

## Statistics Sample Questions: Set 1

---

### Books Legends:

WM = Probability and Statistics for Engineers and Scientists – Walpole, Myers, Myers and Ye.

SCH = Schaum's Outlines Statistics – Murray R. Spiegel and Larry J. Stephens

Mann = Introductory Statistics – Prem S. Mann (Publisher: John Wiley & Sons, Inc.)

SPG = Advanced Practical Statistics – S. P. Gupta

BS = Business Statistics – S. P. Gupta and M. P. Gupta

### Standard Deviation and Central Values

1. [Ref: SPG 163/ Ex 34]

The mean and standard deviation of 100 items are found to be 40 and 10 respectively. If at the time of calculations two items are wrongly taken as 30 and 70 instead of 3 and 27, find the correct mean and standard deviation. [Ans. Mean = 39.3, S.D. = 10.24]

2. [SPG 177/ Ex 49]

The mean and standard deviation calculated from 20 observations are 15 and 10 respectively. If an additional observation 36, left out through oversight, be included in the calculations, find the correct mean and standard deviation. [Ans. Mean = 16.0, S.D. = 10.73]

3. [SCH119/ Ex 5.6]

Calculate the first four moments about the mean from the following data:

Class mark (X):	61	64	67	70	73
Frequency (f):	5	18	42	27	8

[Ans.  $\mu_1 = 0$ ,  $\mu_2 = 8.5275$ ,  $\mu_3 = -2.6932$ ,  $\mu_4 = 199.3759$ ]

SCH 123/ Chap5 : Definition of Moments

4. [SCH 130/ Ex 5.5]

Establish the relation between the first four moments about the mean  $\mu_r$  and the moments about an arbitrary origin  $\mu'_r$  (raw moments).

or, Prove that (a)  $\mu_2 = \mu'_2 - (\mu'_1)^2$ , (b)  $\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3$ , and

(c)  $\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 3(\mu'_1)^4$

5. [SCH 107/ Ex 4.12]

(a) Prove that the standard deviation (i)  $\sigma = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{X^2 - \bar{X}^2}$  and (ii)  $\sigma = \sqrt{d^2 - \bar{d}^2}$

where  $d = X - A$

(b) Use any one of the above formulas to find the standard deviation of the set of numbers

12, 6, 7, 3, 15, 10, 18, 5.

[Ans. Mean = 9.5,  $\overline{X^2} = 114$ , SD = 4.87]

Also practice SCH pages 108-109/ Ex 4.14-4.17]

**6. [SPG 238/ Ex 24]**

From the data given bellow calculate Karl Pearson's coefficient of skewness and comment on the result:

Profits (Tk. lakhs):	10-20	20-30	30-40	40-50	50-60
No. of companies:	18	20	30	22	10

**Hints:** Calculate the mean = 33.6, mode = 35.56 and S.D. = 12.33 and then use formula,  $Sk = (\text{mean} - \text{mode})/\text{SD}$ ; [ Ans. -0.159 ].

**7. [BS 217/ Ex 15(i), NewBS 186/Ex15]**

Calculate coefficient of variation (CV) and Karl Pearson's coefficient of skewness from following data:

Marks less than:	20	40	60	80	100
No. of Students:	18	40	70	90	100

**Hints:** Cumulative frequency is given in the table, so first calculate the frequencies then use the formula for coefficient of variation,  $CV = \frac{\sigma}{\bar{x}} \times 100$  and  $Sk = (\text{mean} - \text{mode})/\text{SD}$ .

[Calculate, mean = 46.4, SD = 24.56, mode = 48.89 , Ans. CV = 52.93, Sk = -0.101. Therefore it is a case of low degree of negatively skewed distribution]

**8. [BS 217/ Ex 15(ii)]**

From the prices of X and Y given below, state which share is more stable in value:

X:	53	54	58	50	61	60
Y:	105	108	104	106	100	102

**9. [SPG 179/ Ex 52]**

From the prices of shares of X and Y below, find out which is more stable in value:

X:	35	54	52	53	56	58	52	50	51	49
Y:	108	107	105	106	100	107	104	103	104	101

[Hints: For Share X,  $CV = (5.92/51) \times 100 = 11.6$ , for Share Y,  $CV = (2.0/105) \times 100 = 1.904$ ; Since CV is less in case of shares Y, hence, they are more stable in value. ]

**10. [SPG 152/ Ex 24]**

Two cricketers scored the following runs in the several innings. Find who is better run-getter and who is more consistent player:

A:	42	17	83	59	72	76	64	45	40	32
B:	28	70	31	0	59	108	82	14	3	95

**Hints** (Calculate): For Cricketer A; mean = 53 runs, SD = 20.09, CV = 37.92

For Cricketer B; mean = 49 runs, SD = 37.06, CV = 75.63

[More correct results: A: SD = 21.18, CV = 39.97 , B: SD = 39.06, CV = 79.72]

Since average is higher in case of cricketer A, hence, cricketer A is better run getter. Also coefficient of variation (CV) is less in case of cricketer A, hence, he is more consistent player.

#### 11. SNB 30/Exr16, 17

Exr16. Compute the arithmetic mean, geometric mean, and harmonic mean of the following set of data.

3, 5, 7, 11, 14, 57

If these data were observations on the time needed to cure a disease, which mean would you think to be most appropriate?

Exr 17. If the weights are 2, 1, 1, 3, 1, and 2 for the numbers 3, 5, 7, 11, 14, and 57 (exercise 16), compute the weighted average and variance.

#### 12. SNB 48/Exr 6, 7

6. The concentration of drug in solution was measured as a function of time:

Time (weeks) :	0	4	8	26	52
Concentration :	100	95	91	68	43

(a) Plot concentration versus time.

(b) Plot log concentration versus time.

7. Plot the following data and label the axes appropriately.  
(See the table from the book)

#### 13. SNB 173/Chap7

### Regression and Correlation

**Simple Regression:** A regression model is a mathematical equation that describes the relationship between two or more variables. A simple regression model includes only two variables: one independent and one dependent. The dependent variable is the one being explained, and the independent variable is the one used to explain the variation in the dependent variable.

**Linear Regression:** A (simple) regression model that gives a straight-line relationship between two variables is called a linear regression model.

Simple linear regression analysis is a statistical technique that defines the functional relationship between two variables, X and Y, by the “best-fitting” straight line. A straight line is described by the equation,  $Y = A + BX$ , where Y is the dependent variable (ordinate), X is the independent variable, and A and B are the Y intercept and slope of the line, respectively

Correlation is a procedure commonly used to characterize quantitatively the relationship between variables. Correlation is related to linear regression, but its application and interpretation are different.

**Formula:**

Regression equation of Y on X is  $Y - \bar{Y} = m(X - \bar{X})$ ; with  $m = \frac{\sum xy}{\sum x^2}$

Regression equation of X on Y is  $X - \bar{X} = m'(Y - \bar{Y})$ ; with  $m' = \frac{\sum xy}{\sum y^2}$

where  $y = Y - \bar{Y}$  and  $x = (X - \bar{X})$

**or,**

Regression equation of y on x is  $y - \bar{y} = m(x - \bar{x})$ ; with  $m = \frac{\sum x'y'}{\sum x'^2}$

Regression equation of x on y is  $x - \bar{x} = m'(y - \bar{y})$ ; with  $m' = \frac{\sum x'y'}{\sum y'^2}$

where  $y' = y - \bar{y}$  and  $x' = (x - \bar{x})$

**14. [SPG 322/ Ex 1]**

Calculate (i) the regression equation of x on y and y on x from the following data and (ii) estimate x when y = 20, (iii) estimate y when x = 30.

x:	10	12	13	17	18
y:	5	6	7	9	13

[Ans. Mean=14, 8;  $x = y + 6$ ,  $y = 0.87x - 4.18$ ; when y = 20,  $x = 20 + 6 = 26$  ]

**15. [SPG 330/ Ex 9]**

From the following data obtain the line of regression of (i) y on x and (ii) x on y and (iii) estimate the value of y when x = 8, 16 and 24.

x:	2	6	8	11	13	13	13	14
y:	8	6	10	12	12	14	14	20

[Ans. mean=10, 12;  $y = 0.8125x + 3.875$ ;  $x = 0.8125y + 0.25$ ;  $y = 10.375, 16.875, 23.375$ , when x = 8, 16 and 24]

**16. [SPG 262/ Ex 1]**

Calculate Karl Pearson's coefficient of correlation between x and y from the following data.

x:	23	27	28	28	29	30	31	33	35	36
y:	18	20	22	27	21	29	27	29	28	29

[ **Hints:**  $r = \frac{\sum x'y'}{\sqrt{\sum x'^2 \sum y'^2}} = \frac{123}{\sqrt{(138)(164)}} = +0.818$  (Ans) ]

SCH 327/Ex13.8 , 329/Ex13.11, 13.13, (read 13.15 , but not for exam), 342/Exr13.32

### 17. SCH 333/ Ex 13.17

Fit a least square line (regression line) of (a) y on x and (b) x on y from the following data and (c) estimate the value of y when x = 80, (d) estimate the value of x when y = 168.

x:	70	63	72	60	66	70	74	65	62	67	65	68
y:	155	150	180	135	156	168	178	160	132	145	139	152

[Ans.  $Y = 3.22X - 60.9$ ;  $X = 31.0 + 0.232Y$ ;  $Y = 196.7$  when  $X = 80$ ;  $X = 70.0$  when  $Y = 168$ ]

Sch345/ Chap14: Correlation  
Formula SCH 350

### 18. [SCH 360/ Ex 14.11]

Find the coefficient of linear correlation between the variables X and Y presented in the following table.

X:	1	3	4	6	8	9	11	14
Y:	1	2	4	4	5	7	8	9

Coefficient of linear correlation is  $r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{84}{\sqrt{(132)(56)}} = 0.977$ ,

here  $y = Y - \bar{Y}$  and  $x = (X - \bar{X})$ .

### 19. [P/ Ex ]

Find (i) the coefficient of linear correlation (r) between the variables x and y from the following data and (ii) calculate the regression of y on x and predict the average value of y when x is 9.

x:	3	6	5	4	4	6	7	5
y:	3	2	3	5	3	6	6	4

[ Ans. Coefficient of linear correlation is  $r = \frac{6}{\sqrt{(12)(16)}} = 0.433$ ,

$y = 0.5x + 1.5$ ; when  $x = 9$ ,  $y = 0.5(9) + 1.5 = 6$ . ]

### 20. [Mann 571/Ex13.1]

Find the (i) least squares regression line for the data on incomes and food expenditures on the seven households given in the following table. Use income as an independent variable and food expenditure as a dependent variable. (The information obtained in hundreds of dollars). (ii) Predict the value of food expenditure if the monthly household income is \$6100.

Income, $x$ :	55	83	38	61	33	49	67
Food Expenditure, $y$ :	14	24	13	16	9	15	17

[ Ans. (i)  $y = 0.2525x + 1.5050$ ; This regression line is called the least squares regression line. It gives the regression of food expenditure on income.

(ii) For monthly income \$6100,  $x = 61$ . So, food expenditure is  $y = 0.2525(61) + 1.5050 = \$16.9075$  hundred = \$1690.75 ]

### 21. [SNB 177/ Table 7.2]

From the following data obtain the line of regression of  $y$  on  $x$  and estimate the value of  $y$  when  $x = 110$  and  $130$ .

Drug potency, $x$ :	60	80	100	120
Assay, $y$ :	63	75	99	116

[Ans.  $Y = 0.977X + 5.9$ ;  $Y = 113.37, 132.91$ , when  $X = 110$  and  $130$ ]

### Formula:

$$\text{Mean, } \bar{x} = \frac{\sum x}{N} = \frac{\sum fx}{N}$$

SCH,4Ed64/ Definitions

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times c ;$$

Where,  $L$  = lower class boundary of the median class (i.e., the class containing the median)

$N$  = number of items in the data (i.e., total frequency)

$cf$  = cumulative frequency of the class next lower to the median class

$f$  = frequency of the median class

$c$  = size of the median class interval

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c ;$$

Where,  $L$  = lower class boundary of the modal class (i.e., the class containing the mode)

$\Delta_1$  = excess of modal frequency over frequency of next-lower class

$\Delta_2$  = excess of modal frequency over frequency of next-higher class

$c$  = size of the modal class interval

### SNB 12/141 Average

$$AM = \text{Mean} = \bar{x}, \quad GM = (x_1 x_2 x_3 \cdots x_N)^{\frac{1}{N}},$$

$$\text{Harmonic mean} = H, \text{ where, } \frac{1}{H} = \frac{\sum \frac{1}{x}}{N} = \frac{\sum f \cdot \frac{1}{x}}{N} = \overline{\left(\frac{1}{x}\right)}$$

### Empirical Relation between Mean, Median and Mode

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

$$= 6 =$$

## 22. SCH,4Ed66/ RMS

The root mean square (RMS), or quadratic mean, of a set of numbers  $X_1, X_2, \dots, X_N$  is sometimes

denoted by  $\sqrt{x^2}$  and is defined by  $\text{RMS} = \sqrt{x^2} = \sqrt{\frac{\sum x^2}{N}}$

This type of average is frequently used in many physical applications.

## 23. SCH,4Ed66/ RMS

**Quartiles, Deciles, and Percentiles**

If a set of data is arranged in order of magnitude, the middle value (or arithmetic mean of the two middle values) that divides the set into two equal parts is the median. By extending this idea, we can think of those values which divide the set into four equal parts. These values, denoted by  $Q_1, Q_2$ , and  $Q_3$ , are called the first, second, and third quartiles, respectively, the value  $Q_2$  being equal to the median.

Similarly, the values that divide the data into 10 equal parts are called deciles and are denoted by  $D_1, D_2, \dots, D_9$ , while the values dividing the data into 100 equal parts are called percentiles and are denoted by  $P_1, P_2, \dots, P_{99}$ .

**The formulas are**

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times c ; Q_2 = L + \frac{\frac{2N}{4} - cf}{f} \times c = \text{Median}, Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times c$$

$$D_3 = L + \frac{\frac{3N}{10} - cf}{f} \times c , D_7 = L + \frac{\frac{7N}{10} - cf}{f} \times c ; P_3 = L + \frac{\frac{3N}{100} - cf}{f} \times c , P_{87} = L + \frac{\frac{87N}{100} - cf}{f} \times c \text{ etc.}$$

**Using formula****24. SCH, P68/3.15, 4Ed73/3.15**

$x$ :	61	64	67	70	73
$f$ :	5	18	42	27	8

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{N} = \frac{6745}{100} = 67.45 \text{ in}$$

SCH, P70/3.20, 4Ed73/3.15

$$\bar{x} = A + \frac{\sum fd}{N} ; d = x - A, \text{ Let } A = 67 \text{ then } \bar{x} = 67 + \frac{45}{100} = 67.45$$

SCH, P71/3.22, 4Ed76/3.22

$x$	$f$	$d = x - A$	$u = (x - A)/c$	$fd$	$fu$
61	5	-6	-2	-30	-10
64	18	-3	-1	-54	-18
67	42	0	0	0	0
70	27	3	1	81	27

$$= 7 =$$

73	8	6	2	48	16
Sum: $N = \sum f = 100$			Sum: $\sum fd = 45$		

$$\bar{x} = A + \frac{\sum fu}{N} \times c ; u = d/c = (x - A)/c, \text{ then } \bar{x} = A + \frac{\sum fu}{N} \times c = 67 + \left(\frac{15}{100}\right)(3) = 67.45$$

## Chap 4 : Standard Deviation and Measures of Dispersion

(SCH, 4Ed Page 95)

### Dispersion, or Variation

The degree to which numerical data tend to spread about an average value is called the dispersion, or variation, of the data. Various measures of this dispersion (or variation) are available, the most common being the range, mean deviation, semi-interquartile range, 10–90 percentile range, and standard deviation.

### The range

The range of a set of numbers is the difference between the largest and smallest numbers in the set.

**Ex.** The range of the set 2, 3, 3, 5, 5, 5, 8, 10, 12 is

$$\text{Range} = 12 - 2 = 10.$$

The range could be indicated as 2 to 12, or 2–12.

### The Mean Deviation

The mean deviation, or average deviation, of a set of  $N$  numbers  $x_1, x_2, \dots, x_N$  is abbreviated MD and is defined by

$$\text{Mean deviation, MD} = \frac{\sum |x - \bar{x}|}{N} = \overline{|x - \bar{x}|} = \overline{|d|}$$

### SCH, 4Ed96/ SD

The standard deviation of a set of  $N$  numbers  $x_1, x_2, \dots, x_N$  is denoted by  $\sigma$ ,  $s$  or SD and is defined

$$\text{SD} = \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}} \quad \text{or,} \quad \text{SD} = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$$

### SNB 16/1.5.1

The sample standard deviation is denoted as s.d. or  $S$ , is calculated as  $s.d = S = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}}$

SCH, P76/3.32, fig 3-4 / Mode

25. SCH, 4Ed86/Ex 3.42

Find the quadratic mean of the numbers 3, 5, 6, 6, 7, 10, and 12.

Solution:

$$= 8 =$$



$$\text{Quadratic mean} = \text{RMS} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\overline{x^2}} = \sqrt{\frac{3^2 + 5^2 + 6^2 + 6^2 + 7^2 + 10^2 + 12^2}{7}} = \sqrt{57} = 7.55$$

## Moments, Skewness and Kurtosis

SCH114, 4Ed123/ Chap 5

### Moments

The  $r$  th moment about the mean  $\bar{x}$  is defined as  $\mu_r = \frac{\sum (x - \bar{x})^r}{N}$  or, as  $\mu_r = \frac{\sum f(x - \bar{x})^r}{N}$

The  $r$  th moment about any origin  $A$  is defined as  $\mu'_r = \frac{\sum (x - A)^r}{N} = \frac{\sum d^r}{N} = \overline{d^r}$ , where  $d = x - A$

### Skewness

SCH116-117, 4Ed 125

$$\text{Sk} = \frac{\text{mean} - \text{mode}}{SD} = \frac{\bar{x} - \text{mode}}{\sigma} = \frac{3(\bar{x} - \text{median})}{\sigma}$$

$$\text{Skewness} = a_3 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\sqrt{\mu_2})^3} = \frac{\mu_3}{\sqrt{\mu_2^3}} ;$$

Another measure of skewness is sometimes given by  $\beta_1 = a_3^2 = \frac{\mu_3^2}{\mu_2^3}$ . For perfectly symmetrical curves, such as the normal curve,  $a_3$  and  $\beta_1$  are zero.

### Kurtosis

$$a_4 = \frac{\mu_4}{(\sqrt{\mu_2})^4} = \frac{\mu_4}{\mu_2^2} ; \text{ Here } \beta_2 = a_4 = 3 \text{ for normal or mesokurtic distribution.}$$

For this reason, the kurtosis is sometimes defined by  $\gamma = \beta_2 - 3$ , which is positive for a leptokurtic distribution, negative for a platykurtic distribution, and zero for the normal distribution.