sheet
2. Determine the range
3. Determine the cell interval
4. Determine the cell mid points
5. Determine the cell boundaries
6. Post the cell frequency

The number of cell or grouped in a frequency distribution

The number of cells should be between 5 and 20.

- use 5 to 9 cells when the number of observations is less than 100.
- use 8 to 17 cells when the number of observations is between 100 and 500.
- use 15 to 20 cells when the number of observations is greater than 500.

There are 45 categories, which are too many and must be reduced by grouping in to cells.

TABLE 3-3 Tally Sheet Shaft Weight (Coded from 2.500 kg).

WEIGHT	TABULATION	WEIGHT	TABULATION	WEIGHT	TABULATION
31	I	46	IIII	61	𝍸
32	II	47	III	62	II
33	I	48		63	I
34	II	49	II	64	III
35		50	IIII	65	II
36	II	51	𝍸 III	66	I
37	II	52	𝍸 I	67	II
38	III	53	III	68	II
39	I	54	II	69	III
40		55	I	70	III
41		56	𝍸	71	II
42	III	57	I	72	II
43	III	58	III	73	
44	II	59	𝍸	74	II
45	IIII	60	𝍸	75	II

Determine the range

$$R = X_H - X_L$$

Lowest number

Highest number

= 2.575-2.531

= 0.044

Determine the cell interval

The cell interval is the distance between adjacent cell midpoints

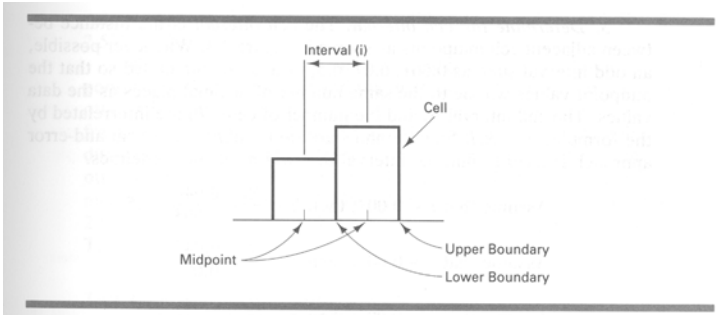


FIGURE 3-3 Cell nomenclature.

When ever possible, an odd interval such as 0.001, 0.07, 0.5, or 3 is recommended so that the mid point values will be to the same number of decimal places as the data values.

The cell interval (i) and the number of cells (h) are interrelated by the formula,

$$h = R / i$$

Assume that i=0.003; then h = 0.044/0.003
= 15

Assume that i=0.005; then h = 0.044/0.005
= 9

Assume that i=0.007; then h = 0.044/0.007
= 6

Give the best presentation

Determine the cell interval

Sturgis' rule,

$$i = \frac{R}{1 + 3.322 \log n}$$

= 0.044/(1+3.322(log110))

= 0.0057 ⇨ 0.005

Determine the cell midpoints

The lowest cell midpoint must be located to include the lowest data value in its cell. The simplest technique is to select the lowest data point as the midpoint value for the first cell. A better technique is to use the formula,

$$MP_L = X_L + \frac{i}{2}$$ **Do not round the answer**

$$MP_L = X_L + \frac{i}{2}$$

$$= 2.531 + \frac{0.005}{2}$$

$$= 2.533$$

The first cell will have data values of 2.531, 2.532, 2.533, 2.534, and 2.535.

TABLE 3-6 Frequency Distribution of Steel Shaft Weight (kilograms).

CELL BOUNDARIES	CELL MIDPOINT	FREQUENCY
2.531–2.535	2.533	6
2.536–2.540	2.538	8
2.541–2.545	2.543	12
2.546–2.550	2.548	13
2.551–2.555	2.553	20
2.556–2.560	2.558	19
2.561–2.565	2.563	13
2.566–2.570	2.568	11
2.571–2.575	2.573	8
Total		110

Determine the cell boundaries

The boundary values are an extra decimal place in accuracy than the observed values.

Some analysts prefer to leave the boundaries at the same number of decimal places as the data.

Therefore, the lower boundary for the first cell is 2.531.

Post the cell frequency

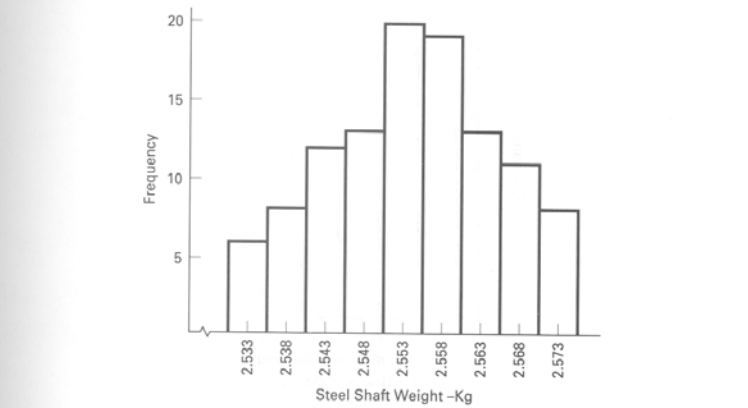


FIGURE 3-4 **Histogram of data given in Table 3-6.**

Measures of Central Tendency

Average

1. Ungrouped data

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

2. Grouped data

$$\bar{X} = \frac{\sum_{i=1}^{h} f_i X_i}{n} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_h X_h}{f_1 + f_2 + \ldots + f_h}$$

Example Problem

A technician checks the resistance value of five coils and records the values in ohms(**Ω**):X_1=3.35, X_2=3.37, X_3=3.23, X_4=3.34, and X_5=3.30. Determine the average.

=3.33 Ω

Example Problem

Given the frequency distribution of the life of 320 automotive tires in 1000 km as shown in Table 3-7, determine the average.

TABLE 3-7 Frequency Distributions of the Life of 320 Tires in 1000 km.

BOUNDARIES	MIDPOINT X_i	FREQUENCY f_i	COMPUTATION f_iX_i
23.6–26.5	25.0	4	100
26.6–29.5	28.0	36	1,008
29.6–32.5	31.0	51	1,581
32.6–35.5	34.0	63	2,142
35.6–38.5	37.0	58	2,146
38.6–41.5	40.0	52	2,080
41.6–44.5	43.0	34	1,462
44.6–47.5	46.0	16	736
47.6–50.5	49.0	6	294
Total		$n = 320$	$\Sigma f_iX_i = 11,549$

3. Weight average

When a number of average are combined with different frequency.

$$\bar{X}_w = \frac{\sum_{i=1}^{n} w_i \bar{X}_i}{\sum_{i=1}^{n} w_i}$$

Example Problem

Tensile tests on aluminum alloy rods are conducted at three different times, which results in three different average values in megapascals (Mpa). On the first occasional five tests are conducted with a average of 207 Mpa; on the second occasion six tests, with a average of 203 Mpa; and on the last occasion three test, with a average of 206 Mpa. Determine the weight average.

Median

is the value which divides a series of ordered observation so that the number of items above it is equal to the number below it.

- Ungrouped technique
  - **n=odd Median is the midpoint of the values.**
  - **n=even Median is the average of the two middle numbers.**
- Grouped technique

$$Md = L_m + \left(\frac{\frac{n}{2} - cf_m}{f_m} \right) i$$

Md	Median
L_m	Lower boundary of the cell with the median
n	Total number of observations
cf_m	Cumulative frequency of all cells below
f_m	Frequency of median cell
i	Cell interval

Example Problem

TABLE 3-7 Frequency Distributions of the Life of 320 Tires in 1000 km.

BOUNDARIES	MIDPOINT X_i	FREQUENCY f_i	COMPUTATION $f_i X_i$
23.6–26.5	25.0	4	100
26.6–29.5	28.0	36	1,008
29.6–32.5	31.0	51	1,581
32.6–35.5	34.0	63	2,142
35.6–38.5	37.0	58	2,146
38.6–41.5	40.0	52	2,080
41.6–44.5	43.0	34	1,462
44.6–47.5	46.0	16	736
47.6–50.5	49.0	6	294
Total		$n = 320$	$\Sigma f_i X_i = 11{,}549$

Mode

is the set of numbers is that value that occurs with the greatest frequency. It is possible for the mode to be nonexistent in a series of numbers or to have more than one value.

Example

3, 3, 4, 5, 5, 5, and 7 has a mode of 5

22, 23, 25, 30, 32, and 36 does not has a mode

105, 105, 105, 107, 108, 109, 109, 109, 110, and 112

has two modes, 105 and 109.

Relationship Among the Measures of Central Tendency

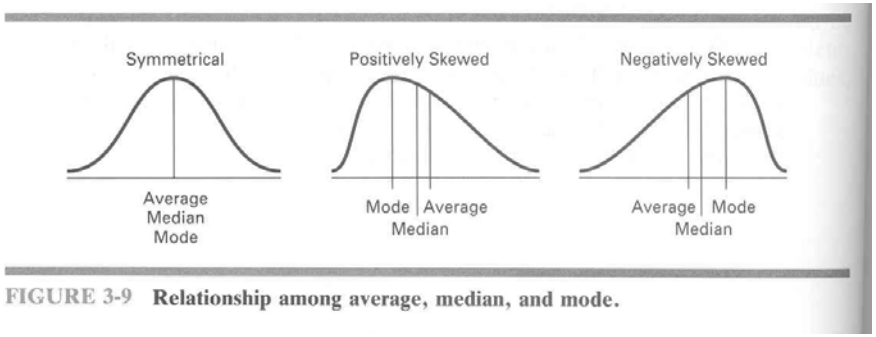


FIGURE 3-9 Relationship among average, median, and mode.

Measure of Dispersion

describe how the data are spread out or scattered on each side of the central value.

- Range
- Standard Deviation

Range

Is the difference between the largest and smallest values or observations.

$$R = X_H - X_L$$

X_L: **Lowest observation in a series**

X_H: **Highest observation in a series**

Example Problem

If the highest weekly wage in the assembly department is $280.79 and the lowest weekly wage is $173.54, determine the range.

$107.25

Standard Deviation

is a numerical value in the units of the observed values that measures the spreading tendency of the data. A large standard deviation shows greater variability of the data than does a small standard deviation.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$$

Example Problem

s=0.13 kg

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
3.2	+0.2	0.04
2.9	-0.1	0.01
3.0	0.0	0.00
2.9	-0.1	0.01
3.1	+0.1	0.01
2.9	-0.1	0.1
$\bar{X}$ =3.0	$\sum$ =0	$\sum$ =0.08

1. Ungrouped technique

$$s=\sqrt{\frac{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}{n(n-1)}}$$

Determine the standard deviation of the moisture content of a roll of Kraft paper. The results of six reading across the paper web are 6.7, 6.0, 6.4, 6.4, 5.9, and 5.8%.

=0.35%

2. Grouped technique

$$s=\sqrt{\frac{n\sum_{i=1}^{h}\left(f_iX_i^2\right)-\left(\sum_{i=1}^{h}f_iX_i\right)^2}{n(n-1)}}$$

Do not round $\sum_{i=1}^{h} f_iX_i$ or $f_iX_i^2$, as this action will affect accuracy.

Example Problem

TABLE 3-9 Passenger Car Speeds (in km/h) During a 15-Minute Interval on I-57 at Location 236.

BOUNDARIES	MIDPOINT X_i	FREQUENCY f_i	COMPUTATIONS f_iX_i	$f_iX_i^2$
72.6–81.5	77.0	5	385	29,645
81.6–90.5	86.0	19	1634	140,524
90.6–99.5	95.0	31	2945	279,775
99.6–108.5	104.0	27	2808	292,032
108.6–117.5	113.0	14	1582	178,766
Total		n = 96	ΣfX = 9354	ΣfX^2 = 920,742

Given the frequency distribution of Table 3-9 for passenger car speeds during a 15 minute interval on I-57, determine the average and standard deviation.

Relationship Between the Measure of Dispersion

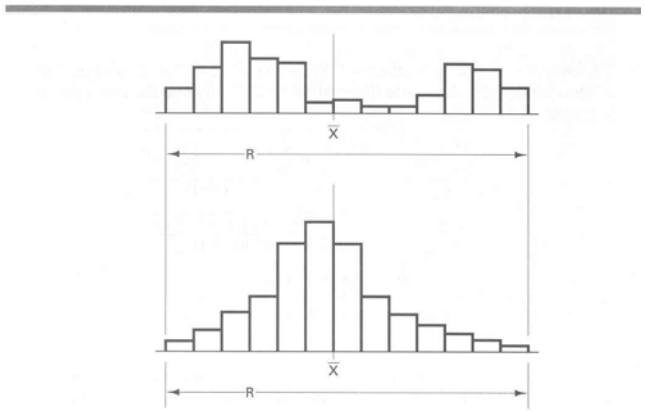


FIGURE 3-10 Comparison of two distributions with equal average and range.

Other Measures

1. Skewness is a lack of symmetry of the data. The formula is given by

$$a_3 = \frac{\sum_{i=1}^{h} f_i \left(X_i - \bar{X} \right)^3 / n}{s^3}$$

0 the data are symmetrical

\+ the data are skewed to the right

\- the data are skewed to the left

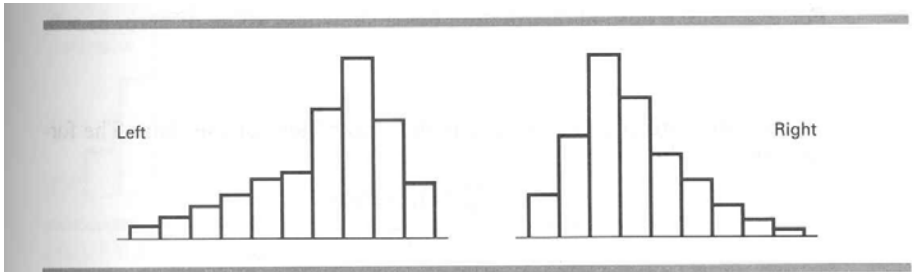


FIGURE 3-11 Left (negative) and right (positive) skewness distributions.

Example Problem

TABLE 3-10 Data for Skewness and Kurtosis Example Problems.

X_i	f_i	$X_i - \bar{X}$	$f_i(X_i - \bar{X})^3$	$f_i(X_i - \bar{X})^4$
1	1	$(1 - 7) = -6$	$1(-6)^3 = -216$	$1(-6)^4 = 1296$
4	6	$(4 - 7) = -3$	$6(-3)^3 = -162$	$6(-3)^4 = 486$
7	16	$(7 - 7) = 0$	$16(0)^3 = 0$	$16(0)^4 = 0$
10	8	$(10 - 7) = +3$	$8(+3)^3 = +216$	$8(+3)^4 = 648$
	$\Sigma = 31$		$\Sigma = -162$	$\Sigma = 2430$

Determine the skewness of the frequency distribution of Table 3-10. The average and sample standard deviation are calculated and are 7.0 and 2.32, respectively.

1. Kurtosis is the peakedness of the data. The formula is given by

$$a_4 = \frac{\sum_{i=1}^{h} f_i \left(X_i - \bar{X} \right)^4 / n}{s^4}$$

Example Problem

TABLE 3-10 Data for Skewness and Kurtosis Example Problems.

X_i	f_i	$X_i - \bar{X}$	$f_i(X_i - \bar{X})^3$	$f_i(X_i - \bar{X})^4$
1	1	$(1 - 7) = -6$	$1(-6)^3 = -216$	$1(-6)^4 = 1296$
4	6	$(4 - 7) = -3$	$6(-3)^3 = -162$	$6(-3)^4 = 486$
7	16	$(7 - 7) = 0$	$16(0)^3 = 0$	$16(0)^4 = 0$
10	8	$(10 - 7) = +3$	$8(+3)^3 = +216$	$8(+3)^4 = 648$
	$\Sigma = 31$		$\Sigma = -162$	$\Sigma = 2430$

Determine the kurtosis of the frequency distribution of Table 3-10, which has $\bar{X}$ = 7.0 and S = 2.32.

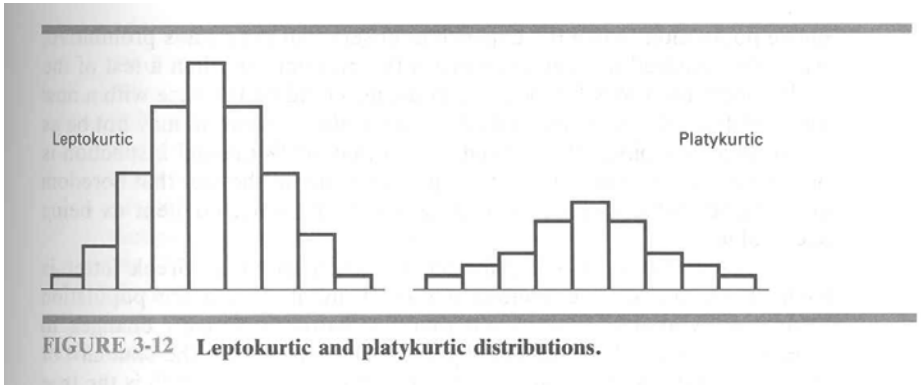


FIGURE 3-12 Leptokurtic and platykurtic distributions.

question

1. **Determine the median of the following number**
   22, 11, 15, 8, 18
   35, 28, 33, 38, 43, 36
2. **The average height of 24 students in sec1 of qc course is 1.75m; the average height of 18 student in sec2 is 1.79 m; and the average height of 29 students in sec3 is 1.68 m. What is the average height of the student in three sections of quality control.**

Concept of a Population and a Sample

TABLE 3-11 Comparison of Sample and Population.

SAMPLE	POPULATION
Statistic	Parameter
$\bar{X}$—average	$\mu(\bar{X}_0)$—mean
s—sample standard deviation	$\sigma(s_0)$—standard deviation

The Normal Curve

Normal curve or Gaussian distribution is one type of population that is quite common. The area under the curve is equal to 1.0000 or 100%.

Standard normal value $Z = \dfrac{X_i - \mu}{\sigma}$

99.73% of item are include between ±3σ

Problem

5. A company that fills bottles of shampoo tries to maintain a specific weight of the product. The table gives the weight of 110 bottles that were checked at random intervals. Make a tally of these weights and construct a frequency histogram. (Weight is in kilograms.)

6.00	5.98	6.01	6.01	5.97	5.99	5.98	6.01	5.99	5.98	5.96
5.98	5.99	5.99	6.03	5.99	6.01	5.98	5.99	5.97	6.01	5.98
5.97	6.01	6.00	5.96	6.00	5.97	5.95	5.99	5.99	6.01	6.00
6.01	6.03	6.01	5.99	5.99	6.02	6.00	5.98	6.01	5.98	5.99
6.00	5.98	6.05	6.00	6.00	5.98	5.99	6.00	5.97	6.00	6.00
6.00	5.98	6.00	5.94	5.99	6.02	6.00	5.98	6.02	6.01	6.00
5.97	6.01	6.04	6.02	6.01	5.97	5.99	6.02	5.99	6.02	5.99
6.02	5.99	6.01	5.98	5.99	6.00	6.02	5.99	6.02	5.95	6.02
5.96	5.99	6.00	6.00	6.01	5.99	5.96	6.01	6.00	6.01	5.98
6.00	5.99	5.98	5.99	6.03	5.99	6.02	5.98	6.02	6.02	5.97