1-

Variables and frequency distribution

1.1 Basic definitions

1.1.1 Variables

In engineering systems, whether we deal with raw materials, equipment, utilities or products, a lot of variables are encountered, such as mass, level of purity, flow rate, temperature, pressure, etc. These variables can be classified into two types:

Discrete variables: These can take only certain specific numerical values. For example, the maximum ambient temperature, as recorded in ^{o}C in five consecutive days whose values belong to a finite set: {25, 26, 26, 24, 27}.

Continuous variables: Here, the variable does not take specific values but can vary within a real interval taking all possible values within this interval. For example, For example, the mole fraction of a volatile component in a distillate fraction will belong to the real interval (0.6, 0.64).

1.1.2 Population and samples

Consider the total production per day of a factory producing sacks of fertilizers. The daily produced sacks represent a **population**. Now, if we choose 50 of these sacks to test them for their weight (For example), this set is called a **sample**. Each of the 50 sacks of this sample is called a **specimen**.

A characteristic variable of a population is called a **parameter** while that of a sample is called a **statistic**.

In the following sections we will be dealing with samples.

1.2 Frequency distribution

Consider the following data, describing the daily productivity of an oil well (bbl) over 30 days:

800	850	790	940	1000	740	820	940	960	940
970	920	850	870	800	760	1030	1010	980	890
960	900	920	990	850	930	860	840	800	780

Given this way, these data are termed **ungrouped**. If, however, a table can be drawn, that shows the different classes of concentration figures in 50 bbl intervals and the frequency of their occurrence, then this table represents **grouped data**. This is shown in Table (1.1) where the mid value of each class is shown in the third raw.

Table (1.1): Class intervals of daily production

Class	**700-750**	**750-800**	**800-850**	**850-900**	**900-950**	**950-1000**	**1000-1050**
Frequency	1	3	5	6	7	5	3
Mid-class	725	775	825	875	925	975	1025

These data can be represented graphically in the form of **histogram** (Figure 1.1).

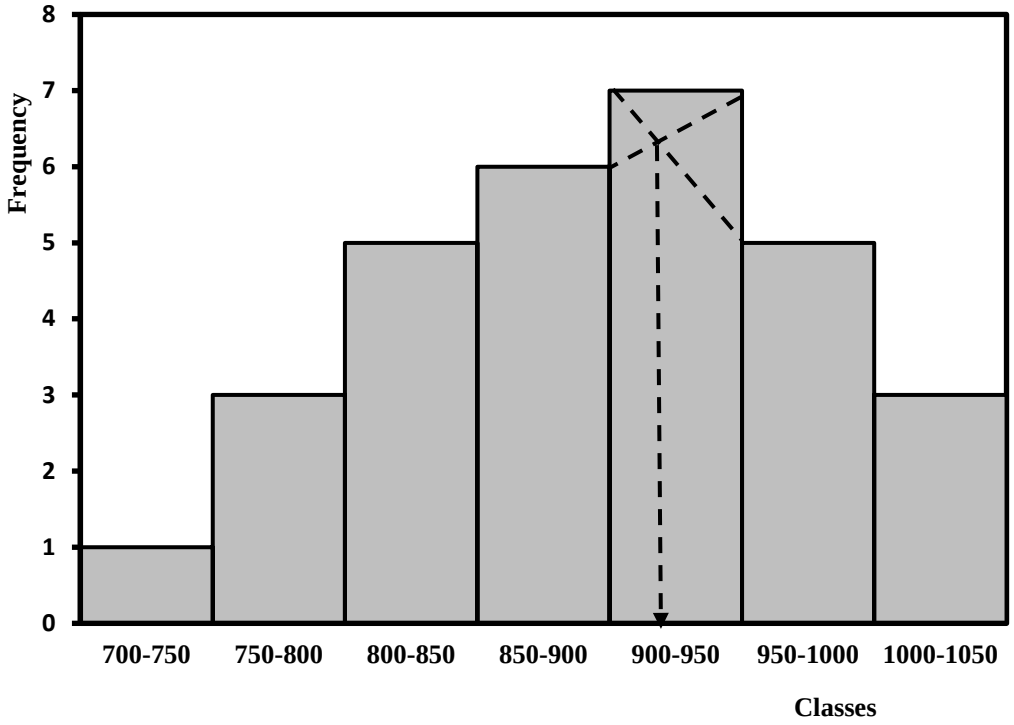

Fig (1.1): Histogram for S.M. Concentration

Plotting the frequency against mid-class values produces a broken line frequency plot, shown in Figure (1.2). It is sometimes referred to as **Frequency polygon**.

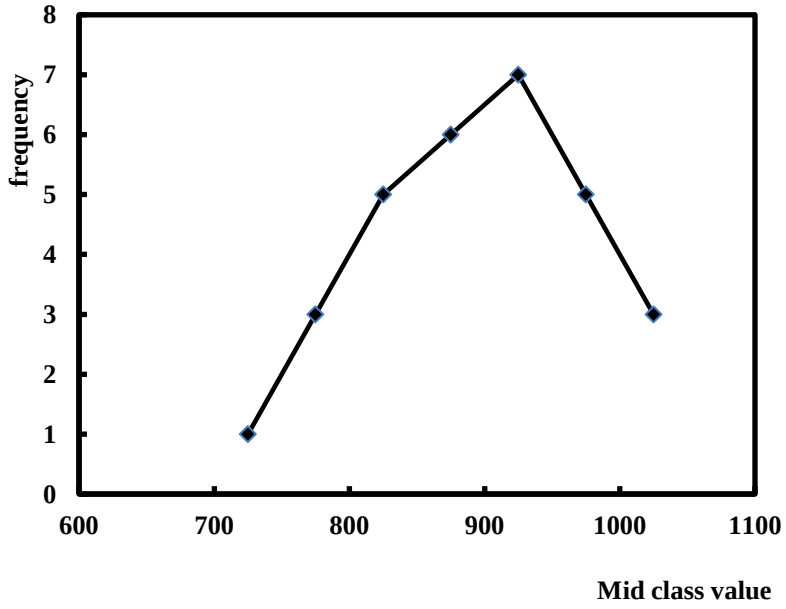

Fig (1.2): Frequency polygon

The above plot shows only one maximum value, which represents the most encountered pollutant concentration. This value is usually known as the **mode** of distribution. This curve is thus known as a **monomoda**l curve. Bimodal curves will have two maximum values, whereas skewed curves will have their mode shifted towards one end of the class values interval.

1.3 Statistical averages

Often, we need to quote a single value that represents some typical average of the statistical distribution. There are several such values, commonly known as central tendencies or averages.

1.3.1 The mean value (Arithmetic average)

The mean (or average) is the most popular and well known measure of central tendency. For a sample of **ungrouped** values such as x_1, x_2, …, x_n, the mean value is simply defined as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

A similar definition for a population of *N* ungrouped values can be written:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1.2}$$

On using EXCEL, it is possible to obtain the mean value by simply highlighting the numbers and use the function **AVERAGE**.

One obtains the value $\bar{x}$ **= 889.7**

For a sample of **grouped** values, the definition becomes:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i \bar{x}_i}{n} \tag{1.3}$$

Whereas for a population, the corresponding expression is:

$$\mu = \frac{\sum_{i=1}^{N} f_i \bar{x}_i}{N} \tag{1.4}$$

In both cases, $\bar{x}_i$ represents the mid-class value. The following table shows the calculations using grouped data from Table (1.1):

Table (1.2): Calculation of mean value

$\bar{x}_i$	725	775	825	875	925	975	1025	**Sum**
f_i	1	3	5	6	7	5	3	**30**
$f_i\bar{x}_i$	725	2325	4125	5250	6475	4875	3075	**26850**

$\bar{x} = \frac{26850}{30} =$ **895**

This value is slightly different from the exact value obtained from raw data (889.7)

1.3.2 The median of a distribution

In an array of **ungrouped data** arranged by ascending order of magnitude, the median is the middle value. For example, the following data show the maximum temperature recorded over one week period in Cairo (in ^{o}C):
26, 28, 29, 27, 25, 24, 24.

The data are then grouped in ascending order.
24, 24, 25, **26**, 27, 28, 29.
As these data are grouped in an ascending order, the median = 26.

If the number of data points is even, then the median value will consist of the arithmetic average of the two middle values. In any case, the EXCEL function **MEDIAN** directly discloses the median value of any set of ungrouped data. For the data in Table (1.1), the median value is ***M* = 895**.

For **grouped data**, a cumulative plot is made, and the median is the value corresponding to one half the sum of all frequencies.

For example, in the preceding example, Table (1.2) can be re-written in terms of "number of days where the production is less than…" as shown in Table (1.3):

Table (1.3): Cumulative distribution of data in Table (1.1)

End value of class	750	800	850	900	950	1000	1050
No of days with production less than	1	4	9	15	22	27	30

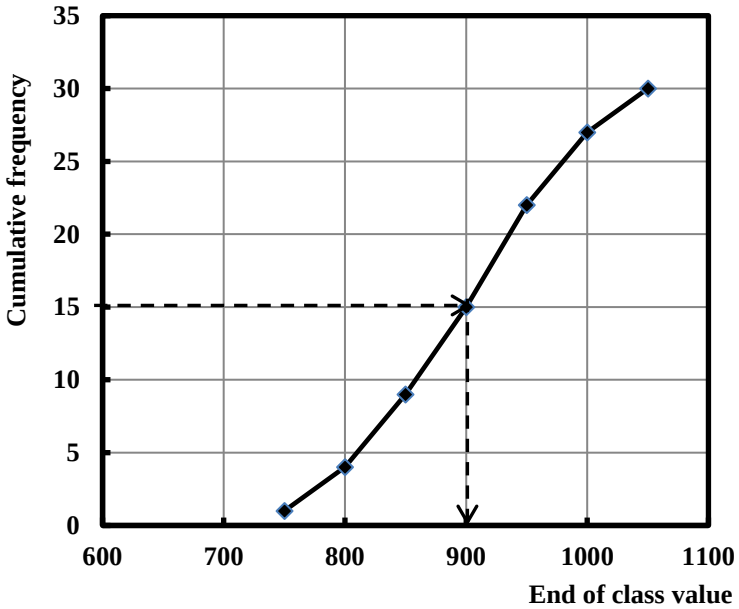

Fig (1.3): Cumulative frequency curve

A cumulative plot represents the cumulative frequency of values less than the entry in the table. This cumulative plot is called an **ogive**. The plot is performed for cumulative frequency against end of class values.

This cumulative plot is shown in Figure (1.3) and the median value is obtained at a frequency = $0.5 \times 30 = 15$. Its value is about **900**.

1.3.3 The mode of a distribution

In its simplest form, the mode can be defined as the value which occurs at the highest frequency among the data. In Table (1.2), this value is **940.** However, this value is only approximate, since it corresponds to the mid – value of the interval.

To get a more accurate value, the following interpolation can be made within the class corresponding to the maximum frequency (modal class):

$$Mo = L_1 + \frac{D_1 \times i}{D_1 + D_2} \qquad (1.5)$$

Where:

L_1 represents the lower value of modal class
D_1 is the difference in the frequency of modal class and the previous one
D_2 is the difference in the frequency of modal class and the next one
i is the size of class interval.

For data of Table (1.1), the modal class is 900 –950, hence L_1 = 900
D_1 = 7 – 6 = 1 and D_2 = 7 – 5 = 2 and i = 50

Substituting in equation (1.6), one gets:

Mo ≈ 916.7

A simpler graphical method relying on the above equation can be applied to get the value of the mode with reasonable accuracy. This can be followed in Figure (1.1) where the interpolation is made graphically to give: Mode ≈ 920.

1.3.4 Level of measurements of variables

In some cases we are confronted with non – numerical data. In this respect the following represents the classical **level of measurement** of data that are treated statistically:

These are categorized as **nominal**, **ordinal, interval, or ratio** variables.

Nominal variables are those variables that have no specific numerical value. They may be numbered only for convenience. For example, chemicals in a warehouse are categorized as: 1. Inorganic solids, 2. inorganic solutions, 3. Organic solids, 4. Organic liquids, etc. Note that the numbering system is extremely arbitrary and the value of the number holds no indication about its magnitude. Nominal variables do not have a median or mean value. They may however have a mode. For example, if we consider a class list the most frequent surname will be the mode.

Ordinal variables on the other hand, possess numerical values that can only help in ordering them in an ascending or descending way. For example, the Moh's scale of hardness sorts minerals according to their hardness in a 1 to 10 scale. Talc, which is assigned the ordinal 1 is the mineral of least hardness while quartz is much harder (7). On this scale, diamond is the hardest mineral with an ordinal of 10. Such numbers bear only relative numerical significance since quartz is not 7 times as hard as talc. For such variables a median and a mode can be defined but not a mean value.

Interval variables: Quantitative attributes are all measurable on interval scales, as any difference between the levels of an attribute can be multiplied by any real number to exceed or equal another difference. A highly familiar example of interval scale measurement is temperature with the Celsius scale. In this scale, the unit of measurement is 1/100 of the temperature difference between the freezing and boiling points of water under a pressure of 1 atmosphere. The "zero point" on an interval scale is arbitrary; and negative values can be used. Variables measured at the interval level are called "interval variables" or sometimes "scaled variables" as they have units of measurement.

Ratios between numbers on the scale are not meaningful, so operations such as multiplication and division cannot be carried out directly. But ratios of differences can be expressed.

The central tendency of a variable measured at the interval level can be represented by its mode, its median, or its arithmetic mean. Statistical dispersion can be measured in most of the usual ways, such as range and standard deviation. Since one cannot divide, one cannot define measures that require a ratio, such as coefficient of variation.

Ratio variables: This represents the highest level of measurement and includes common variables that can be subjected to all numerical operations.

1.4 Measures of dispersion

1.4.1 The concept of dispersion

The following data represents the daily consumption of electrical energy of a small factory over a one-week period (in kWh): 310, 330, 400, 290, 320, 290, 300. Their mean value = 320.

On the other hand, another factory yields the following figures: 310, 330, 300, 360, 290, 350, 300. The mean value = 320.

Although both distributions have the same mean value, we can see that the difference between the highest and the lowest values in the first is 110, while it is 70 in the second. This suggests that the first distribution is more dispersed than the second.

The difference between the highest and lowest values is known as **the range** and represents the crudest measure of dispersion.

This measure can be misleading: For, if we discard the number 400 from the first set of data, we get a range of 40.

Another commonly used measure of dispersion is called the interquartile range (IQR). This will not, however, be discussed in this work.

The most reliable measure of dispersion used in statistical analyses is the **standard deviation**.

1.4.2 The Standard Deviation

(a) For **ungrouped data**, the standard deviation of **sample** data is defined by:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.6}$$

For a **population** the corresponding expression is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{1.7}$$

Equation (1.7) can be simplified as follows:

$$\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{i=1}^{N}x_i^2 - 2\mu.x_i + \mu^2 = \sum_{i=1}^{N}x_i^2 - 2\mu.\sum_{i=1}^{N}x_i + N.\mu^2$$

Since: $\sum_{i=1}^{N}x_i = N.\mu$, the previous equation simplifies to:

$$\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{i=1}^{N}x_i^2 - N.\mu^2$$

Hence equation (1.7) simplifies to:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}x_i^2}{N} - \mu^2} \tag{1.8}$$

For a sample of size *n* the equation is slightly different:

$$s = \sqrt{\frac{\sum_{i=1}^{n}x_i^2 - n.\bar{x}^2}{n-1}} \tag{1.9}$$

On using EXCEL, the function **STDEV.S** is applied for a sample and **STDEV.P** for a population. For the sample data in Table (1.1), one obtains: ***s* = 81.1**

(b) For **grouped data**, the standard deviation of **sample** data is defined by:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \bar{x}_i^{\,2} . f_i - n.\bar{x}^{\,2}}{n-1}} \qquad (1.10)$$

For a **population** the corresponding expression is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \bar{x}_i^{\,2} . f_i}{N} - \mu^2} \qquad (1.11)$$

Calculations of the standard deviation using the grouped data in Table (1.2) are displayed in Table (1.4)

Table (1.4): Calculation of standard deviation from grouped data

$\bar{x}_i$	725	775	825	875	925	975	1025	Sum
f_i	1	3	5	6	7	5	3	30
$f_i\bar{x}_i$	725	2325	4125	5250	6475	4875	3075	26850
$f_i\bar{x}_i^2$	525625	1801875	3403125	4593750	5989375	4753125	3151875	24218750

Hence $s = \sqrt{\frac{24218750 - 30 \times 895^2}{30-1}} =$ **80.5**

This value is close to the exact value of 81.1 obtained from raw data.

It is to be finally noted that the units of standard deviation are identical to those of the variable under consideration.

1.4.3 The Variance

The variance can be thought of as the square of the standard deviation, implying its being a measure of dispersion as well.

For a population and a sample of ungrouped data respectively:

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2 \qquad (1.12)$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \bar{x}^{\,2} \qquad (1.13)$$

These can be directly obtained by using the EXCEL function **VAR.P** or **VAR.S** respectively.

The corresponding formulae for grouped data are:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \bar{x}_i^{2} . f_i}{N} - \mu^2 \qquad (1.14)$$

And

$$s^2 = \frac{\sum_{i=1}^{n} \bar{x}_i^{2} . f_i - n.\bar{x}^{2}}{n-1} \qquad (1.15)$$

1.4.4 The coefficient of variation (CV)

When comparing dispersions in two distributions using their standard deviations as measure, the variables must have the same units and their mean values should be equal. Since this is not usually the case, the comparison is rather based on the **coefficient of variation**, defined for populations and samples respectively as:

$$\text{CV} = \frac{\sigma}{\mu} \times 100\% \qquad (1.16)$$

$$\text{CV} = \frac{s}{\bar{x}} \times 100\% \qquad (1.17)$$

For example, for the ungrouped data of Table (1.1), since $\bar{x}$ = 889.7 and s = 81.1, the coefficient of variation is:

$$\text{CV} = \frac{81.1}{889.7} \times 100\% = \mathbf{9.11\%}$$

The coefficient of variation allows investors to determine how much risk is assumed in comparison to the amount of expected return. The lower the coefficient of variation, the better is the risk-return trade-off.

1.5 Skewness of a distribution

An ideal distribution where the mean, medium and mode are equal, is a symmetrical distribution. Figure (1.4)

On the other hand, some distributions are not symmetrical as can be seen from the same figure. These are called **skewed distributions**. This is calculated from the following formulas ofr ungrouped data:

$$S_k = \frac{\sum_{i=1}^{N}(x_i - \mu)^3}{N.\sigma^3} \text{ (For a population)} \qquad (1.18)$$

As for samples:

$$S_k = \frac{\sum_{i=1}^{n} n.(x_i - \bar{x})^3}{(n-1)(n-2).s^3} \text{ (For a sample)} \qquad (1.19)$$

For ungrouped data, skewness is readily obtained from the EXCEL function: **SKEW.P** or **SKEW** that uses equations (1.18) or (1.19) respectively. For the data in Table (1.1), this function yields a low skewness of $\mathbf{-0.094}$. This is emphasized by the high symmetry observed in Figure (1.2).

For grouped data, similar formulas can be applied:

$$S_k = \frac{\sum_{i=1}^{N} f_i(x_i - \mu)^3}{N.\sigma^3} \quad \text{(For a population)} \qquad (1.20)$$

$$S_k = \frac{\sum_{i=1}^{n} nf_i.(x_i - \bar{x})^3}{(n-1)(n-2).s^3} \quad \text{(For a sample)} \qquad (1.21)$$

A positive value of skewness means that the distribution possesses a "tail" to the right while the tail is at the left for negative values of skewness.

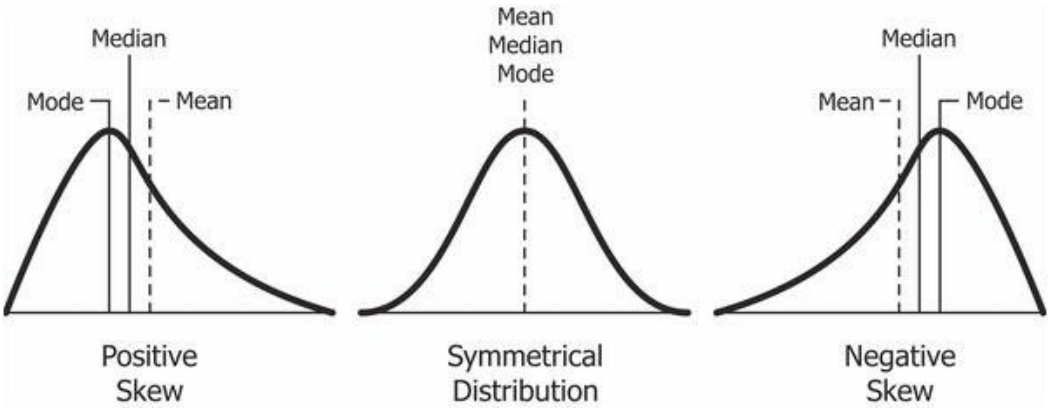

Fig (1.4): Symmetrical and skewed distributions

1.6 Kurtosis

Kurtosis defines the flatness of a distribution rather than its being symmetrical or not. It is defined by a formula using ungrouped data. It shows how far the distribution is "tailed". A positive value of kurtosis indicates an elongated distribution (Leptokurtic) while a negative value will denote a flat distribution (Platykurtic) (Figure 1.5).

The value of kurtosis can be obtained for ungrouped data from the EXCEL function KURT. For the data of Table (1.1), $= \mathbf{-1.094}$, which is considered to denote a moderate platycurtic distribution (Figure 1.5).

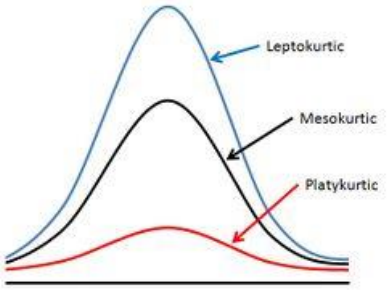

Fig (1.5): Kurtosis as related to curve shapes

1.7 Statistical moments

Equation (1.4) can be rewritten in the form:

$\sum_{i=1}^{N} \mu . f_i = \sum_{i=1}^{N} x_i f_i$

Hence: $\sum_{i=1}^{N} (x_i - \mu) f_i = 0$

The expression $\frac{\sum_{i=1}^{N} (\bar{x}_i - \mu) f_i}{N}$ is called the **First Moment about the mean**

It is obvious that its value is zero.

The **second moment** is similarly defined by:

$$\frac{\sum_{i=1}^{N} (\bar{x}_i - \mu)^2 f_i}{N}$$

This is the definition of **variance** of a population which is usually simplified to the form previously obtained (Equation 1.15)

Similarly **Skewness,** as defined by equation (1.18) refers to the **third moment**. And although the defining expression was not stated, the **fourth moment** refers to **kurtosis.**

1.8 Exercise problems

(1) The temperature inside an isothermal reactor was measured every hour over a 9 hour period. Calculate the mean value and the standard deviation from these data**:** 67, 68, 70, 69, 67, 66, 64, 65, 67.

(2) The following data were collected over a two weeks period as a sample for COD of waste water ($mg.L^{-1}$) of a food processing plant.

1550 2070 1800 2020 1560 2700 2530 2100 1890 2050 2450 1720 1050 1400

Estimate the mean value of COD, its standard deviation, median value, the skewness of the distribution and its kurtosis.

(3) The following table represents the ages of a random sample of 130 persons:

Age	[0 –10)	[10 – 20)	[20 – 30)	[30 – 40)	[40 – 50)	[50 – 60)	[60 – 70)	[70 – 80)
Frequency	2	18	26	37	22	14	9	2

Calculate:

a- The mean age of sample
b- The standard deviation of the sample
c- The median value
d- Skewness
e- The modal value

(4) The following table shows the incidence of stoppages occurring in a production line along a 320 days year

Nº of stoppages	0	1	2	3	4	5	6	7
Days	84	88	64	38	25	11	8	2

Calculate the average number of stoppages and the variance.

(5) The following data represent the scores of a class of 40 students in a test. (Out of 20)

16	14	15	9	3	16	11	19	5	13
20	5	6	17	13	14	14	17	11	10
4	6	18	20	16	17	11	8	9	1
20	14	19	18	12	19	7	13	7	6

Obtain the average score and the standard deviation out of these raw data. Also determine the median and skewness of the distribution.

(6) The following table summarizes the chemical analyses related to the purity of samples of an ore obtained from two different quarries A and B. Compare the mean values of ore purity and their COV. In your opinion which quarry is more reliable?

Sample number	1	2	3	4	5	6	7	8	9	10
% Purity (A)	39	44	52	37	40	45	32	55	48	43
% Purity (B)	40	42	46	39	37	41	45	44	44	39