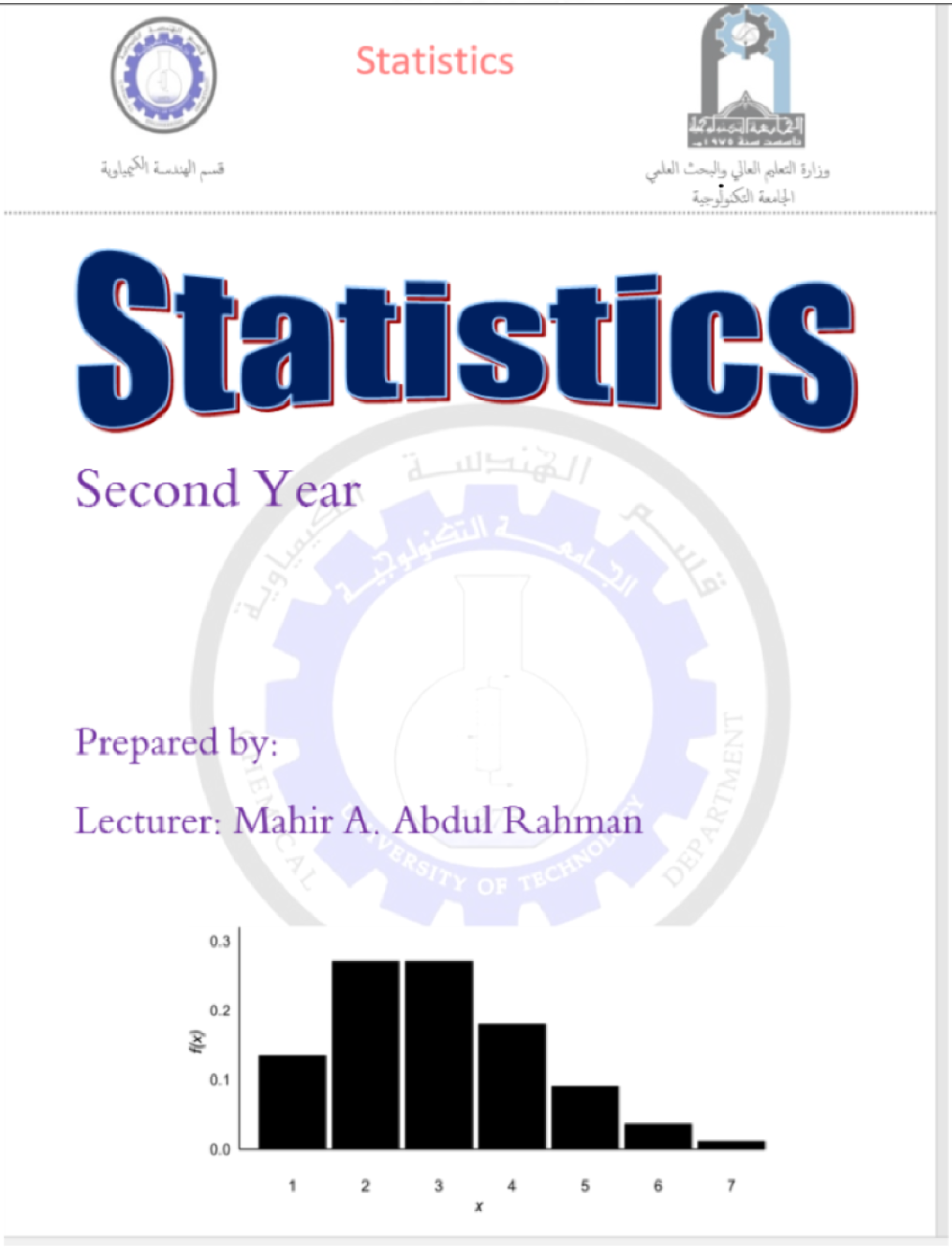
Statistics



قسم الهندسة الكيمياوية

Statistics

وزارة التعليم العالي والبحث العلمي
الجامعة التكنولوجية

Statistics

Second Year

Prepared by:

Lecturer: Mahir A. Abdul Rahman

Save from: http://www.uotechnology.edu.iq/dep-chem-eng/index.htm

Statistics

Chapter (1) Introduction

- **Statistics :** Is concerned with scientific methods for collecting, organizing, summarizing, presenting and analysis data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

- **Population :** Set of all possible measurements.

- **Finite :** All bots produced in a factory, catalyst pellets.

- **Infinite :** All possible outcomes (heads, tails) in successive tosses of a coin.

- **Sample :** A set of measurements taken to represent an infinite or large finite population, which is selected randomly.

- **Random sample :** Is selected so that all elements of the pop. have an equal chance of being measured.

- **Sample array :** Is the set of measurements of sample elements.

- **Inductive or statistical inference :** If a sample is representative of a pop., important conclusions about the pop. can often be inferred from the analysis of the sample. The phase of statistics dealing with the conditions under which such inference is valid is called inductive statistics.

- Deductive or descriptive statistics : The phase of statistics which seeks only to describe and analysis a given group without drawing any conclusion or inferences about a large group.

- Variable : Is a symbol, such as X , Y, which can assume any of a prescribed set of values, called the **domain** of the variables.

- If the variable can assume only one value is called a **constant**.

- A variable which can theoretically assume any value between two given values is called a **continuous variable**, otherwise it is called a **discrete variable**.

- The no. N of children in a family, which can assume any of the values 0,1,2,3,... but cannot be 2.5 or 3.842, is a **discrete variable**.

- The age A of an individual, which can be 62 years, 63.8, ... depending on accuracy of measurements is a **continuous variable.**

- **Size of data** : Number of measurements.

- **Range** : Highest – Lowest measurements.

Chapter (2)

Frequency Distribution

- When elements of population are unequal in a certain parameter and/or measurement error is involved a statistical estimation is needed. This involves:

1. Data sampling for repeated measurements.
2. Classification of data (frequency dist.).
3. Presentation of classified data.
4. Estimation of statistical parameters.
5. Analysis of statistical parameters and hypotheses.

- Frequency distribution into classes:

The sample range is sub-divided in to a number of classes. Usually:

For size $\nearrow$ > 50 10-20 classes
$\searrow$ ≤ 50 5-10 classes are used

Example :

The life of electric bulbs in hours was sampled :

690	701	722	684	680
728	705	693	691	688
740	663	676	738	714
698	687	703	726	699
694	705	717	682	717
712	733	705	673	694
679	680	664	691	669
689	702	710	696	697
685	724	726	698	688
702	696	708	696	710

- sample size = 50 measurements.

- sample range = 740 – 663 = 77 hr.

- class limits = highest and lowest measurements in the class.

- class interval = upper limit – lower limit.

- class boundaries = limits $\mp \frac{1}{2}$ unit in LSD.

- class width = upper – lower boundaries.

- class mark = mid point of class.

- e. g. if the class limits are 670 → 678.

* class interval = 678 – 670 = $\underline{8}$.

* class boundaries are : lower bound. = 670 – 0.5 = $\underline{669.5}$.

upper bound. = 678 + 0.5 = $\underline{678.5}$.

* Class width = 678.5 – 669.5 = $\underline{9}$

* class mark $= \frac{670+678}{2} = \underline{674}$

Or :

Class mark $= \frac{669.5+678.5}{2} = \underline{674}$

e. g. if class limits are 5.87 → 6.32 :

* class interval $= 6.32 - 5.87 = \underline{0.45}$

* class boundaries are : lower bound. $= 5.87 - 0.005 = \underline{5.865}$.

upper bound. $= 6.32 + 0.005 = \underline{6.325}$.

* Class width $= 6.325 - 5.865 = \underline{0.46}$

* class mark $= \frac{5.87+6.32}{2} = \underline{6.095}$

Or :

Class mark $= \frac{5.865+6.325}{2} = \underline{6.095}$

Determination of classes :

1. Determine the range.
2. Determine the total width

   Total width = range + one unit in LSD.
3. Divided the total width into a convenient no. of classes

   Class width $= \frac{total\ width}{no.of\ classes}$

Note :

(Adjust the total width by adding one or two units in LSD if necessary, to select a suitable no. of classes, so that the class width is of a similar accuracy to the measurements).

4. Determine class interval :

Class interval = class width – one unit in LSD

5. Starting at lowest measurements, calculate the limits of successive classes.

Solution of example (electric bulb sample) :

1. range = 740 – 663 = $\underline{77}$ hr

2. one unit in LSD = $\underline{1}$

$\therefore$ total width = 77 +1 = $\underline{78}$

3. select no. of classes (for example take 5 classes) so the class width = $\frac{78}{5} = \underline{15.6}$

Which is not the same accuracy as the data.

So take 6 classes :

class width = $\frac{78}{6} = \underline{13}$

Which is the same accuracy as the data.

4. class interval = 13 – 1 = 12

Freq. dist. Table

			x_i	f_i
	Class limit	Class bound.	Class mark	Freq.
1.	663-675	662.5-675.5	669	4
2.	676-688	675.5-688.5	682	10
3.	689-701	688.5-701.5	695	15
4.	702-714	701.5-714.5	708	11
5.	715-727	714.5-727.5	721	6
6.	728-740	727.5-740.5	734	4
				N = 50

Types of Frequency :

1. Numeric frequency : $f \rightarrow \sum f_i = N$

2. Relative frequency : $f_r = \frac{f}{N} \rightarrow \sum f_r = 1$

3. Percent frequency : $f_p = f_r * 100 \rightarrow \sum f_p = 100$

4. Cumulative frequency : The freq. is also expressed cumulatively of : f , f_r , f_p.

- Cumulative freq. of class K is the sum of frequencies of all classes up to K.

$$fc_K = \sum_{i=1}^{K} f_i \quad , \quad fc_{rK} = \sum_{i=1}^{K} f_{ri} = 1$$

$$fc_{pK} = \sum_{i=1}^{K} f_{pi} = 100$$

f	f_r	f_p
4	0.08	8
10	0.2	20
15	0.3	30
11	0.22	22
6	0.12	12
4	0.08	8
$\sum f = 50$	$\sum f_r = 1$	$\sum f_p = 100$

$$* fc = \sum_{i=1}^{i=6} f_i = 50$$

$$* fc_r = \sum_{i=1}^{6} f_{ri} = 1$$

$$* fc_p = \sum_{i=1}^{6} f_{pi} = 100$$

<u>Graphical presentation of freq. dist. :</u>

- Classified data may be presented as graphical plot with freq. as vertical axis Versus measurement as horizontal axis :

1) Histogram : Is a bar chart, in which each class is represent by a rectangle, whose base extends between the class boundaries and the area proportional to frequency.

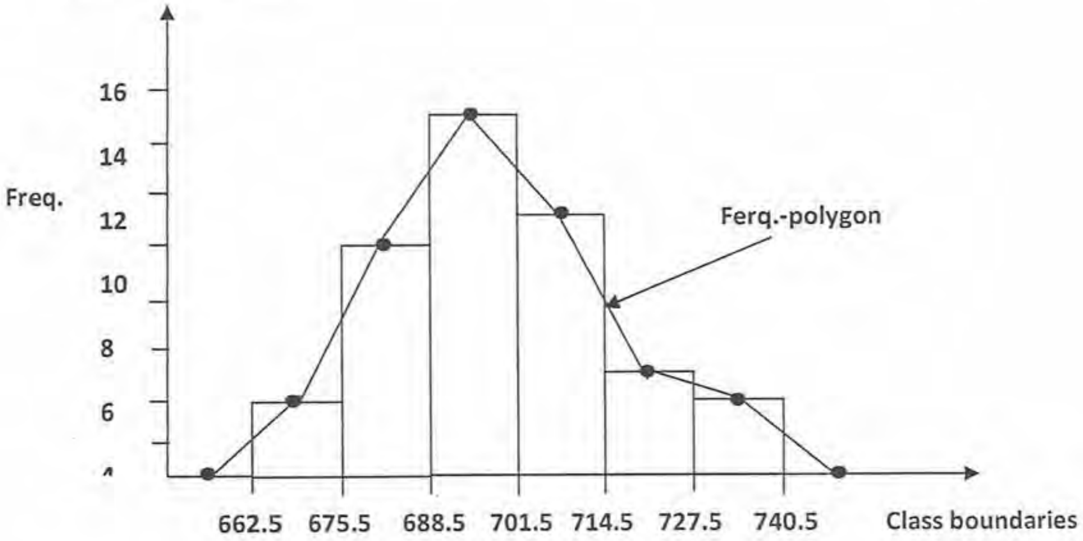
2) Frequency polygon : Consists of lines joining the class mark with freq., it may be obtained from the histogram by joining the mid-points of the bar tops.

3) Frequency curve : Is a smoothed frequency polygon in to a continuous curve.

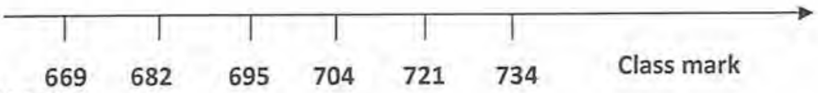
4) Cumulative freq. curve (Ogive) : Is a smoothed cumulative freq. polygon. It is usually S-shape.

Graphical presentation of frequency dist. Table :

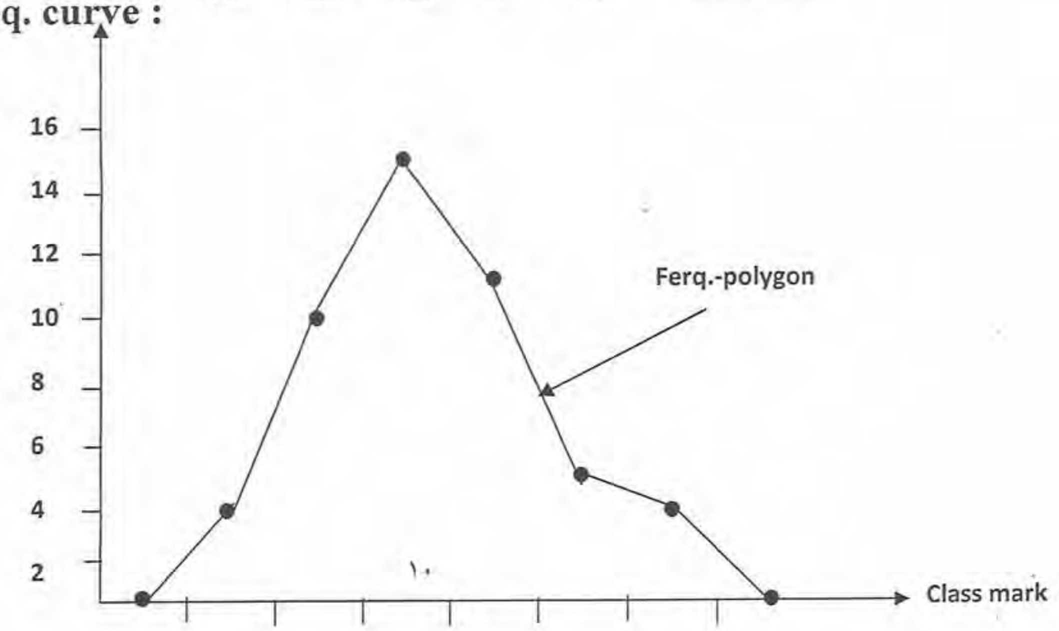
1) Histogram :



2) freq. polygon :

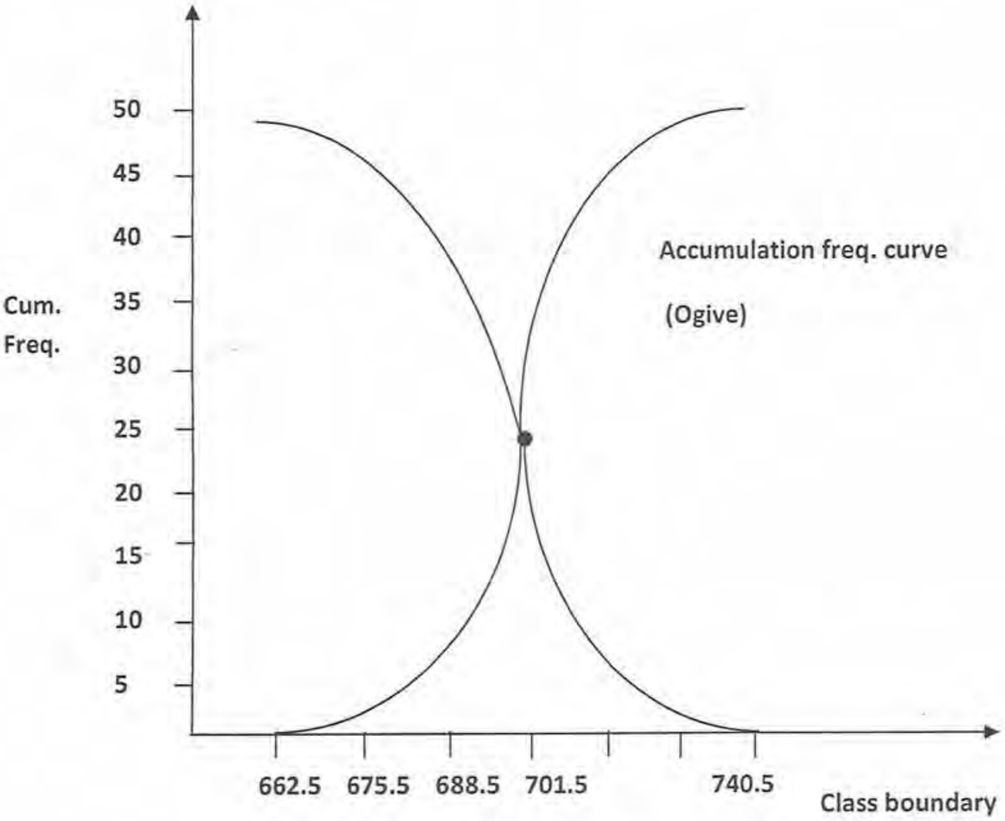


3) freq. curve :



4) Cumulative freq. :

Ascending cum. Freq.		Desending cum.	
Upper boundaries	cum. Freq.	Upper boundaries	cum. Freq.
Less than 662.5	0	Greater than 662.5	50
Less than 675.5	4	Greater than 675.5	46
Less than 688.5	14	Greater than 688.5	36
Less than 701.5	29	Greater than 701.5	21
Less than 714.5	40	Greater than 714.5	10
Less than 727.5	46	Greater than 727.5	4
Less than 740.5	50	Greater than 740.5	0



Tutorial Sheet No. (1)

For the following data groups obtain :

1) Frequency distribution table.

2) f_r , f_p , f_c

3) Histogram, freq. polygon, and Ogives

Data 1)

6.3	7.0	7.5	9.0	7.7	7.8	7.1	8.1
6.6	7.2	8.3	8.5	6.9	7.7	8.0	7.3
8.6	7.1	8.7	6.4	7.7	7.4	8.0	7.6
7.5	7.2	7.5	8.8	7.8	7.9	7.3	7.0
6.8	8.1	8.4	6.7	7.1	8.2	8.1	7.7

Data 2)

5.4	4.1	5.2	2.8	4.9	5.6	4.0	4.1	4.3
3.9	4.5	6.1	3.7	2.3	4.5	4.9	5.6	4.3
4.2	3.2	5.0	4.8	3.7	4.6	5.5	1.8	5.1
5.1	6.3	3.3	5.8	4.4	4.8	3.0	4.3	4.7

Data 3)

12.16	12.38	12.21	12.55	12.22	12.40	12.43	12.35
12.31	12.07	12.31	12.33	12.56	12.41	12.42	12.44
12.30	12.39	12.10	12.37	12.18	12.48	12.19	12.43
12.25	12.37	12.47	12.49	12.35	12.28	12.30	12.31
12.35	12.20	12.39	12.54	12.59	12.29	12.46	12.09

Chapter (3)

Measures of Location

When raw data is classified in to a frequency distribution table and presented graphically, the major features of the sample become apparent. However, to make quantitative decisions, further condensation in to a number of statistical parameters is needed.

Measures of location are statistical parameters, giving an estimate of the data centre, being typical of all measurement.

Mode : Is the measurement that occurs with the greatest frequency.

e. g. for sample :

14 , 19 , 16 , 21 , 19 , 24 , 18 , 19

Mode = 19

For sample : 6 , 7 , 7 , 3 , 8 , 3 , 9 , 5

Mode = 3 , 7

(bimodal)

For grouped data, the mode corresponds to the top of the frequency curve.

$$mode = L_m + \frac{\Delta L}{\Delta L + \Delta H} C_m$$

Where:

L_m is lower boundary of modal class

$\Delta L = f_m - f_{\text{lower class}}$

$\Delta H = f_m - f_{\text{higher class}}$

C_m = width of modal class

e. g. for electric bulbs sample :

$$mode = 688.5 + \frac{15-10}{(15-10)+(15-11)}(13) = 695.7$$

<u>Median :</u> Is the middle measurement of an ordered array (odd). Or the arithmetic mean of the two middle values (even).

e. g. for sample : 3 , 4 , 4 , 5 , 6 , 8 , 8 , 10 , 11

median = 6

for sample : 5 , 5 , 7 , 9 , 11 , 12 , 15 , 18

median = 10

* For grouped data, the median line halves the area under the frequency curve.

$$median = L_m + \frac{\frac{N}{2} - f_{CL}}{f_m} C_m$$

Where :

L_m is lower boundary of median class

N is sample size

F_{CL} is cumulative frequency of lower class

f_m is frequency of median class

C_m is width of median class

e. g. for electric bulbs sample :

3rd class is median class, since $f_c = 29 > \frac{N}{2}$

median = $688.5 + \frac{\frac{50}{2} - 14}{15}(13) = 689.0$

Arithmetic Mean: is the sum of measurements divided by sample size.

$$\bar{x} = \frac{\sum x_i}{N}$$

For grouped data :

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

e. g. for electric bulbs sample:

$\bar{x}$ = [(4)(669) + (10)(682) + (15)(695) + (11)(708) + (6)(721) + (4)(734)] / 50

$\bar{x}$ = 699.4

Relation between mode/median/mean :

For symmetrical distributions, the three measures coincide. Else, the mean is further removed from mode than is the median. For moderately skewed unimodal distributions :

Mean – mode ≈ 3 (mean – median)

Other Mean Measures :

** Geometric Mean* $G = (\pi x_i)^{\frac{1}{N}}$, $\log G = \frac{\sum f_i \, log X_i}{N}$

** Harmonic Mean* $H = \frac{N}{\sum \frac{1}{x_i}}$, $H = \frac{N}{\sum \frac{f_i}{x_i}}$

** Root Mean Square*

$$RMS = \sqrt{\frac{\sum x_i^2}{N}} \quad , \quad RMs = \sqrt{\frac{\sum f_i \, x_i^2}{N}}$$

For a sample of positive measurements,

$H \le G \le \bar{x} \le RMS$

e. g. for electric bulbs sample :

$\bar{x}$ = 699.4

G = 699.2

H = 699.0

RMS = 699.6

Properties of the Arithmetic Mean :

1. The sum of deviations of the data from their arithmetic mean is zero.

$\sum(x_i - \bar{x}) = 0$ (prove)

2. For several samples, the combined mean is given by:

$$\bar{x} = \frac{N_1\overline{x_1} + N_2\overline{x_2} + \ldots}{N_1 + N_2 + \ldots}$$

3. If the deviations (d$_i$) from any value (A) are available, then :

$\bar{x} = A + \frac{\sum d_i}{N}$ $where\ d_i = x_i - A$ (prove)

Or $\bar{x} = A + \frac{\sum f_i d_i}{N}$ (grouped data)

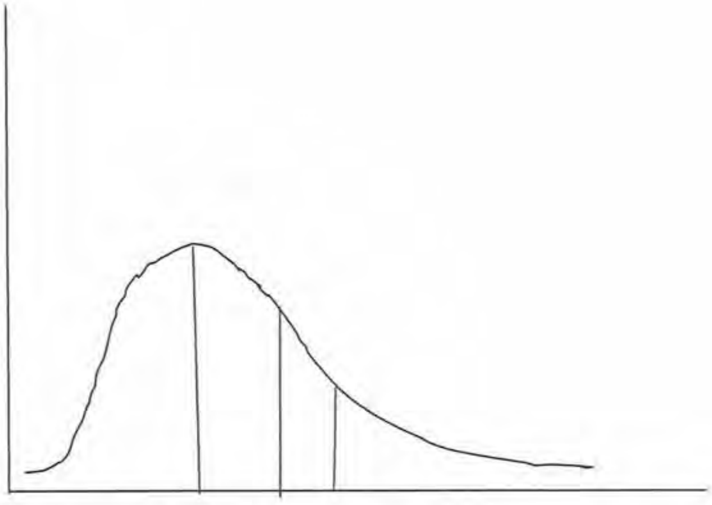
Empirical Relation between mean, median and mode :

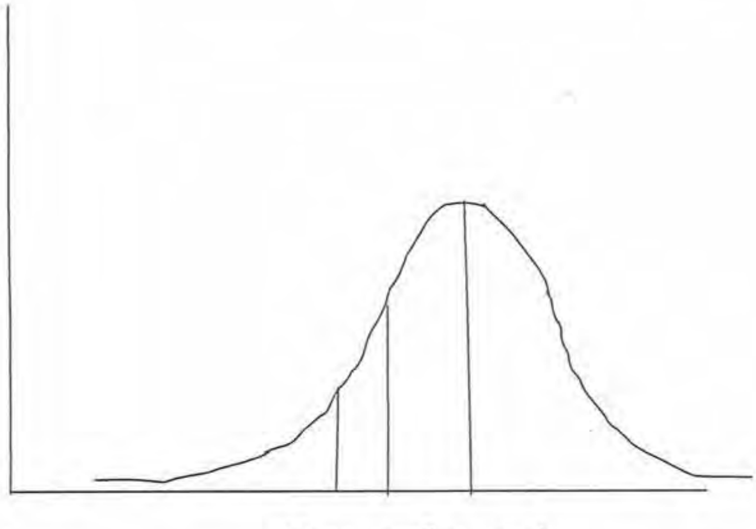
For unimodal freq. curves which are moderately skewed (asymmetrical), where the empirical relation.

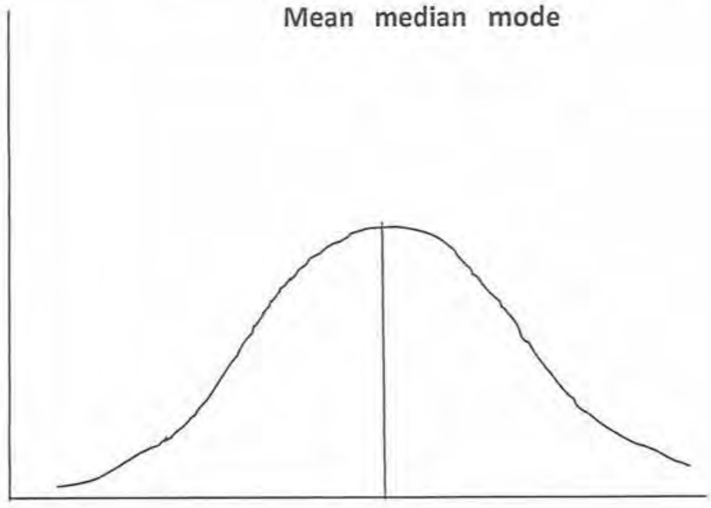
Mean – mode = 3 (mean – median)

In figs. Below are shown the relative position of the mean, median and mode for freq. curve which are skewed to the right and left resp.

For symmetrical curves the mean, mode and median al coincide.


Mode median mean


Mean median mode


Mode
mean
Median

Chapter (4)

Measures of Dispersion

Dispersion is the degree of data spread about an average. Several measures are used including:

Range, mean absolute deviation , standard deviation , variance and coefficient of variation.

Mean Absolute Deviation :

Is the arithmetic mean of the absolute deviations.

$$M.A.D = \frac{\sum |x_i - \bar{x}|}{N} \qquad \text{for raw data}$$

$$= \frac{\sum f_i \, |x_i - \bar{x}|}{N} \qquad \text{for grouped data}$$

Mean other than $\bar{x}$ may be used to obtain M.A.D from the respective mean.

Standard Deviation :

Is the root mean square of the deviation.

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \qquad \text{for raw material}$$

$$S = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}} \qquad \text{for grouped data}$$

Means other than $\bar{x}$ may be used to obtain the standard deviation from the respective mean .

Standard deviation of a sample (S) is related to the standard deviation of the population (σ) by :

$$\sigma = S\sqrt{\frac{N}{N-1}} = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N-1}}$$

Variance :

Is the square of the standard deviation i.e. S^2 for sample, σ^2 for population.

Coefficient of Variation :

Is a relative dispersion measure (dimension less).

$$Relative\ Dispersion = \frac{absolute\ dispersion}{average}$$

$$Coefficient\ of\ Variation = \frac{S}{\bar{x}}$$

Properties of Standard Deviation :

* Of all standard deviations, the min. is that from the arithmetic mean.

* For ideal normal distributions:

With in $\bar{x} \mp S$	$68.27\%\ of\ data$
$\bar{x} \mp 2\,S$	$95.45\%\ of\ data$
$\bar{x} \mp 3\,S$	$99.73\%\ of\ data$

* For several samples, the combined S is given by :

$$S^2 = \frac{N_1\, S_1^2 + \, N_2\, S_2^2 + \, ...}{N_1 + \, N_2 + \cdots}$$

Standard Variable :

The dimensional measurements x_i may be expressed as dimension less standardized variables Z_i

$$Z_i = \frac{x_i - \bar{x}}{S} = \frac{(\bar{x} + S) - \bar{x}}{S} = 1$$

i.e. when Z=1 , the measurement is removed by one standard deviation from the mean.

Properties of Z :

1. The arithematic mean for the standard scores equal to zero.

$$\acute{Z} = \frac{\sum f_i Z_i}{N} = 0$$

2. The standard deviation (or variance) for the standard scores equal to **one**.

$$S_Z = \sqrt{\frac{\sum f_i (Z_i - \bar{Z})^2}{N}} = 1$$

$$S_Z^2 = \frac{\sum f_i (Z_i - \bar{Z})^2}{N} = 1$$

Tutorial Sheet No. (2)

Q 1) Prove the following

A. $$S = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

B. $$S = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Where $d_i = x_i - A$ where A is constant

Q 2) For the following data :

Class limits	f
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 - 74	8

Obtain :

A) S , S^2 B) Z C) $\acute{Z}$, S_z

Chapter (5)

Probability Distribution

* Probability : When an event may happen in (x) ways out of a total of (n) possible equally likely ways, the probability of occurrence (success) is given by :

$$p = Pr(E) = \frac{x}{n}$$

Hence the prob. Of non-occurrence (failure) is :

$$q = \Pr(\tilde{E}) = \frac{n-x}{n} = 1 - \frac{x}{n} = 1 - p$$

Thus $\quad p + q = 1$

Discrete Prob. Distribution :

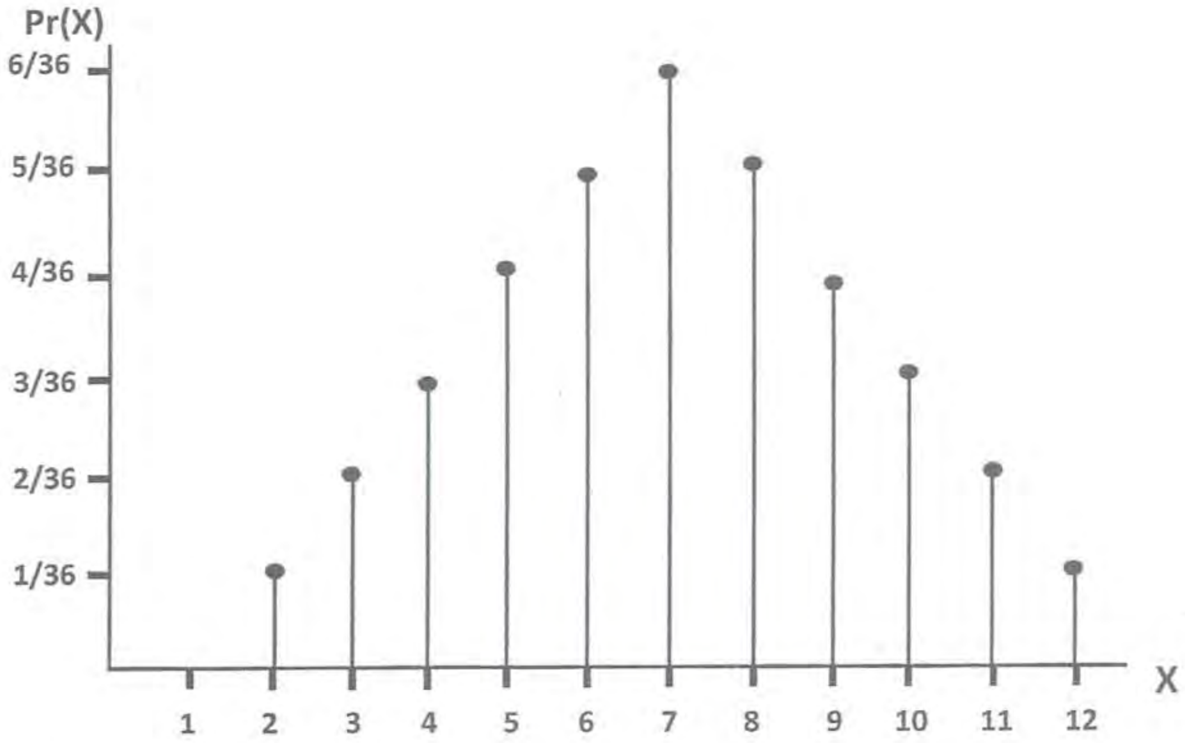
If a variable X may assume a set of discrete values (x_i) with respective prob. p_i , where $\sum p_i = 1$, this defines a discrete prob. distribution for X.

Example :

Let X be the sum of points obtained on a throw of two dice. The prob. or frequency distribution is given as :

X :	2	3	4	5	6	7	8	9	10	11	12
Pr(x):	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$$\sum \Pr(X) = 1$$



* Relative frequency distribution of (N) throws is thus related to a sample of size (N) drawn out of an infinite population.

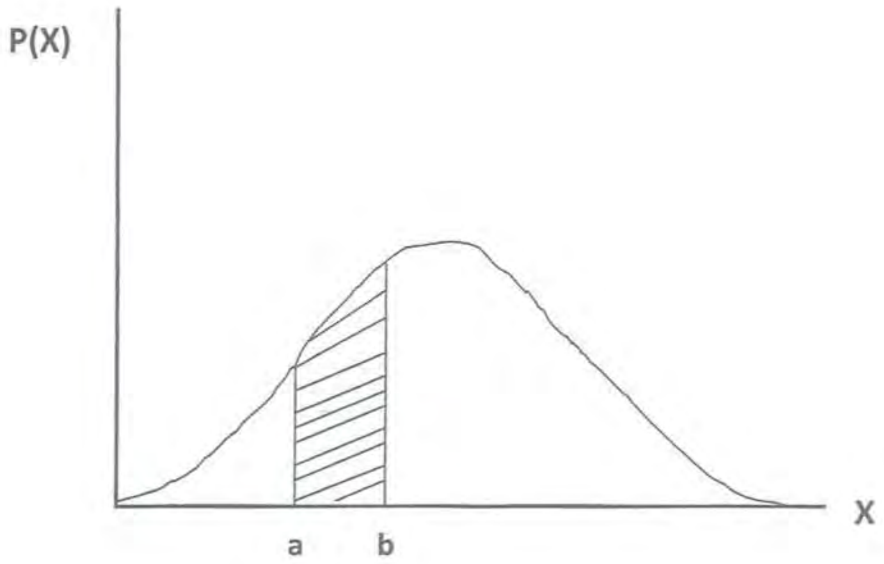
As $N \rightarrow \infty$, the relative freq. dist. Approaches the prob. dist. Of the population .

Continuous Probability Distribution :

* If a variable X may assume a continuous set of values, the prob. dist. Is a frequency curve where p(x)=fr .

Total area under the curve = $\sum fr = \sum p = 1$.

* Prob. that X may lie between a and b ; Pr [$a < X < b$] = area under curve from a to b .



The Normal Distribution

* The normal distribution is the most important of all probability distribution. It is applied directly to many practical problems, and several very useful distributions are based on it .

It is some times called the Gaussin dist. .

Characteristics :

Many empirical freq. dist. Have the following characteristics :

1. They are approximately symmetrical, and the mode is close to the centre of the dist.
2. The mean, median, and mode are close together.
3. The shape of the dist. Can be approximated by a bell.

*** The prob. density function for the normal dist. Is given by:**

$$f(x) = \frac{1}{\sigma\sqrt{2\,\pi}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}$$

Where :

μ : is the mean of the theoretical dist.

σ : is the standard deviation, and $\pi = 3.14$

*** This function extends from $-\infty$ to ∞**

Let $z = \frac{X-\mu}{\sigma}$,

$$f(z) = \frac{1}{\sigma\sqrt{2\,\pi}}\, e^{-\frac{1}{2}z^2}$$

* The total area bounded by the curve and the X axis is one. Hence, the area under the curve between X=a and X=b , Where a < b represent the prob. that X lies between a and b.

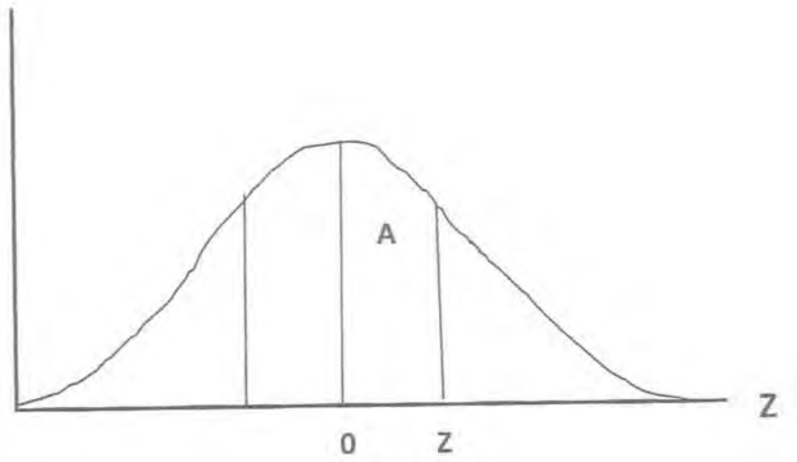
* Areas under the normal dist. Curve between 0 and Z are given in a :

Table (*). (Given below)

The prob. that Z lies between 0 and Z :

Pr [0 < Z < z] = A

From table (*) the area between any two ordinates can be found by using the symmetry of the curve about Z=0 .



*** Some properties of the normal dist. :**

$Mean = \mu$

$Variance = \sigma^2$

$Standard\ dev. = \sigma$

$Mean\ dev. = \sigma \sqrt{\frac{2}{\pi}} = 0.7979\ \sigma$

*** Areas under the normal curve :**

When $Pr[0 < Z < Z_1] = A$

$Pr[-Z_1 < Z < 0] = A$ (symmetrical curve).

$Pr[Z_1 < Z] = 0.5 - A$
$Pr[-Z_1 < Z] = 0.5 + A$ } Tatal area=1 so that area from $0 \rightarrow \infty$ is 0.5

When Z_1 and Z_2 are of same signs :

$Pr[Z_1 < Z < Z_2] = A_{Z_2} - A_{Z_1}$

When Z_1 and Z_2 are of different signs :

$Pr[Z_1 < Z < Z_2] = A_{Z_2} + A_{Z_1}$

* For bound of measurements, the bound in actual value is ∓ 0.5 units in L.S.D.

Example :

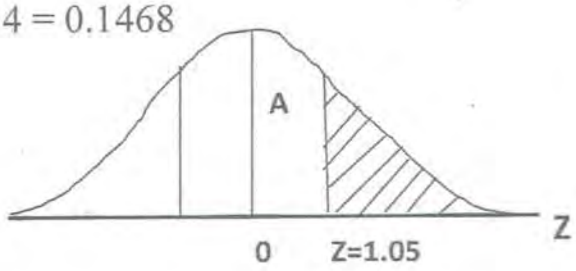
For measurements of $\mu = 160$, $\sigma = 10$, obtain the following :

1. Pr[X greater than 170] = Pr[X >170]

$$Z = \frac{X-\mu}{\sigma} = \frac{170.5-160}{10} = 1.05$$

$\therefore$ Pr[Z > 1.05]= = 0.5 − 0.35314 = 0.1468

Where $A_{Z=1.05} = 0.35314$

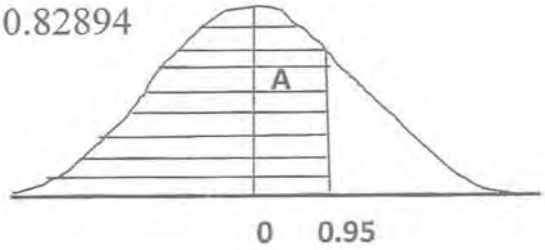


2. Pr[X less than 170] = Pr[X<170]

$x = 169.5 \rightarrow Z = 0.95$

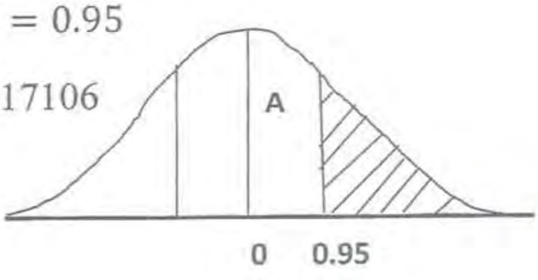
Pr [Z < 0.95] = 0.5 + 0.32894 = 0.82894

Where $A_{Z=0.95} = 0.32894$



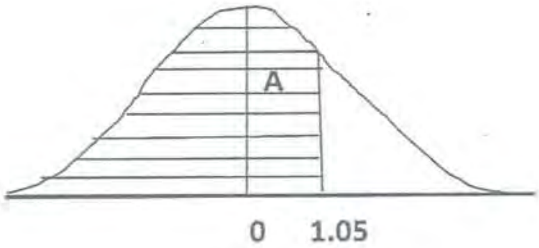
3. Pr[X ≥ 170] $\rightarrow x = 169.5 \rightarrow Z = 0.95$

Pr [Z > 0.95] = 0.5 − 0.32894 = 0.17106



4. Pr[X ≤ 170] → $x = 170.5$ → $Z = 1.05$

Pr [Z > 1.05] = 0.5 + 0.35314 = 0.85314



Linear inter polation :

When Z lies between successive Z_1 and Z_2 with respective A_1 and A_2 , A is obtained by linear interpolation "

$$\frac{Z - Z_1}{Z_2 - Z_1} = \frac{A - A_1}{A_2 - A_1} \qquad , Z_1 < Z < Z_2$$

e.g. for $Z_1 = 2.32$ $A_1 = 0.48983$

$Z_2 = 2.33$ $A_2 = 0.49010$

Then when Z = 2.327 → A = ?

$$A = \frac{Z - Z_1}{Z_2 - Z_1}(A_2 - A_1) + A_1 = 0.49002$$

Example 1)

For a measurement of size (N)=500

$\mu = 151$, $\sigma = 15$, assuming normal dist. Find how many measurements :

a) between 120 and 155 = $\Pr[120 \le X \le 155]$

$x_1 = 119.5$ → $z_1 = -2.1$ → $A_1 = 0.4821$

$x_2 = 155.5$ → $z_2 = 0.30$ → $A_2 \doteq 0.1179$

Pr $[\overset{z_1}{-2.1} < Z < \overset{z_2}{0.3}] = 0.4821 + 0.1179 =$ No. of meas. = 500[0.4821+0.1179]=300

b) more than 185 = Pr[Z>185]

$x = 185.5 \rightarrow Z = 2.3 \rightarrow A = 0.4893$

Pr [Z > 2.3] = 0.5 − 0.4893 = No. of meas. = 500[0.51−0.4893]=5.

c) Less than 128 = Pr [X < 128]

$x = 127.5 \rightarrow Z = -1.57 \rightarrow A = 0.4418$

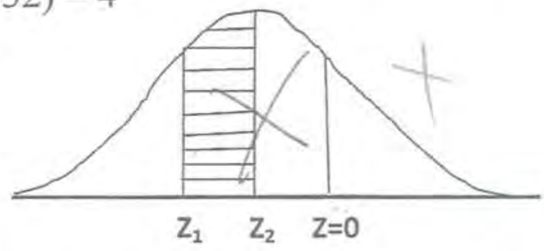
No. of meas. = 500[0.5−0.4418]=29

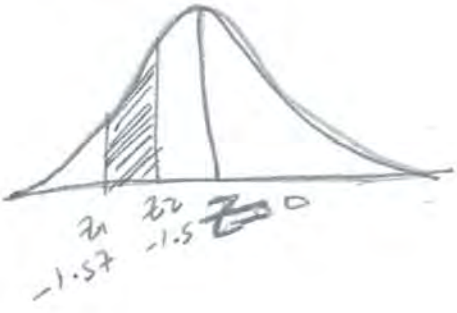
d) equal to 128 = Pr [X=128]

$x_1 = 127.5 \rightarrow z_1 = -1.57 \rightarrow A_1 = 0.4418$

$x_2 = 128.5 \rightarrow z_2 = -1.5 \rightarrow A_2 = 0.4332$

Pr [−1.57 < Z < −1.5] = 0.4418−0.4332 = No. of meas. = 500(0.4418−0.4332) = 4





e) Less than or equal to 128 = Pr [X≤128]

$x = 128.5 \rightarrow Z = -1.5 \rightarrow A = 0.4332$

No. = 500[0.5−0.4332]=33

f) Less than or equal to 185 = Pr [X≤185]

$x = 185.5 \rightarrow Z = 2.3 \rightarrow A = 0.4893$

No. = 500[0.5 + 0.4893] = 495

Example 2)

For a sample of washers produced by a machine the mean inside dia. (μ) is 5.02 mm and the standard deviation is 0.05 mm. The max. useful tolerance in the dia. Is 4.96 to 5.08 mm, otherwise the washers are considered defective. Determine % of defective washers.

Solu.)

Pr of max. to lerance = Pr $(4.96 \le X \le 5.08)$

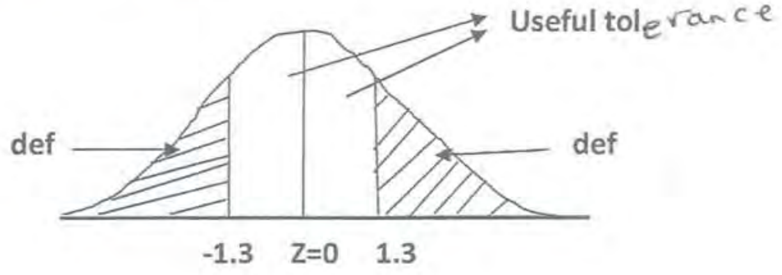
One unit in L.S.D. = 0.01

$x_1 = 4.955 \rightarrow z_1 = -1.3 \rightarrow A_1 = 0.4032$

$x_2 = 5.085 \rightarrow z_2 = +1.3 \rightarrow A_2 = 0.4032$

Pr $[-1.3 < Z < 1.3] = 2 * 0.4032 = 0.8064$

$\therefore$ % of defective washers = $(1-0.8064) * \ = 19.4$ %



Example 3)

Out of a large No. of examination applicant a sample of size 50 gave a mean mark of 64 and a standard dev. of 14 . What is the expected % of applicants achieving a min. pass mark of 50 ?

Solu. :

one unit = 1

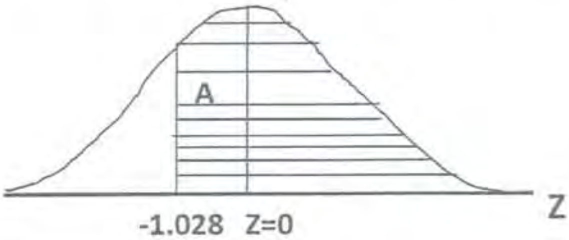
$\mu = 64$

$$\sigma = S\sqrt{\frac{N}{N-1}} = 14\sqrt{\frac{50}{50-1}} = 14.1$$

Pr [app. Have a min. pass mark of 50]= Pr [50≤X]

$x = 49.5 \rightarrow Z = -1.028 \rightarrow A = 0.3480$

Pr $[-1.028 < Z] = 0.348 + 0.5 = 0.848 = 84.8\ \%$



Example 4)

The strength of individual bars made by a certain manufacturing process are approximate normally distributed with mean 28.4 and standard dev. 2.95 . To ensure safety, a customer requires at least 95% of the bars to be stronger than 24.0 . (one unit =0.1)

a) Do the bars meet the specification ?

b) By improved manufacturing techniques, the manufacturer make the bars more uniform (that is, decrease the standard dev.) what value of standard dev. will just meet the specification if the mean stays the same ?

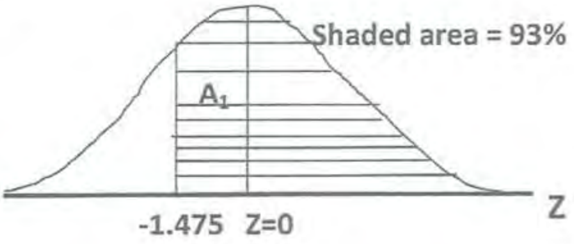
Solu. :

Pr [X > 24.0]→ $Z_1 = \frac{X-\mu}{\sigma} = \frac{24.05-28.4}{2.95}$

$$Z_1 = -1.475 \rightarrow A_1 = 0.4299$$

Pr [Z > - 1.475] = 0.5 + 0.4299 = 0.9299 ≃93%

Since (93%) less than 95% , the bars do not meet the specification.



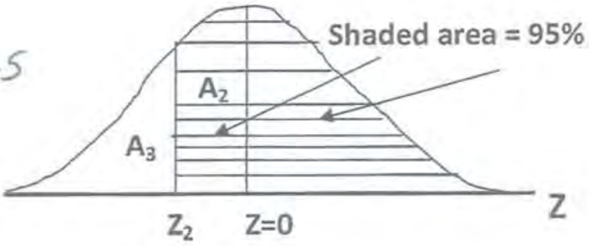
b) The specification is at least 95% of bars > 24.0

$A_3 = 1 - 0.95 = 0.05$

$\therefore\ A_2 = 0.5 - 0.05 = 0.45$

oR: $A_2 : 0.95 - 0.5$

$A_2 = 0.45$



At $A_2 = 0.45 \rightarrow Z_2 = -1.645$ (from table)

$$\underline{Z_2} = \frac{X-\mu}{\sigma} \rightarrow -1.645 = \frac{24.05 - 28.4}{\sigma}$$

$\sigma = 2.644$ (if the σ can be reduced to 2.644 while keeping the mean constant, the specification will just be met)

Normal Distribution

Tutorial sheet (4)

Q. 1) Diameters of bolts produced by a machine are normally distributed with $\mu = 0.76$ *and* $\sigma = 0.012$ cm. Specifications call for dia. From 0.72 cm to 0.78 cm.

a) What percentage of bolts will meet there specification ?

b) What percentage of bolts will be smaller than 0.73 cm. ?

Q. 2) The diameters of screws are normally distribution with $\mu = 2.1$ *and* $\sigma = 0.15$ cm .

a) What proportion of screws are expected to have dia. Greater than 2.5 cm. ?

b) A specification calls for screw dia. Between 1.75 cm and 2.5 cm . What proportion of screws will meet the specification ?

Q. 3) Diameters of ball bearings produced by a company follow a normal distribution. If the mean dia. is 0.4 cm and stand. dev. is 0.001 cm.

a) What percentage of the bearings can be used o a machine specifying a size of 0.399 $\mp$0.0015 cm. ?

b) What is the upper bound of the size range that has a lower bound of 0.398 cm . and include 80% of the bearings ?

Q. 4) The probability that a river flow exceeds 2000 m^3/sec is 15% . The coefficient of variation of these flows is 20% . Assuming a normal distribution, calculate :

a) The mean of the flow ?

b) The stand. dev. of the flow ?

c) The prob. That the flow will be between 1300 or 1900 m^3/sec?

Q. 5) A water quality parameter monitored in a lake is normally dist. With $\mu = 24.3$. It is also known that there is 70% probability that the parameter will exceed 17.6 :

a) Find the stand. dev. of the parameter ?

b) If the parameter exceeds the 95% an investigation of a local industry begins. What is this critical value ?

The Binomial Distribution

* If P is the prob. That an event will happen in any single trial (called the prob. Of succeed)and q=1−P is the prob. That it will fail to happen in any single trial (called the prob. Of a failure), then the prob. That the event will happen exactly X times in N trials (X success and N−X failures will occur) is given by :

$$P(X) = {}_N C_X \, P^X \, q^{N-X} = \frac{N\,!}{X\,!\,(N-X)!} \; P^X \, q^{N-X}$$

Where :

X = 0 , 1 , 2 , … , N

N ! = N (N−1) (N−2) …1

0 ! = 1

Example 1)

The prob. Dist. For getting heads in N tossed of a coin is :

$$P(X) = {}_N C_X \left(\frac{1}{2}\right)^X \left(\frac{1}{2}\right)^{N-X}$$

For example N=3 ?

Pr (0 heads) = ${}_3C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{3-0}$

$$= \frac{3\,!}{0\,!\,(3-0)!} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{3-0} = \frac{1}{8}$$

Pr (1 heads) = ${}_3C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{3-1} = \frac{3}{8}$

Pr (2 heads) = $_3C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{3-2} = {_3C_2}\left(\frac{1}{2}\right)^3 = \frac{3}{8}$

Pr (3 heads) = $_3C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{3-3} = {_3C_3}\left(\frac{1}{2}\right)^3 = \frac{1}{8}$

H = head , T = tail

HHH , HHT , HTT , TTT , THH , TTH , THT , HTH

Total out coins = 8 = $(2)^N = 2^3 = 8$

Where 2 = P(head + tail)

Example 2)

Find the prob. Of getting :

a) Pr (2 H in 6 tosses) = $_6C_2\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{6-2}$

$$= \frac{6\,!}{2\,!\,(6-2)!}\left(\frac{1}{2}\right)^6 = \frac{15}{64}$$

Total out coins : $(2)^6 = 64$

b) Pr (at least 4 H in 6 tosses) = Pr (X≥4)

= Pr (X=4) + Pr (X=5) + Pr (X=6)

∴ $\Pr(X\geq4) = [\,{_6C_4} + {_6C_5} + {_6C_6}]\,[\frac{1}{2}]^6 = \frac{22}{64}$

Example 3)

If 10% of items produced by a machine are defective, find the prob. that out of 5 items :

a) None are defective

P = prob. of success (defection).

P = 0.1 → q + p =1 → q=0.9

Q = 0.9 = prob. of failure (non def.)

Pr (o def.) = Pr (X=0) = $_5C_0\ (0.1)^0\ (0.9)^5$

Pr (X=0) = $\frac{5\,!}{0\,!\,(5-0)!}\ (0.1)^0\ (0.9)^5$ = 0.5905

b) All are def.

Pr (5 def.) = Pr (X=5) = $_5C_5\ (0.1)^5\ (0.9)^5$

$$= \frac{5\,!}{5\,!\,(5-5)!}\ (0.1)^5\ (0.9)^0 = 0.00001$$

c) At most 2 def.

Pr (at most 2 def.) = Pr (X≤2) = Pr (X=2) + Pr (X=1)+ Pr (X=0)

Pr (X≤2) = $_5C_2\ (0.1)^2\ (0.9)^3 +\ _5C_1\ (0.1)^1\ (0.9)^4 +$

$$_5C_0\ (0.1)^0\ (0.9)^5$$

$$= 0.00729 + 0.32805 + 0.59049 = 0.9258$$

Some properties of Binomial dist :

Mean = μ = NP

Variance = σ^2 = NPq

Stand. dev. = $\sigma = \sqrt{N\ P\ q}$

The Binomial

Tutorial sheet (5)

Q. 1) Out of 800 families with 5 children howmany would you expect to have : a) 3 boys , b) 5 girls , c) either 2 or 3 boys . Assume equal probabilities for boys and girls ?

Q. 2) Find the prob. of getting a total of 11 : a) once , b) twice , in two tosses of a pair dice ?

Q. 3) What is the prob. of getting a 9 exactly once in 3 throws with a pair of dice ?

Q. 4) Find the prob. of guessing correctly at least6 of the 10 answers on a atrue – false examination ?

Q. 5) An insurance salesman sells policies to 5 men the prob. that a mean will be a live in 30 years is 2/3. Find the prob. that in 30 years :

a) all 5 men , b) at least 3 men , c) only 2 men , d) at least 1 man will be a live ?

Relation between Binomial and Normal Distributions

* If N is large and if neither P nor q is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized variable given by :

$$Z = \frac{X - \mu}{\sigma} = \frac{X - NP}{\sqrt{N\,P\,q}}$$

The approximation becomes better with increasing N , . In practice, the approximate is good when : NP > 5 , Nq > 5 .

Example 1)

Find the prob. of obtaining 3 – 6 heads in 10 tosses of a coin :

a) using the Binomial dist.

Pr(3-6 heads) = Pr(3) + Pr(4) + Pr(5) + Pr(6)

$$= \left[{}_{10}C_3 + {}_{10}C_4 + {}_{10}C_5 + {}_{10}C_6\right]\left[\frac{1}{2}\right]^{10} = 0.7734$$

b) using the normal dist.

one unit= 1.0

Pr (3-6 heads) = Pr $(3 \le X \le 6)$

$$Z = \frac{X-\mu}{\sigma} = \frac{X-NP}{\sqrt{N\,P\,q}}$$

μ = NP = 10*0.5 = 5

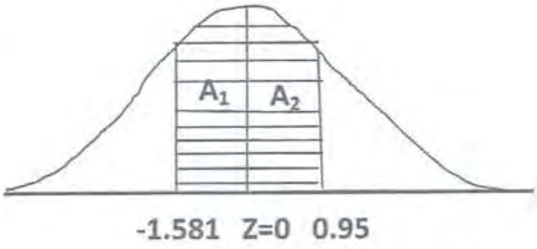
$\sigma = \sqrt{N\,P\,q} = \sqrt{10 * 0.5^2} = 1.581$

For $X_1 = 2.5 \rightarrow Z_1 = -1.581 \rightarrow A_1 = 0.4431$

$X_2 = 6.5 \rightarrow Z_2 = 0.949 \rightarrow A_2 = 0.3287$

Pr $(-1.581 < Z < 0.949) = A_1 + A_2 = 0.7718$

% of error between the binomial and normal = 0.0016



Example 2)

What is the prob. That at most 90% of 20 students will graduate ? Given % of graduate 70% ?

a) using the Binomial dist.:

Pr (at most $\frac{90}{100} * 20$) = Pr (at most 18)

Pr $(X \leq 18)$ = Pr(0) + Pr(1) + … + Pr(18)

$= 1-[\Pr(19) + \Pr(20)]$

Pr (X=19) = ${}_{20}C_{19}\ (0.7)^{19}\ (0.3)^{1} = 6.84*10^{-3}$

Pr (X=20) = ${}_{20}C_{20}\ (0.7)^{20}\ (0.3)^{0} = 7.98*10^{-4}$

Pr (X≤18) = $1-(6.84*10^{-3} + 7.98*10^{-4}) = 0.992$

b) using the normal dist.

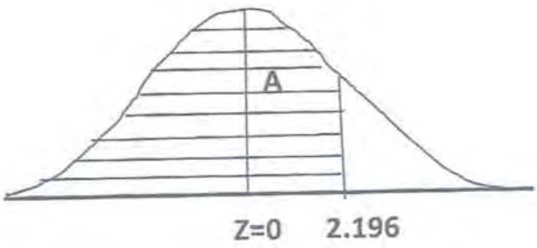
$$Z = \frac{X - \mu}{\sigma} = \frac{X - NP}{\sqrt{N\ P\ q}}$$

μ = NP = 20*0.7 = 14

$\sigma = \sqrt{N\ P\ q} = 2.049$

Pr (X≤18) → X = 18.5 → Z = 2.196 → A = 0.48

Pr (Z < 2.196) = 0.5 + 0.486 = 0.986



Relation between Binomial and Poisson :

Poisson dist. :

Is a discrete prob. dist. Defined by :

$$Pr(X) = \frac{\lambda^x e^{-\lambda}}{x\,!}$$

Where :

X = 0 , 1 , 2, … , N

e = 2.71828

With properties :

$\mu = \lambda \quad , \sigma = \sqrt{\lambda}$

Example 1)

Product of a machine is 10% defective, Find the prob. of obtaining at most 2 def. items out of 10 :

a) using Binomial :

Pr [at most 2] = Pr($X \leq$ 2)=Pr(0) + Pr(1) + Pr(2)

$$= {}_{10}C_0\,(0.1)^0\,(0.9)^{10} + {}_{10}C_1\,(0.1)^1\,(0.9)^9 + {}_{10}C_2\,(0.1)^2\,(0.9)^8$$

Pr ($X \leq 2$) = 0.9298

b) using Poisson :

Poisson is applicable for large N , While P is close to zero.

In practice , N≥50 , NP < 5 .

Use $\mu = NP = \lambda = 10 * 0.1 = 1$

Pr (at most 2) = Pr(X≤ 2) = Pr(0) + Pr (1) +Pr(2)

$$= \frac{1^0 e^{-1}}{0\,!} + \frac{1^1 e^{-1}}{1\,!} + \frac{1^2 e^{-1}}{2\,!} = \left[1 + 1 + \frac{1}{2}\right] e^{-1}$$

$$= 0.9197$$

Example 2)

The prob. of failure of a certain process is 3% . Determine the prob. of 3 failures at most in 100 repetition of the process :

a) using binomial dist.

Pr (at most 3) = Pr($X \leq 3$)=

$${}_{100}C_0\,(0.03)^0\,(0.97)^{100} + \cdots + {}_{100}C_3\,(0.03)^3\,(0.97)^{97}$$

= 0.6474

b) using Poisson dist. :

N = 100 , NP = = $100 * \frac{3}{100}$ = 3 = λ

$$\text{Pr (X≤3)} = \frac{3^0 e^{-3}}{0\,!} + \cdots + \frac{3^3 e^{-3}}{3\,!}$$

$$= \left[\frac{3^0}{0\,!} + \frac{3^1}{1\,!} + \frac{3^2}{2\,!} + \frac{3^3}{3\,!}\right] [e^{-3}] = 0.6472$$

Use Normal probability plots to assess normality :

- A normal prob. plot is a graph that plots observed data versus normal scores, (expected Z – scores).

- Drawing a normal prob. plot requires the following step :

1. Arrange the data in ascending order.

2. Compute $\left(f_i = \frac{i-0.375}{n+0.25}\right)$ where i is the index of the data, n . number of observation.

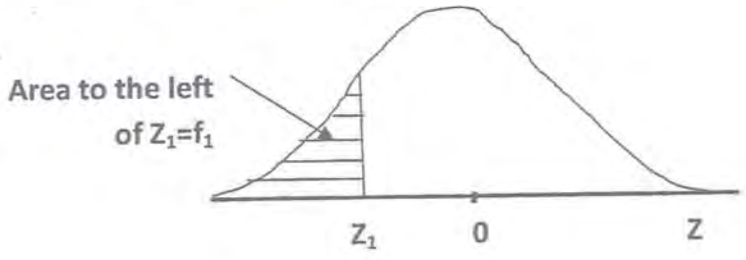
3. Find the Z – score corresponding to f_i from the table of the normal curve.

4. plot the observed values on the horizontal axis and the corresponding expected Z-scores on the vertical axis.

- The value of f_i represents the expected area to the left of the ith observation when the data come from a population that is normally distributed.

- For example, f_1 is the expected area to the left of the smallest data value.

- Values of normal random variables and their Z-scores are linearly related $(X = \mu + Z\sigma)$, so a plot of observation of normal variable against their expected Z-scores will be linear. We conclude the following :



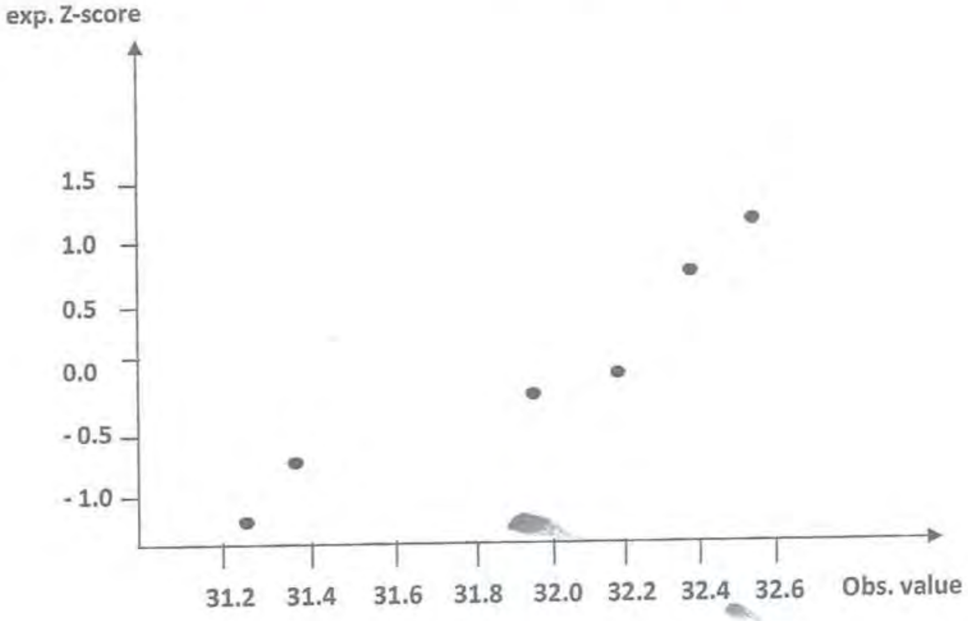
(If the sample data are taken from a population that is normally distributed, a normal prob. plot of the observed values vs. The expected Z-scores will be approximately linear).

Example : For the following data, construct the normal prob. plot. :

31.35 35.52 32.06 31.26 31.91
32.37

Solution :

Index, i	Observed value	Fi	Expected Z-score
1	31.26	$\frac{1-0.375}{6+0.25}=0.1$	- 1.28
2	31.35	$\frac{2-0.375}{6+0.25}=0.26$	- 0.64
3	31.91	0.42	- 0.20
4	32.06	0.58	0.20
5	32.37	0.74	0.64
6	35.52	0.9	1.28



* Although the normal prob. plot does show some curvature, it is roughly linear. We conclude that the data are approximately normally distributed.

Examples :

A random sample of college students aged 18 to 24 years was obtaind, the no. of hours of television watched was recorded:

36.1	30.5	2.9	17.5	21.0
23.5	25.6	16.0	28.9	29.6
7.8	20.4	33.8	36.8	0.0
9.9	25.8	19.5	19.1	18.5
22.9	9.7	39.2	19.0	8.6

Determine if the data come from a normal dist.

Data for a normal prob. plot. :

1)

0.276	0.274	0.275	0.274	0.277
0.273	0.276	0.276	0.279	0.274
0.273	0.277	0.275	0.277	0.277
0.276	0.277	0.278	0.275	0.276

2)

26	24	22	25	23
24	25	23	25	22
21	26	24	23	24
25	24	25	24	25
26	21	22	24	24

3)

24.0	7.9	1.5	0.0	0.3
0.4	8.1	4.3	0.0	0.5
3.6	2.9	0.4	2.6	0.1
16.6	1.4	23.8	25.1	1.6
12.2	14.8	0.4	3.7	4.2

Examples :

1) Steel rods are manufactured with a mean length of 25 cm, and standard deviation of 0.07 cm.

a) What proportion of rods has a length less than 24.9?

b) Any rods that are shorter than 24.85 cm or longer than 25.15 cm are discarded. What proportion of rods will be discarded?

c) Using the results of part (b), if 5000 rods are manufactured in a day, how many should the plant manager expected to discard?

d) If an order comes for 10000 steel rods, how many rods should the plant manager manufacture if the order states that all rods must be between 24.9 cm and 25.1 cm?

2) Ball bearing are manufactured with a mean dia. of 5 mm and stand. dev. of 0.02 mm.

a) What proportion of ball bearings ~~gas~~ has a dia. more than 5.03 mm?

b) Any ball bearing that have a dia. less than 4.95 mm or greater than 5.05 mm are discarded. What proportion of ball bearing will be discarded?

c) Using the results of (b) if 30000 ball bearings are manufactured in a day, how many should the plant manager expected to discard?

d) If any order comes in for 50000 ball beaning, how many bearing should the plant manager manufacture if the order stares that all ball bearing must be between 4.97 and 5.03 mm?

Tutorial Sheet No. (2)

Q 1) Prove the following

A. $$S = \sqrt{\frac{\sum f x^2}{N} - \left(\frac{\sum f x}{N}\right)^2}$$

B. $$S = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2}$$

Where $d_i = x_i - A$ where A is constant

Q 2) For the following data :

Class limits	f
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 - 74	8

Obtain :

A) S , S^2 B) Z C) $\acute{Z}$, S_z