Statistics

CA Sumit L. Sarda

Definition of statistics

- **Statistics** is the study of the collection, organization, analysis, and interpretation of data
- Used as a plural noun as well as singular noun
- Plural noun- numerical facts or information which can be used to draw some conclusion
- Singular noun- scientific method with the help of which numerical facts can be used to treat these data and draw conclusions from them

Collection of data

- Quantitative information about some particular characteristics under consideration
- Data collected for the first time-primary data
- Already collected and used by a different person-secondary data

Collection of primary data

- Interview method
  - Personal interview
  - Indirect interview
  - Telephone interview
- Mailed questionnaire
- Observation method

Sources of secondary data

- International sources- WHO, ILO, IMF, World bank
- Government sources- CSO, Indian Agri Stats
- Private and quasi governmental- ISI, ICAR, NCERT
- Unpublished-research institutes, researchers

Classification of data

- Process of arranging data into different groups
- Chronological or time series data- successive time points or intervals
- Geographical data- when data is arranged region wise
- Qualitative data- attribute
- Quantitative- variable

Methods of classification

- In terms of nature of variables
- Individual series- items are listed singly
- Discrete series- showing finite values of different items of the series
- Continuous- continuous variables showing range of values of different items of the series

Presentation of data

- Modes of presentation
  - Textual presentation
    - Paragraph or no. of paragraphs
  - Tabulation
    - Help of statistical table
  - Diagrammatic presentation
    - Charts, diagrams and pictures
    - Used for both educated and uneducated section of society

Types of diagrams

- Line diagram or histogram
  - Data that varies over time
- Bar diagram
  - Horizontal bar diagram
    - Qualitative data
  - Vertical bar diagram
    - Quantitative data or time series
- Pie chart
  - Different components to the total

Frequency distribution

- Statistical table that distributes total frequency to a no. of classes
- Statistical class is an interval in a frequency distribution.

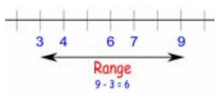
Terms associated with frequency distribution

- Class Limit
  - Minimum value and the maximum value of the class interval
  - Minimum known as LCL
  - Maximum known as UCL
- Class boundary
  - Actual class limit of the class interval

- Mid point
  - Total of two class limits of class intervals divided by two
- Width
  - Difference between UCL and LCL of class interval
- Cumulative frequency
  - No. of observations less than the value
  - Less than the value no. of observations

- Range:
  - The difference between the lowest and highest values.

    In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9, so the range is 9-3 = 6.



- ***Inclusive Class Interval***:
  When the lower and the upper class limit is included, then it is an **inclusive class interval**. For example - 220 - 234, 235 - 249 ..... etc. are inclusive type of class intervals. Usually in the case of discrete variate, inclusive type of class intervals are used.
- ***Exclusive Class Interval***:
  When the lower limit is included, but the upper limit is excluded, then it is an **exclusive class interval**. For example - 150 - 153, 153 - 156.....etc are exclusive type of class intervals. In the class interval 150 - 153, 150 is included but 153 is excluded.
  Usually in the case of continuous variate, exclusive type of class intervals are used.

Graphical presentation

- Histogram or Area diagram
- Frequency polygon
- Ogives or cumulative frequency graphs
  - Used for obtaining quartiles graphically

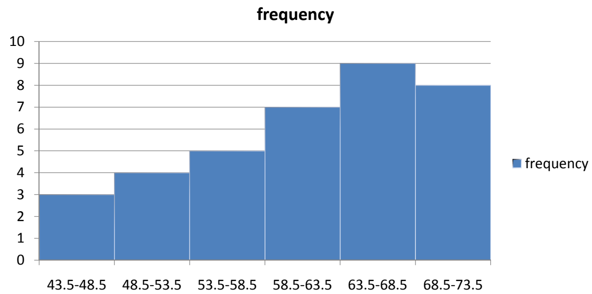
example

- Weight of 36 BBA students
  - 70,73,49,61,61,47,57,50,59,73,68,45,55,65,68,56,68,55,70,70,57,44,69,73,64,49,63,65,70,65,62,64,73,67,60,50
  - Range- 73-44=29kg
  - No. of intervals= range/class length
    = 29/5
    = 6

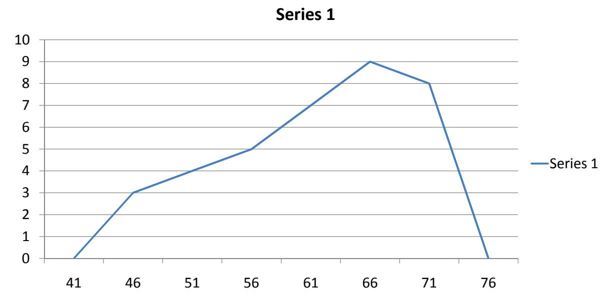
Weight	Tally marks	frequency	Cumulative frequency	Percentage frequency
44-48	III	3	3	8.33
49-53	IIII	4	7	11.11
54-58	IIIII	5	12	13.89
59-63	IIIIIII	7	19	19.45
64-68	IIIIIIIII	9	28	25
69-73	IIIIIIII	8	36	22.22
TOTAL		36		100

Weight	LCL	UCL	LCB	UCB	Mid point
44-48	44	48	43.5	48.5	46
49-53	49	53	48.5	53.5	51
54-58	54	58	53.5	58.5	56
59-63	59	63	58.5	63.5	61
64-68	64	68	63.5	68.5	66
69-73	69	73	68.5	73.5	71

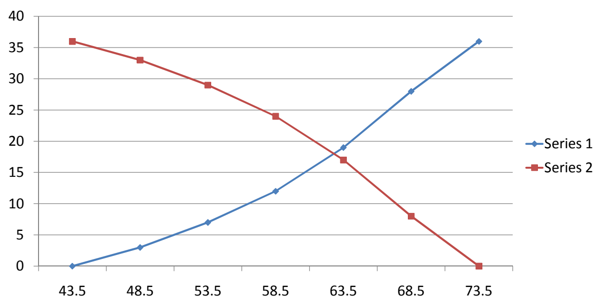
HISTOGRAM



Frequency polygon



ogives



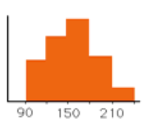
Difference between discrete and continuous variable

- Discrete data have finite values, or buckets. You can count them. Continuous data technically have an infinite number of steps, which form a continuum. The number of questions correct would be discrete--there are a finite and countable number of questions. Time to complete a task is continuous since it could take 178.8977687 seconds.

Difference between bar graph and histogram

Histogram

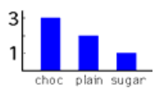
- To present continuous data



- histograms the bars are adjacent

Bar graph

- To present variable data
- E.g. offer to buy donuts for six people, 3 take choco, 2 plain, 1 sugar



- bar graphs are usually separated

Other terms

- Caption- description of column and sub column
- Stub- left part describing rows
- Box head- entire top portion of the table
- Body
- Footnote-place where source of data is written
- Ratio chart
- Hidden trend

- Frequency density
  - Frequency of a class interval divided by class length
- Relative frequency
  - Frequency of a class interval divided by total no. of frequency

Measures of Central Tendency
माप की केंद्रीय प्रवृत्ति

- CA Sumit L. Sarda

26

Definition

- A single expression representing the whole group, is selected which may convey a fairly adequate idea about the whole group
- Tendency of a given set of observations to cluster around a single central value.
- The single expression is known as average
- Averages are the central part of the distribution

27

Types of central tendency

- Arithmetic mean
- Median
- Mode
- Geometric mean
- Harmonic mean

28

Features of a good average

- Should be rigidly defined
- Easy to understand and easy to calculate
- It should be based on all the observations of the data
- Easily subjected to further mathematic calculation
- Least affected by fluctuations of sampling

29

Arithmetic mean

- Average of group of scores refers to mean
- Dividing sum of variables by total no. of observations

$$\overline{X} = \frac{x_1+x_2+x_3+\ldots\ldots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

30

- Find mean for the data
- Wages of 10 workers
- 60,70,80,95,100,110,115,130,140,160

$$= \frac{60+70+80+95+100+110+115+130+140+160}{10}$$

=106

31

Arithmetic mean for discrete series

$$\overline{X} = \frac{f_1x_1+f_2x_2+f_3x_3+\ldots+f_nx_n}{f_1+f_2+f_3+\ldots+f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}$$

32

- Distribution of fifty males given below

Age in yrs	20	21	22	23	24	25
No. of males	1	2	4	5	15	23

Class(x)	20	21	22	23	24	25
Frequency	1	2	4	5	15	23
X*f	20	42	88	115	360	575

33

- Total= 1200
- A.M= 1200/50
  = 24

34

Arithmetic mean of continuous series

$$\overline{X} = A + \frac{\sum fd}{N} \times C$$

35

- Mean weight of group of students

Weight	no.	mid point	f*m.p
44-48	3	46	138
49-53	4	51	204
54-58	5	56	280
59-63	7	61	427
64-68	9	66	594
69-73	8	71	568

36

- Total= 2211
- A.M.=2211/36=61.42kg

37

Alternate method

- Assumed mean= 56
- Thus above can be re written as

46*3	(56-10)*3	56*3-10*3
51*4	(56-5)*4	56*4-5*4
56*5	56*5	56*5
61*7	(56+5)*7	56*7+5*7
66*9	(56+10)*9	56*9+10*9
71*8	(56*15)*8	56*8+15*8

38

56*3	-10*3	=56(3+4+5+7+9+8)/36
56*4	-5*4	=56
56*5	0	
56*7	+5*7	
56*9	+10*9	
56*8	+15*8	

39

Weight	no.	mid point	d	f*d
44-48	3	46	-10	-30
49-53	4	51	-5	-20
54-58	5	56	0	0
59-63	7	61	5	35
64-68	9	66	10	90
69-73	8	71	15	120

40

- Total is 195
- A.M.=56+195/36
  =56+5.42
  =61.42

41

Arithmetic mean of less than series

Marks less than/up to	frequency
10	10
20	30
30	60
40	110
50	150
60	180

42

Marks	students	mid point
0-10	10	5
10-20	20	15
20-30	30	25
30-40	50	35
40-50	40	45
50-60	30	55

43

M.p	f	d	f*d
5	10	-3	-30
15	20	-2	-40
25	30	-1	-30
35	50	0	0
45	40	1	40
55	30	2	60

Total=0
A.M.=35+0*10/180
=35

44

A.M. for more than series

Marks more than	frequency
0	180
10	170
20	150
30	120
40	70
50	30
60	0

45

A.M. for unequal intervals

Marks	no. of students
0-10	10
10-30	60
30-40	50
40-50	40
50-60	20

46

- Since class intervals are unequal, frequencies have been distributed to make the class intervals equal on the assumption that they are equally distributed throughout the class

47

Marks	f	m.p.	D	f*d
0-10	10	5	-3	-30
10-20	30	15	-2	-60
20-30	30	25	-1	-30
30-40	50	35	0	0
40-50	40	45	1	40
50-60	20	55	2	40

48

- Total = (-40)
- A.M.= 35+ (-40)*10/180
  = 35-2.22
  = 32.78

Combined arithmetic mean

- Two groups containing n1 and n2 observations x1 and x2 are their respective arithmetic mean, then combined mean is
- (n1*x1+n2*x2)/(n1+n2)

- Average wage of 10 workers in factory A is 30 and average wage of 20 workers in factory B is 15.
- Combined A.M. will be
  =(10*30+20*15)/(10+20)
  =(300+300)/30
  =20

Correction of incorrect mean

- Correct mean=incorrect sum of product(x)-incorrect value + correct value/ frequency
- =(sum of x- In. V+ C. V)/f

- Average height of group of 40 students is 155cms. It was discovered that height of one student was recorded incorrectly as 157 instead of 137. Calculate average height
- Correct height= (40*155-157+137)/40
  =(6200-20)/40
  =154.5

Properties of A.M.

- Sum of product of deviations from A.M. will always be zero $\sum(x-\overline{x})=0$
- Sum of product of squared deviations of items form A.M. is less than equal to zero
- $\sum(x-\overline{x})^2 \leq 0$
- If each value of x is increased or decreased by some constant C, arithmetic mean also increased or decreased by C.

- Similarly when value of variable X are multiplied or divided by constant k, arithmetic mean can also be multiplied or divided by k

55

Limitations of A.M.

- Affected by extreme values
- Not useful for studying qualitative phenomenon

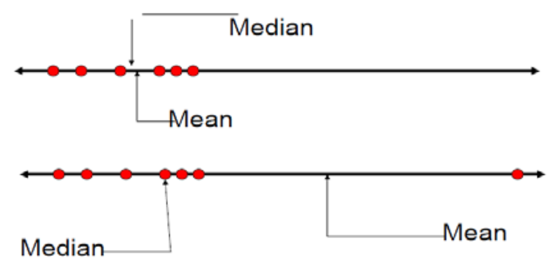
56

Median

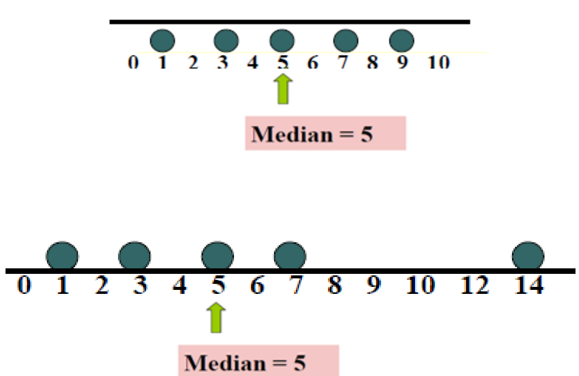
- Middle score of the distribution when all the scores have been ranked
- Value of a variable which divides a series into 2 equal parts so half of the observations have value greater than and other half have values less than the median value
- If there are even no. of scores the median is the average of the two middle scores.
- Also called as **positional average** because it refers to place of a value in the series irrespective of all the other variables

57

- In an ordered array, median is the middle number



58



59

Formula

- Odd= value of (n+1)/2th item
- Even= average of $\frac{N}{2}$ and $\frac{N}{2}+1$

- Find median of raw data
- 25,55,5,45,15,35

60

- Determine the median for the following data

Value	5	7	9	11	13	15
frequency	1	2	9	13	15	8

61

Median for continuous series

$$M = L + \left(\frac{\frac{N}{2} - m}{f} \right) \times c$$

- L= lower limit of median class
- M=cumulative frequency of above median class
- f= frequency of median class
- N= no. of frequencies
- c= no. of class length

62

- From the following data, calculate the median

Marks	0-10	10-20	20-30	30-40	40-50	50-60
students	10	20	30	50	40	30

Marks	students	Cumulative frequency
0-10	10	10
10-20	20	30
20-30	30	60
30-40	50	110
40-50	40	150
50-60	30	180

63

- N/2=180/2=90
- Thus Median= 30+(90-60)*10/50
  36

64

Merits of median

- Not affected by extreme values
- More suitable average for dealing with qualitative data
- It can be determined graphically

65

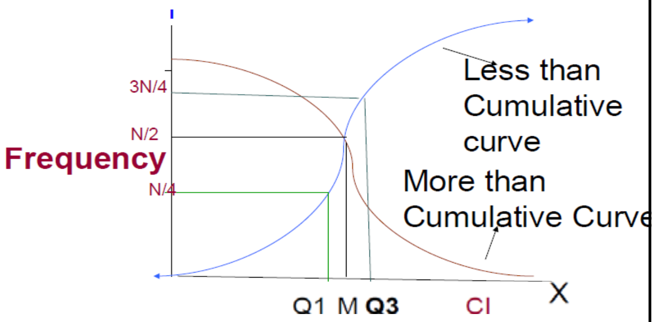
Limitations

- Not based on the items of the series
- Not capable of algebraic treatment. Its formula cannot be extended to calculate combined median of 2 or more related groups

66

Sumit.sarda@hotmail.com
CA. Sumit L. Sarda
+918600364185

Median by graph



67

Quartiles

- Values of variates that divides the series or the distribution into four equal parts are known as quartiles
- First quartile, known as lower quartile is the value of variate below which 25% of the observations lie
- Second quartile, known as middle quartile or median, below which 50% of the observations lie

68

- Third quartile also known as upper quartile, 75% of the observations lie below them

$$Q_1 = Size\frac{N+1}{4}th \text{ Item}$$

$$Q_3 = Size\frac{3(N+1)}{4}th \text{ Item}$$

69

Octiles

- Divides series into 8 equal parts-12.5% each
- Since 7 points are required to divide the data into eight equal parts, we have 7 octiles

$$O_j = Size\frac{j(N+1)}{8}th \text{ Item}$$

$$O_4 = Size\frac{4(N+1)}{8}th \text{ Item}$$

70

- Deciles divides into 10 equal parts-10% each
- Percentile divides into 100 equal parts
  - 99 percentiles exist

71

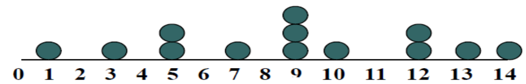
Relationship between partition values

- Q1=O2=P25
- Q2=
- Q3=

72

Mode

- That value which occurs greatest number of times, i.e. with the greatest frequency
- Not effected by extreme values
- Used either for numerical or categorical data
- There may be no mode or several mode
- Most frequent score in distribution



- Mode= 9

- No mode

- A distribution where a single score is most frequent has one mode and is uni modal
- A distribution that consists of only one of each score, the distribution is bi modal if two scores tie or multimodal if more than two scores tie
- The idea of mode is very useful in expressing most common size of readymade garments or shoes etc.

- Mode is calculated in case of discrete and individual series by inspection
- Calculate the mode from the following data of marks obtained by 10 students
- 20,30,31,32,25,25,30,31,30,32

- Mode=30

- Calculate the mode for the following data

Collar size of shirt	35	36	37	39	40	41	43
Persons buying shirt	10	20	30	50	20	12	10

Mode for continuous series

$$Z = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times c$$

- Where
- Z=mode, L= lower limit of the mode class
- f_0= frequency of the pre modal class
- f_1= frequency of the modal class
- f_2=frequency of post modal class
- C= class interval of modal class

- Following is the age distribution of the employees of the firm
- Age no. of persons
- 20-25 5
- 25-30 7
- 30-35 8
- 35-40 18
- 40-45 25
- 45-50 12
- 50-55 7
- 55-60 5

79

- L= 40
- f0=18
- f1=25
- f2=12
- C=5

- Mode=40+(25-18)*5/(2*25-12-18)
  =41.75

80

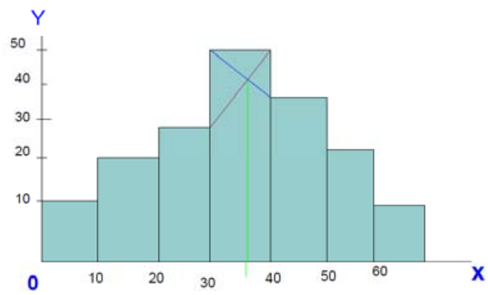
- From the following data, calculate mode

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	10	20	30	50	40	30

- Here L=30, f0=30,f1=50, f2=40, C=10
- Mode=36.667

81

Calculating mode graphically



82

Relationship between mean, median and mode

- Distance between mean and median is about one third of distance between the mean and the mode
- Karl Pearson has expressed the relationship as follows
- Mean-Mode=(mean-median)/3
- Solving we get
- Mode= 3Median-2 Mean

83

Example

- For a monetary skewed distribution of marks in statistics for a group of 200 students, the mean and median mark were found as 55.6 and 52.4. what is the modal mark
- Mode= 3median-2mean
  = 3*52.4+2*55.6
  = 46

84

- If Y= 2+1.5X and mode of X is 15, what is the mode of Y
- Y= 2+1.5*15=24.5

85

Merits of mode

- Mode is the only suitable average
- E.g.. Garment, shoes etc.
- Not affected by extreme values
- Values can be defined graphically

86

Limitations of mode

- In case of bimodal/multimodal series, mode cannot be determined
- Not capable of further algebraic treatment, combined mode of two or more series cannot be determined
- Not based on all items in the series
- Value is significantly affected by size of the class intervals

87

Geometric mean

- Defined as nth root of product of observations

$$\overline{x_G} = \sqrt[n]{x_1 x_2 \cdots x_i \cdots x_n}$$
$$= \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

88

- Take the logarithm of each item of variable and obtain their total
- i.e. ∑log X
- Calculate G.M. as follows

$$G.M = Anti\log\left(\frac{\sum \log X}{n}\right)$$

89

- From the following data, find the geometric mean

Roll no.	1	2	3	4	5	6
Marks	5	15	25	35	45	55

- Alternate 1:
- (5*15*25*35*45*55)^1/6
- (162421875)^1/6
- 23.3582

90

- Alternate 2
- Antilog (∑log X/n)

X	5	15	25	35	45	55
Log X	0.6990	1.17610	1.3979	1.5441	1.6532	1.7404

- Antilog(8.2107/6)
- Antilog(1.3685)
- 23.36

- Average rate of increase in population which in the first decade has increased by 10%, in the second decade by 20% and third decade by 30%

decade	% rise	Population(%age)	Log X
1	10	110	2.0414
2	20	120	2.0712
3	30	130	2.1139

- Antilog(log6.2345/3)
- Antilog(2.0782)
- 119.8

- Alternate 2
- 100*1.1*1.2*1.3=100(X)^3
- X^3=171.6/100
- X=(1.716)^1/3
- X=1.198

Geometric mean for discrete series

- G.M= $(x1^{f1} * x2^{f2} * \ldots\ldots xn^{fn})^{1/n}$
- Alternatively take the logarithm of each item of variable and multiply with the respective frequencies obtain their total
- i.e. ∑f*log X
- Calculate G.M. as

$$G.M = Anti\log\left(\frac{\sum f.\log X}{N}\right)$$

- Find the geometric mean for following distribution

X	2	4	8	16
f	2	3	3	2

- G.M.=$(2^2 * 4^3 * 8^3 * 16^2)^{1/10}$
- =$(2^{2+2*3+3*3+4*2})^{1/10}$
- =$2^{2.5}$
- 5.66

- Alternate2

x	2	4	8	16
F	2	3	3	2
Log X	0.3010	.6020	.9030	1.2041
F*log X	.6020	1.806	2.709	2.4082

- G.M.=antilog(∑f*log X/n)
- G.M.=antilog(7.5252/10)
- G.M.=antilog(0.75252)=5.63

Weighted geometric mean

$$G.M = Anti\log\left(\frac{\sum w.\log X}{\sum w}\right)$$

97

Merits of geometric mean

- Based on all items of the series
- It is rigidly defined
- It is capable of algebraic treatment
- Useful for averaging ratios and percentages rates are increased or decreased

98

Limits of geometric mean

- Difficult to understand and calculate
- Cannot be computed when there are both negative and positive values in the series
- Biased for small values as it gives more weight to small values

99

Harmonic mean

- Reciprocal of arithmetic mean of their respective resiprocals
- Symbolically

$$H.M = \frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+\ldots\ldots+\frac{1}{X_n}}$$

- Where X1, X2, X3 refers to value of various series

100

- From the following data, calculate Harmonic mean

Roll no.	1	2	3	4	5	6
marks	5	15	25	34	45	55

Marks	1/marks
5	0.2
15	0.0666
25	0.04
35	0.0286
45	0.0222
55	0.0182

- H.M.=6/0.3576=15.9744

101

Harmonic mean for grouped distribution

$$HM = \frac{\sum w_i}{\sum(\frac{w_i}{X_i})}$$

102

- Find the harmonic mean for the following

X	2	4	8	16
f	2	3	3	2

x	F	f/x
2	2	1
4	3	0.75
8	3	0.375
16	2	0.125

- H.M.=10/2.25=4.44

103

Merits of Harmonic mean

- Based on all the items of the series
- It is rigidly defined
- It is capable for algebric treatment
- Useful for averaging measures of time, speed etc.

104

Limitations of H.M.

- It is difficult to understand and calculate
- Cannot be computed when one or more items are zero
- Gives more weight to smaller values. Hence not suitable for analyzing economic data

105

- Compute AM, GM and HM for the numbers
- 6,8,12,36
- AM=(6+8+12+36)/4=15.5
- G.M.=$(6*8*12*36)^{1/4}$=12
- H.M.=4/[(1/6)+(1/8)+(1/12)+(1/16)]=9.93

106

- Find the arithmetic mean and harmonic mean of first n natural numbers, weights being equal to squares of corresponding numbers

X	1	2	3	N
f	1^2	2^2	3^2	n^2

- A.M.=$(1^3+2^3+3^3+n^3)/(1^2+2^2+3^2+n^2)$
- =$[n^2(n+1)^2/4]/[n(n+1)(2n+1)/6]$
- =3n(n+1)/2(2n+1)

107

- Harmonic mean
- $(1^2+2^2+3^2+n^2)/(1+2+3+n)$
- [n(n+1)(2n+1)/6]/[n(n+1)/2]
- (2n+1)/3

108

- Arithmetic mean and geometric mean of two observations are 5 and 4 respectively, find the two observations
- Let the two observations be a and b
- (a+b)/2=5 thus a+b=10
- G.M.=4, thus a*b=16
- Solving the two, a=2,b=8

109

- Relationship between AM, GM and HM
- **G.M.^{2}=A.M.*H.M.**

110

Measures of dispersion

- CA Sumit L. Sarda

111

Meaning

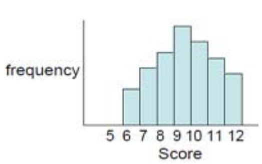
- tells us how compactly the individual values are distributed around the average
- An average such as mean or median only locates the centre of the data
- Average does not tell us anything about the spread of the data
- Dispersion(also known as scatter, spread or variation) measures the items vary from central value
- Measures degree of variation

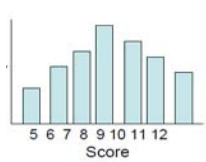
112

Purpose of dispersion

1. Describing distribution of spread

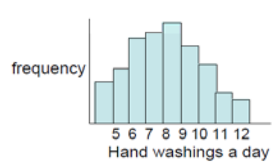
- Two distribution with same mean and different spread

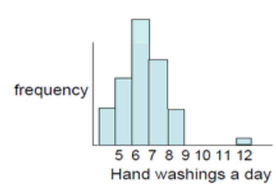




113

2. To compare a given score to the rest of the distribution

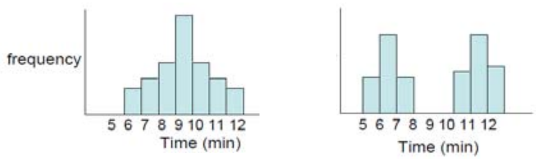




114

3. To take informed decision

- Two routes to campus- same average time



115

Significance of measuring dispersion

- Determine the reliability of the average
- Facilitate comparison
- Facilitate control
- Facilitate the use of other statistical measures

116

Properties of good measure of dispersion

- Simple to understand and easy to calculate
- Rigidly defined
- Based on all items
- Sampling stability
- Not unduly effected by extreme items

117

Absolute measures of dispersion

Based on selected items	Based on all items
Range Inter quartile range	Mean deviation Standard deviation

118

Relative measures of dispersion
-defines variability of data

Based on selected items	Based on all items
Co efficient of Range Co efficient of QD	Co efficient of MD Co efficient of SD Co efficient of Variation

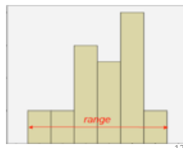
119

- A small value of the measure of dispersion indicates that the data are clustered closely(mean is therefore representative of data)
- A large measure of dispersion indicates that mean is not reliable

120

Range

- describes the spread by giving contrasting the two most extreme scores
- Only takes into account the most extreme values
- May not be the representative of the population



121

Co efficient of range

- Calculated as

$$\frac{L-S}{L+S}$$

- If range of two data are same, more is the co-efficient, greater is the variability in data
- Range is the absolute measure of dispersion, it can not be used for comparison. To make it comparable, we find its coefficient

122

From the following data, calculate range and co efficient of range

X	5	15	25	35	45	55
f	10	20	30	50	40	30

- Range= 55-5=50
- Co efficient of range=
- (55-5)/(55+5)=0.833

X	100	110	120	130	140	150
f	40	35	36	78	92	33

- Range=150-100
- Co efficient of range=
- (150-100)/(100+150)
- =0.2

123

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No .of Students	10	20	30	50	40	30

- Range=60-0=60
- Co efficient of range=(60-0)/(60+0)=1

124

Merits of range

- Easy to understand and calculate
- Does not require special knowledge
- It takes minimum time to calculate

125

Limitations of range

- Does not take into account all items of distribution
- Only two extreme values are taken into consideration
- It is effected by extreme values
- Does not indicate direction of variability
- Not very accurate picture of variability

126

Uses of range

- Facilitates to study quality control
- Variations of price of shares, debentures, bonds and agricultural commodities
- Facilitates weather forecasts

127

Inter quartile range

- The range is based on extreme values. It ignores the deviation in between values. In order to study variation among values, we study Inter-quartile range
- Measures the range of middle values only
- Inter-quartile range is the difference between the third quartile and the first quartile.
- Inter-quartile range = Q_3-Q_1
- Sometimes referred to as quartile deviation or semi-inter quartile range

128

- Quartile Deviation = $\frac{Q_3-Q_1}{2}$
- Coefficient of Q.D. = $\frac{Q3-Q1}{Q3+Q1}$

129

Example

- Number of complaints received by a manager of a super market was recorded for each of the last 10 working days
- 21,15,18,5,10,17,21,19,25,28

- Sorted data(data array)
- 5,10,15,17,18,19,21,21,25,28

130

- Q1=(N+1)/4=11/4=2.75
  = 2^{nd} item +.75(15-10)
  =10+3.75=13.75
- Q3=3(N+1)/4=33/4=8.25
  =8^{th} item+.25(25-21)
  =21+1=22
- Inter quartile range= 22-13.75=8.25

131

- Quartile deviation= 8.25/2=4.125
- Co efficient of QD=(22-13.75)/(22+13.75)
  =8.25/35.75=0.23

132

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No .of Students	10	20	30	50	40	30

- Q1=N/4=180/4=45th item
- Q3=3N/4=3*180/4=135th item

$$Q_1 = L + \left(\frac{\frac{N}{4} - c.f}{f}\right) \times C$$

$$Q_3 = L + \left(\frac{3.\frac{N}{4} - c.f}{f}\right) \times C$$

133

X	F	c.f.
0-10	10	10
10-20	20	30
20-30	**30**	**60**
30-40	50	110
40-50	**40**	**150**
50-60	30	180

134

- Q1=20+(45-30)*10/30=25
- Q3=40+(135-110)*10/40=46.25
- Inter quartile range=46.25-25=21.25
- Quartile deviation=21.25/2=10.625
- Co efficient of QD=(46.25-25)/(46.25+25)
- =0.2982

135

Merits of QD

- Easy to understand and calculate
- Least effected by extreme values
- Used in open end frequency distribution

136

Limitations of QD

- Not suited to algebraic treatment
- Very much affected by sampling fluctuations
- Not based on all items of the series
- Ignores 50% of the distribution

137

Mean deviation

- Takes into consideration all of the values
- Arithmetic mean of the absolute values of deviation from arithmetic mean
- In other words, it is the average distance of the data set from its mean
- It is the arithmetic average of the deviation of all the values taken from some average value(mean, mode, median) of the series, ignoring signs (+ or -) of the deviations.

138

- The deviations are noted as |d| and sum of deviations as Σ |d| is taken. The sum of deviations is divided by the number of items to find the mean deviation.

$$MD = \frac{\sum |x - \bar{x}|}{n}$$

- X= value of each observation, n is no. of observations
- II= absolute value(signs are disregarded)

139

- Remains unchanged due to change of origin but change in the same ratio due to change in scale
- If y=a+bx, M.D. of y= IbI * (M.D. of x)

140

- Steps:
- Find out Mean or median from which deviations are to be taken out.
- Deviation of different items in the series are taken from mean/median, and signs of + or – of the deviations are ignored.
- Each deviation value is multiplied by the frequency facing it and sum of the multiples is obtained.

141

Example

- Weights of sample of crates containing books of the book store are 103,97,101,106,103
- A.M.=510/5=102
- ∑|x-A.M.|=1+5+1+4+1=12
- M.D.=12/5=2.4
- Typically the weights are 2.4kgs from the mean weight.

142

Frequency distribution Mean deviation

- If the data are in the form of frequency distribution, mean deviation can be calculated using

$$MD = \frac{\sum f\,|x - \bar{x}|}{\sum f}$$

- Steps:
- Continuous series are first converted into discrete series by finding mid values of the class interval.
- The same procedure is used for calculation of mean deviation and its coefficient as in case of discrete series.

143

X	F	Fx	Ix-A.M.I	fIx-A.M.I
0	1	0	2	2
1	9	9	1	9
2	7	14	0	0
3	3	9	1	3
4	4	16	2	8

A.M.=48/24=2
M.D.=(2+9+0+3+8)/24=0.92

144

Mean deviation

Calculation of Mean Deviation from Median

- Individual series
- M.D.=∑|x-Median|/n

- Continuous series
- M.D.=∑f|x-Median|/n

- Co efficient of M.D.=
- M.D./Median

Calculation of Mean Deviation from Mean

- Individual series
- M.D.=∑|x-A.M.|/n

- Continuous series
- M.D.=∑f|x-A.M.|/n

- Co efficient of M.D.=
- M.D./A.M.

145

X	Mid point	F	Fx	c.f.	Ix-A.M.I	FIx-A.M.I	Ix-MedianI	fIx-MedianI
0-10	5	10	50	10	30	300	31	310
10-20	15	20	300	30	20	400	21	420
20-30	25	30	750	60	10	300	11	330
30-40	**35**	**50**	**1750**	**110**	**0**	**0**	**1**	**50**
40-50	45	40	1800	150	10	400	9	360
50-60	55	30	1650	180	20	600	19	570

146

- Median
- N/2=90th item
- =30+(90-30)*10/50
- =30+6=36

- M.D.= ∑f|x-Median|/n
- =2040/180=11.333

- Co efficient of M.D.
- =11.333/36=0.3148

- Mean
- A.M.=6300/180
- =35

- M.D.= ∑f|x-A.M.|/n
- =2000/180=11.11

- Co efficient of M.D.
- =11.11/35=.3174

147

Merits of mean deviation

- Easy to understand
- Based on all items of series
- Less effected by extreme values
- Useful small samples when no detailed analysis is required

148

Limitations of M.D.

- Lacks properties such as + or – signs which are not taken into consideration
- Not suitable for mathematical treatment
- Will not give accurate result when the degree of variability is high

149

Standard deviation

- Most commonly used measure of dispersion
- Similar to mean deviation takes into account value of every observation
- Value of mean deviation and standard deviation is relatively similar
- This concept was introduced by Karl Pearson in 1893.
- It is also called Mean Error or Mean Square Error or Root-Mean Square Deviation.

150

- A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values.
- Standard deviation uses squares of the residuals

151

- Steps
- First find out Mean value of series (X)
- Deviation of each item from mean is determined ie We find out x=X-X
- Each value of deviations is squared.
- The sum total of the square of deviations is obtained, Σx^2
- Σx^2 is divided by the number of items in the series
- Square root of $\frac{\Sigma x^2}{N}$ will be Standard Deviation.

152

- The formula is valid *only* if the N values we began with form the *complete* population. If they instead were a random sample, drawn from some larger, "parent" population, then we should have used *n* − 1 instead of *n* in the denominator of the last formula, and then the quantity thus obtained would have been called the **sample standard deviation**

153

Formula

- Full population
- S.D.= $\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{N}}$
- Or $\sqrt{\frac{n\Sigma x^2-(\Sigma x)^2}{n^2}}$

- Sample of population
- $\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{(N-1)}}$
- Or $\sqrt{\frac{n\Sigma x^2-(\Sigma x)^2}{n(n-1)}}$

154

- From the following data calculate standard deviation
- 5,15,25,35,45,55
- Mean=30

X	X-A.M.	(x-A.M.)²
5	-25	625
15	-15	225
25	-5	25
35	5	25
45	15	225
55	25	625

- S.D.=$(1750/6)^{1/2}$=17.078

155

Frequency distribution SD

$$\sigma=\sqrt{\frac{\sum f.(X-\bar{X})^2}{N}}$$

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

156

- From the following data, calculate Standars deviation

Marks	5	15	25	35	45	55
No. of students	10	20	30	50	40	30

Marks	No. of students	fX	X-35	$F(x-35)^2$	fx^2
5	10	50	-30	9000	250
15	20	300	-20	8000	4500
25	30	750	-10	3000	18750
35	50	1750	0	0	61250
45	40	1800	10	4000	81000
55	30	1650	20	12000	90750

Standard deviation by two formula

- S.D.=$(36000/180)^{1/2}$
- =$200^{1/2}$
- 14.142

- S.D.=$[(256500/180)-(6300/180)^2]^{1/2.}$
- $(1425-1225)^{1/2}$
- 14.142

Continuous series

Marks	0-4	4-8	8-12	12-16
students	4	8	2	1

Marks	Mid point	Students	Fx	X-A.M.	$F(x-A.M.)^2$	fx^2
0-4	2	4	8	-4	64	16
4-8	6	8	48	0	0	288
8-12	10	2	20	4	32	200
12-16	14	1	14	8	64	196

Mean= 90/15=6

Standard deviation by two formula

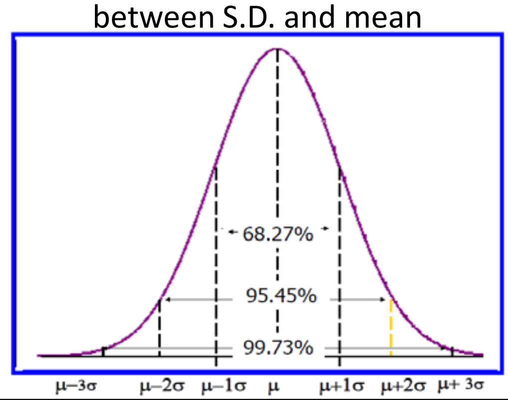
- S.D.=$(160/15)^{1/2}$
- =$10.667^{1/2}$
- 3.265

- S.D.=$[(700/15)-(90/15)^2]^{1/2.}$
- $(46.667-36)^{1/2}$
- 3.265

Properties of Standard Deviation

- Independent of change of origin
- Not independent of change of scale
  - If x and y are related as y=a+bx, then S.D. of **Y=IbI*S.D. of x**
- Fixed relationship of measures of dispersion
  - Q.D.=2S.D./3
  - M.D.=4S.D./5
  - Thus S.D. is never less than QD and MD

Bell shaped curve showing relationship between S.D. and mean



- Minimum sum of squares, sum of squares of deviations of items in series from their arithmetic mean is minimum
- Standard deviation of N natural numbers is

$$= \sqrt{\frac{N^2 - 1}{12}}$$

163

Combined standard deviation

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

- Where N1 is number of items of first group, N2 is number of items in second group

$$d_1 = \overline{X}_1 - \overline{X}_{12} \quad d_2 = \overline{X}_2 - \overline{X}_{12}$$

- Where $\overline{X}_{12}$ is combined mean of 2 groups

164

- A shopkeeper mixes a batch of 200 apples of mean 150g and S.D. 30g with another batch of 300 apples with mean of 100g and S.D. of 20g. Find combined mean and S.D.
- Combined mean=(200*150+300*100)/500

  =120g
- d1=I120-150I=30
- d2=I120-100I=20

165

- Combined S.D.=
- $[200(30^2+30^2)+300(20^2+20^2)/500]^{1/2}$
- $[(200*1800+300*800)/500]^{1/2}$
- $(1200)^{½}$
- 34.64g

166

- Mean and S.D. of marks obtained by two groups of students, consisting of 50 each are given below. Calculate mean and S.D.

Group	Mean	S.D.
1	60	8
2	55	7

167

Merits of standard deviation

- Based on all the items of the distribution
- Mean able to algebraic treatment since actual + or – signs deviations are taken into consideration
- Least affected by sampling fluctuations
- Facilitates calculation of combined standard deviation, co efficient of variation, used to compare variability of distributions

168

- Facilitates other statistical calculations like skewness, correlation
- Provides unit of measurement for the normal distribution

169

Limitations of standard deviations

- Very much affected by extreme values
- Standard deviation cannot be computed for a distribution with open end classes

170

Variance

- Arithmetic mean of squares of deviations of all the items of the distribution from arithmetic mean
- Square of standard deviation
- Smaller the variance, greater the uniformity in population
- Larger the variance, greater the variability

171

Co efficient of variation

- Measure of relative variability
  - Used ot measure change that have taken place in the in a population over time
  - Compare variability of two populations that are expressed in different units of measurement
  - Expressed as a percentage

$$CV = \frac{\sigma}{\overline{X}} \times 100$$

172

- If arithmetic mean and co efficient of variance of x are 10 and 40 respectively, what is the variance of (15-2x)?
- Co efficient of variance=S.D.*100/mean
- Thus S.D.=40*10/100=4
- S.D. of 15-2x=I-2I*4=8
- Thus variance=64

173

Question

- Quaterly profits of two companies are given below, calculate standard deviation and co efficient of variation

Quarters	Company A	Company B
1	9.5	15.5
2	13.7	21.2
3	10.4	23.4
4	8.6	18.8

174

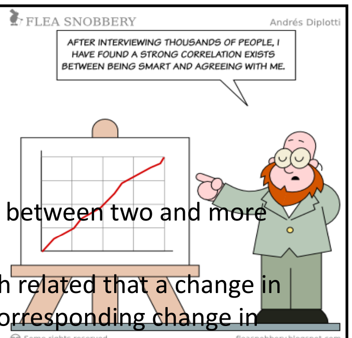
- After settlement of average weekly wage in a factory has increased from 8000 to 12000 and S.D. from 100 to 150. after settlement wage has become higher and more uniform. Do u agree?

175

Correlation

- CA Sumit L. Sarda

176



- Relationship that exists between two and more variables
- If two variables are such related that a change in one variable causes a corresponding change in another, then variables are said to be related

177

Examples

- Relationship between height and weight
- Quantum of rainfall and yield of wheat
- Price and demand of commodity
- Dose of insulin and blood sugar
- is there a relationship between poverty levels and crime
- Attachment level in children and future behavior.
- Are basketball players heights associated with number of points scored?

178

Uses of correlation

- Economic theory and business studies relationship between variables like price and quantity demand
- Helps in deriving precisely the degree and direction of such relationships
- Effect of correlation is to reduce the uncertainty of our prediction
- Analysis on the basis of correlation will be more reliable and near to reality

179

Independent and Dependent Variables

- Independent variable: is a variable that can be controlled or manipulated.
- Dependent variable: is a variable that cannot be controlled or manipulated. Its values are predicted from the independent variable.

180

Example

- Independent variable in this example is the number of hours studied.
- The grade the student receives is a dependent variable.
- The grade student receives depend upon the number of hours he or she will study.
- Are these two variables related?

Student	Hours studied	% Grade
A	6	82
B	2	63
C	1	57
D	5	88
E	3	68
F	2	75

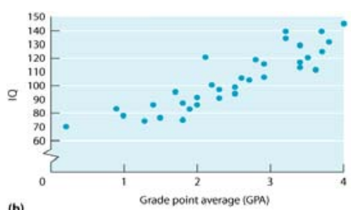
181

Positive correlation

- If both variables vary in the same direction, correlation is said to be positive
- If one variable increases, the other also increases or, if one variable decreases, the other also decreases, then the variables are said to be positive
- E.g.. Horsepower and speed

182

Positive Correlation



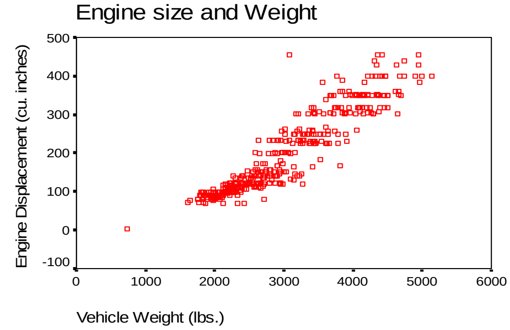
Concurrent Change Same Direction

- The value of the correlation represents the strength of the relationship.
- +1 represents a perfect positive relationship.
- 0.9 is an extremely high correlation, 0.2 isn't as strong.

183

Strong positive correlation:

Engine size and Weight



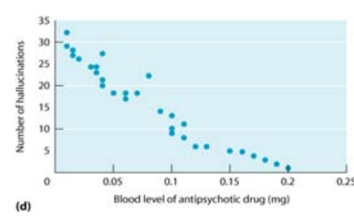
184

Negative correlation

- If both the variables vary in opposite direction, correlation is said to be negative
- If one variable increases, other decreases, of if one variable decrease, the other variable increases, then the two are said to be negative
- Example: Horsepower and miles per gallon

185

Negative Correlation

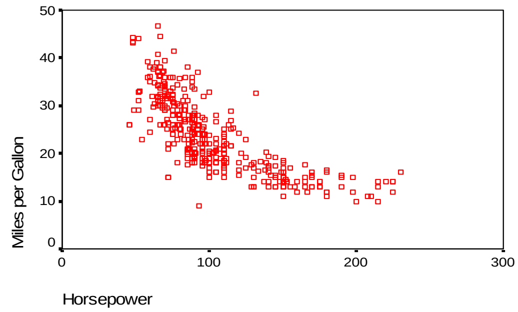


Concurrent Change in Opposite Directions

186

Strong negative correlation:

Horsepower and miles per gallon



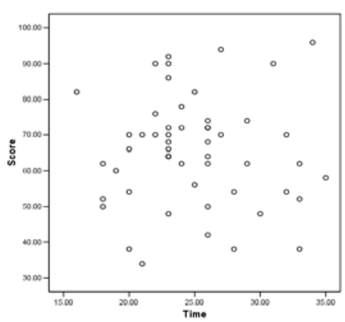
187

Zero correlation

- Zero correlations – as X increases, we have no idea what happens to Y.
  - Values around 0
  - Examples: length of hair and test scores

188

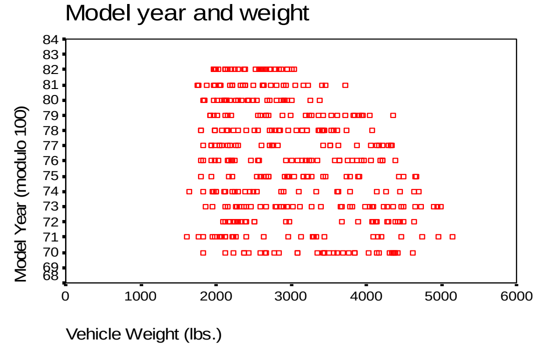
Zero correlation: Exams scores and time to finish exam



189

Very weak correlation

Model year and weight



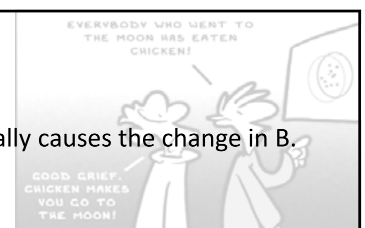
190

Correlation does not mean causation

- A correlation tells you that a relationship exists between 2 variables (aside from the 3rd variable problem), but tell you **absolutely nothing** about cause and effect.
- Ex. Significant correlation between ice cream sales and murder rates – ice cream sales and shark attacks
- The number of cavities in elementary school children and vocabulary size have a strong positive correlation.

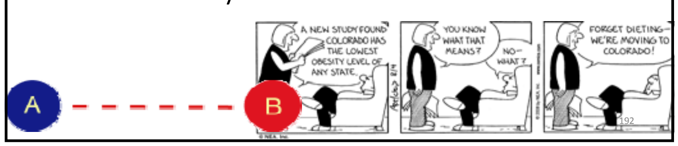
191

CAUSALITY



- When variable A actually causes the change in B.

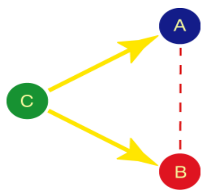
- Variables A and B really do NOT have anything to do with each other but happen to go up or down simultaneously.



192

Common Underlying Cause(s)

- Variable A is correlated with variable B but there is a third factor C (the common underlying cause) that causes the changes in both A and B.



193

Types of correlation

- Simple correlation
- Multiple correlation
  - , which measures the degree to which one variable is correlated with two or more other variables
- Partial correlation
  - which measures the degree of association between two variables when the effects on them of a third variable is removed

194

Methods of studying correlation

- Graphic
  - Scatter diagram
- Algebraic
  - Karl Pearson
  - Rank method
  - Concurrent deviation

195

Scatter diagram method

- Used to demonstrate co relation between two quantitative variables
- The strength of the linear relationship between X and Y is stronger as the swarm of points in the scatter plot more closely approximates a diagonal "line" across the graph

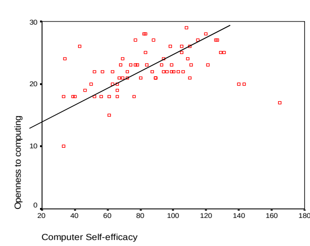
196

- The purpose of the scatter plot is to visualize the relationship between the two variables represented by the horizontal and vertical axes. Note that although the relationship is not perfect, there is a tendency for higher values of openness to computing to be associated with larger values of computer self-efficacy, suggesting that as openness increases, self-efficacy increases. This indicates that there is a positive correlation

197

Drawing A Possible Regression Line

- Let's draw a line through the swarm of points that best "fits" the data set (minimizes the distance between the line and each of the points). This is imposing a linear description of the relationship between the two variables



198

Covariance

- Given a pair of observations (x1,y1) (x2,y2) ......(Xn,Yn) relating to two variables x and y, co variance of x and y is usually represented by Cov(x,y)

$$Cov(X,Y)=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{N}$$

$$=\frac{\sum xy}{N}$$

- When X and Y : cov (x,y) = pos.
- When X and Y : cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

199

Variance vs. Covariance

- Do two variables change together?

Variance:

- Gives information on variability of a single variable.

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$$

Covariance:

- Gives information on the degree to which two variables vary together.
- Note how similar the covariance is to variance: the equation simply multiplies x's error scores by y's error scores as opposed to squaring x's error scores.

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

200

Properties of co variance

- Independent of choice of origin, not independent of choice of scale
- i.e. if two variables x and y are related as y=a+bx and u and v is related to x and y such that x=c+du and y=e+fv, then
- Co variance(u,v)= d*f*co variance(x,y)

201

- From the following data, calculate co variance
- X 1 2 3 4 5
- Y 10 20 30 50 40

X	X-A.M.	Y	Y-A.M.	(X-AM)*(Y-A.M.)
1	-2	10	-20	40
2	-1	20	-10	10
3	0	30	0	0
4	1	50	20	20
5	2	40	10	20
=15	=0	=150	=0	=90

- A.M. of X=15/5=3 A.M. of Y=150/5=30
- Cov(x,y)=90/5=18

202

Limitation of covariance

- One limitation of the covariance is that the size of the covariance depends on the variability of the variables.
- As a consequence, it can be difficult to evaluate the magnitude of the co variation between two variables.
  - If the amount of variability is small, then the highest possible value of the covariance will also be small. If there is a large amount of variability, the maximum covariance can be large.
- Ideally, we would like to evaluate the magnitude of the covariance relative to maximum possible covariance

203

- Let's first note that, of all the variables a variable may co vary with, it will co vary with itself most strongly
- Thus, if we were to divide the covariance of a variable with itself by the variance of the variable, we would obtain a value of 1. This will give us a standard for evaluating the magnitude of the covariance.

$$\frac{\sum(X-M_X)(X-M_X)}{Ns_X s_X}$$

204

- However, we are interested in evaluating the covariance of a variable with another variable (not with itself), so we must derive a maximum possible covariance for these situations too.
- By extension, the covariance between two variables cannot be any greater than the product of the SD's for the two variables.
- Thus, if we divide by $s_x s_y$, we can evaluate the magnitude of the covariance relative to 1.

$$\frac{\sum (X - M_X)(Y - M_Y)}{N s_X s_Y}$$

205

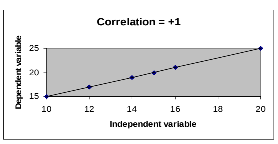
Correlation coefficient

- A correlation coefficient indicates the extent to which two variables are related.
- It can range from -1.0 to +1.0
- A positive correlation coefficient indicates a positive relationship, a negative coefficient indicates an inverse relationship
- Correlation CANNOT be equated with causality.
- Ideally, we would like to evaluate the magnitude of the covariance relative to maximum possible covariance

206

Range of correlation coefficient

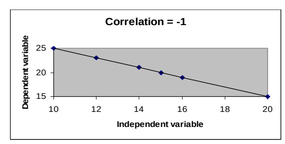
- In case of exact positive linear relationship the value of r is +1.
- In case of a strong positive linear relationship, the value of *r* will be close to + 1.



207

Range of correlation coefficient

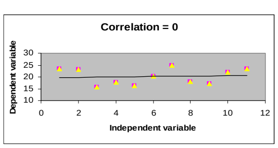
- In case of exact negative linear relationship the value of *r* is –1.
- In case of a strong negative linear relationship, the value of *r* will be close to – 1.



208

Range of correlation coefficient

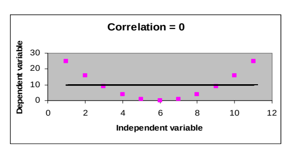
In case of a weak relationship the value of *r* will be close to 0.



209

Range of correlation coefficient

In case of nonlinear relationship the value of *r* will be close to 0.



210

Correlation Coefficient Interpretation

Coefficient Range	Strength of Relationship
0.00 - 0.25	Very Low
0.25 - 0.50	Low
0.50 - 0.75	Moderate
0.75 - 1	High

211

Karl Pearson 's Correlation

- Most widely used mathematical method for measuring the intensity or the magnitude of linear relationship between two variables was suggested by Karl Pearson's
- Measures strength of linear relationship between two quantitative variables

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

212

Properties of Karl Pearson's Coefficient of Correlation

- Independent of origin and scale
- Independent of units of measurement
- Assumptions
  - Linear relationship between variables
  - Cause and effect relationship
- Correlation coefficient between -1 and +1
- Coefficient of correlation is a geometric mean of two regression coefficient

213

Merits

- Gives direction as well as degree of relationship between variables
- Helps in determining value of dependent variable from known value of independent variable

214

Limitations

- Assumption of linear relationship
- Time consuming
- Affected by extreme values
- Requires careful interpretation

215

- Calculate coefficient of correlation
- X 1 2 3 4 5
- Y 10 20 30 50 40

X	Y	X-A.M.	(X-A.M.)²	Y-A.M.	(Y-A.M.)²	X-A.M)(Y-A.M.)
1	10	-2	4	-20	400	40
2	20	-1	1	-10	100	10
3	30	0	0	0	0	0
4	50	1	1	20	400	20
5	40	2	4	10	100	20
=15	=150	=0	=10	=0	=1000	=90

216

- N=5
- A.M. of X=15/5=3
- A.M. of y=150/5=30
- r=90/(10*1000)$^{1/2}$=90/100=+0.9
- r=0.9, there exists a high degree of positive correlation

217

Alternate formula

$$r = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sqrt{\sum(X_i - \bar{X})^2 \times \sum(Y_i - \bar{Y})^2}}$$

$$r = \frac{\sum f d_x d_y - \frac{(\sum f d_x)(\sum f d_y)}{N}}{\sqrt{\sum f d_x^2 - \frac{(\sum f dx)^2}{N}}\sqrt{\sum f d_y^2 - \frac{(\sum f dx)^2}{N}}}$$

218

Standard error

- Used for ascertaining probable error of coefficient of correlation

$$SE = \frac{1-r^2}{\sqrt{N}}$$

- Probable error
  - Amount which is added to or subtracted from values of r which gives upper and lower limit within which coefficient of correlation is expected to lie. It is 0.6745 times S.E.

$$ProbableError = 0.6745.\frac{1-r^2}{\sqrt{N}}$$

219

Uses of probable error

- Used for determining reliability of r in so far as it depends on random sampling

Case	Interpretation
1.If |r |< 6 PE	The value of r is not at all significant. There is no evidence of correlation.
2. 1.If |r | >6 PE	The value of r is significant. There is evidence of correlation

220

Example

- If r= -0.8 and N=36, calculate SE,PE, limits of population, significance of r
- S.E.=(1-(-0.8)^2)/36$^{1/2}$
- =(1-0.64)/6=0.36/6=0.06
- Probable error=0.6745*0.06=0.04
- Limits of population= -0.8+(-)0.04=-0.84 to -0.76
- Since the value of r more than is 6 times PE, the value of r is significant.

221

Coefficient of determination

- Ratio of explained variance to total variance
- Calculated by squaring coefficient of correlation
- Example:
- If r=0.8, what is the proportion of variable in dependent variable which is explained by the independent variable?
- If r=0.8, r^2=0.64
- That means 64%variation in dependent variable is explained by independent variable

222

Co efficient of non determination

- Ratio of unexplained variance to total variance
- Calculated by subtracting co efficient of determination from 1
- Example: in previous example
- Coefficient of non determination=1-.64=36%
- Means 36% variation in dependent variable not explained by independent variable

223

Spearman's rank correlation

- Uses ranks than actual observations and makes no assumption about the population from which actual observation are drawn

$$r = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

224

- For any set of N paired bi variate ranks, the minimum possible value of-$\sum \mathbf{D}^2$ occurs in the case of perfect positive correlation. In this case, rank 1 for X is paired with rank 1 for Y, rank 2 for X with rank 2 for Y, and so on. Each value of **D** will accordingly be equal to zero, and so too will be the sum of the squared values of **D**.
- Conversely, the maximum possible value of-$\sum \mathbf{D}^2$ occurs in the case of perfect negative correlation. This maximum possible value is in every instance equal to maximum-
- $\sum \mathbf{D}^2 = N(N^2 - 1)/3$

225

- Thus, for N=8 with perfect negative correlation:

item	X	Y	D	D²	
a	1	8	−7	49	
b	2	7	−5	25	$\sum \mathbf{D}^2 = 168$
c	3	6	−3	9	
d	4	5	−1	1	
e	5	4	1	1	$8(8^2-1)/3 = 168$
f	6	3	3	9	
g	7	2	5	25	
h	8	1	7	49	

- The ratio of the observed-$\sum \mathbf{D}^2$ to its maximum possible value will therefore be equal to zero in the case of perfect positive correlation, to +1.0 in the case of perfect negative correlation, and to +.50 in the case of zero correlation.
- Double this ratio, subtract it from 1, and voila! you have a quantity that will be equal to +1.0 in the case of perfect positive correlation, to —1.0 in the case of perfect negative correlation, and to zero in the case of zero correlation.

226

- Two judges in the beauty contest ranked entries as follows
- X 1 2 3 4 5
- Y 5 4 3 2 1

X	Y	d=X-Y	d²
1	5	-4	16
2	4	-2	4
3	3	0	0
4	2	-2	4
5	1	-4	16

- $r=1-6*40/5(5^2-1)= -1$

227

Spearman's rank correlation for repeated marks

- Where m= no. of times the rank are repeated
- N=no. of observations
- r=correlation coefficient

$$r = 1 - \frac{6\left[\sum D^2 + \frac{m^3 - m}{12} + \ldots\right]}{n(n^2 - 1)}$$

228

Features

- Based on ranks rather than actual observations
- Distribution free and non parametric because no strict assumptions are made about the form of population from which sample observations are drawn
- Sum of differences between ranks of variables is zero
- Suitable for qualitative data
- Suitable for abnormal data
- Only method for ranks

229

When to be used

- Distribution is not normal
- Behavior of distribution is not known
- Only qualitative data is given

230

Concurrent deviation method

- Based on direction of change in two paired variables

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

- r_c= coefficient of concurrent deviation
- C= no. of positive signs after multiplying change in direction of change of X series and Y series
- N=no. of pairs of observations computed

231

Limitations

- Does not differentiate between small and big changes
- Works on approximations

232

- Compute concurrent deviation
- X 59 69 39 49 29
- Y 79 69 59 49 39

X	Direction of change	Y	Direction of change	D_x*D_y
59		79		
69	+	69	-	-
39	-	59	-	+
49	+	49	-	-
29	-	39	-	+
	n=4			C=2

- $r=[(2c-n)/n]^{1/2}=[(2*2-4)/4]^{1/2}=0$

233

- Calculate co efficient of concurrent deviation

Supply	65	40	35	75	63	80	35	20	80
Demand	60	55	50	56	30	70	40	35	80

Demand	Cx	Supply	Cy	Cx*Cy
65		60		
40	-	55	-	+
35	-	50	-	+
75	+	56	+	+
63	-	30	-	+
80	+	70	+	+
35	-	40	-	+
20	-	35	-	+
80	+	80	+	+

- n=8, C=8
- r=1

234

- Supply 150 154 160 172 160 165 180
- Price 200 180 170 160 190 180 172

Supply	Cx	Price	Cy	Cx*Cy
150		200		
154	+	180	-	-
160	+	170	-	-
172	+	160	-	-
160	-	190	+	-
165	+	180	-	-
180	+	172	-	-

- $r_c = \underline{+}\sqrt{\underline{+}[(2*0-6)/6]} = -\sqrt{-(-1)} = -1$

235

Regression
पीछे की ओर हटाना

-CA Sumit L. Sarda

236

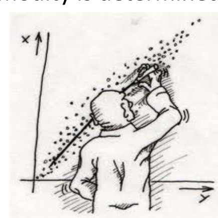
regression

- Measure of average relationship between two or more variables in terms of original units of the data
- Regression analysis is a statistical tool to study the nature and extent of functional relationship between two or more variables and to estimate the unknown values of independent variables

237

Example

- Price of a commodity is determined through its demand



- A line drawn to pass as closer as possible to all the plotted points on a scatter graph is called the line of best fit or the regression line
- Equation of this line is in the form of linear equation

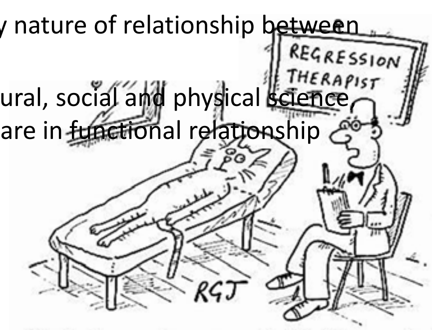
238

Uses of regression analysis

- Regression line facilitates to predict the values of dependent variable from the given value of independent variable
- Cause and effect relationship are indicated from the study of regression analysis
- Useful in economic analysis as regression equation can determine an increase in cost of living index for a particular increase in general price level

239

- Helps in determining the co efficient of correlation as $r^2 = b_{yx} * b_{xy}$
- Enables to study nature of relationship between two variables
- Useful to all natural, social and physical science where the data are in functional relationship



"Under hypnosis you revealed that in your last eight lives you were ... er ... a cat."

240

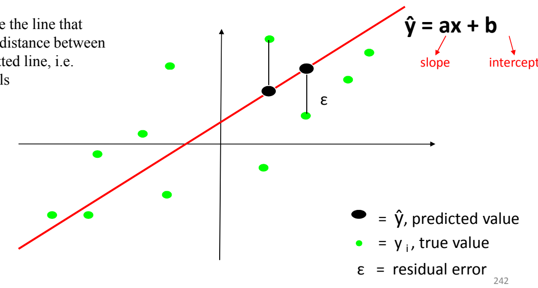
Distinction between correlation and regression

Correlation	Regression
Measures degree and direction of relationship between variables	Measures nature and extent of relationship between two and more variables
Relative measure showing association between variables	Absolute measure relationship
Correlation co efficient is independent of both origin and scale	Regression co efficient is independent of origin but not scale
Correlation co efficient is independent of unit of measurement	Regression co efficient is not independent of unit of measurement
Correlation co efficient lies between -1 and 1	Regression co efficient maybe linear or non linear
It is not forecasting device	It is a forecasting device

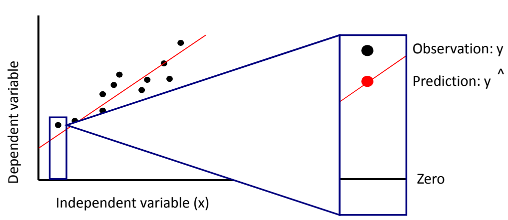
241

Best-fit Line

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x
- This will be the line that minimises distance between data and fitted line, i.e. the residuals



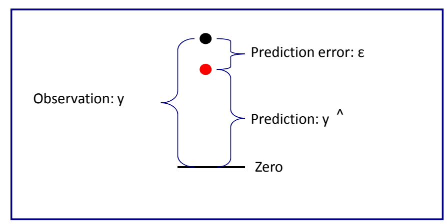
242



The function will make a prediction for each observed data point.

The observation is denoted by y and the prediction is denoted by $\hat{y}$.

243

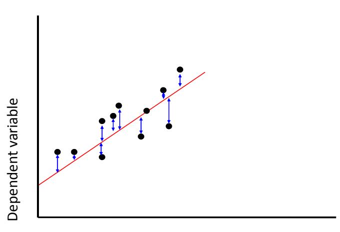


For each observation, the variation can be described as:

$$y = \hat{y} + \varepsilon$$

Actual = Explained + Error

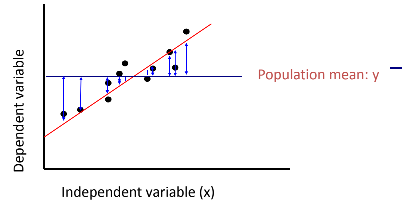
244



A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.

245



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

246

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

$SSR = \sum (\hat{y} - \bar{y})^2$ (measure of explained variation)

$SSE = \sum (y - \hat{y})^2$ (measure of unexplained variation)

$SST = SSR + SSE = \sum (y - \bar{y})^2$ (measure of total variation in y)

247

The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as R^2.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

The value of R^2 can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often referred to as a percentage.

248

Regression lines

- Regression line X on Y

$$X = a + bY$$

- Where X is dependent variable and Y is independent variable
- A=intercept and b=slope

249

Normal equations

- Regression lines X on Y

$$X = a + bY$$

- Then two normal equations are

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

250

Regression lines

- Regression line Y on X

$$Y = a + bX$$

- Where Y is dependent variable and X is independent variable
- A=intercept and b=slope

251

Normal equation

- Regression line Y on X

$$Y = a + bX$$

- Two normal equations are

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

252

- Another way of regression line X on Y

$$X - \overline{X} = b_{xy}\left(Y - \overline{Y}\right)$$

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$$

253

- Another way of regression line Y on X

$$Y - \overline{Y} = byx\left(X - \overline{X}\right)$$

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

254

Regression co efficient

- There are two regression co efficient, b_{yx} and b_{xy}
- The regression co efficient x on y is

$$b_{xy} = r.\frac{\sigma_x}{\sigma_y}$$

- The regression co efficient y on x is

$$b_{yx} = r.\frac{\sigma_y}{\sigma_x}$$

- Where r is correlation co efficient

255

Angles between regression lines

- If r=0
- Regression lines are perpendicular to each other
- If r=+1 or -1
- Regression lines coincide to become identical

256

Advertisement exp.	1	2	3	4	5
Sales	10	20	30	40	50

- Find out two regression equations
- Calculate co efficient of correlation
- Estimate likely sales when advertisement expenditure is Rs.7 lakhs
- What would be the expenditure if the firm wants to attain sales target of Rs.80 lakhs

257

X	Y	X^2	Y^2	XY
1	10	1	100	10
2	20	4	400	40
3	30	9	900	90
4	40	16	1600	160
5	50	25	2500	250

Regression equation of X on Y
X=a+bY
then the normal equations are
$\sum X = Na + b\sum Y$ i.e. $15=5a+150b$
$\sum XY = a\sum Y + b\sum Y^2$ i.e. $550=150a+5500b$

solving we get X=0.01Y

258

- Regression equation of Y on X
  Y=a+bX
  then the normal equations are
  $\sum Y=Na+b\sum X$ i.e.150=5a+15b
  $\sum XY=a\sum X+b\sum X^2$ i.e. 550=15a+55b

  solving we get Y=10X

- Co relation coefficient= $\sum XY/(\sum X^2 * \sum Y^2)^{1/2}$
  $=550/(55*5500)^{1/2}$
  =1

259

- Sales when advertising expenditure is Rs.7 lakhs
- Y=10*7=70lakhs
- Advertising expenditure to obtain sales target of 80lakhs
- X=0.1*80=8lakhs

260

Properties of regression co efficient

- Same sign
- Both Cannot be greater than one
- Arithmetic mean of regression co efficient are greater than correlation co efficient
- R, bxy and byx have same sign
- Correlation coefficient is geometric mean of regression coefficients

261

Independent of origin but not of scale

- If original pair of values x and y are changed to (u,v) such that x=a+pu and y=c+qv

$$b_{uv} = \frac{q}{p} \times b_{xy} \text{ and}$$

$$b_{vu} = \frac{q}{p} \times b_{yx}$$

262

Example

- If the relationship between x and u is u+3x=10 and between y and u is 2y+5v=25 and the regression coefficient of y on x is 0.8, what would be the regression coefficient of v on u

263

- u+3x=10
- Rearranging x=(10/3)-(1/3)u
- 2y+5v=25
- Rearranging y=(25/2)-(5/2)v
- byx=0.8
- bvu= $\frac{-1}{3} \div \frac{-5}{2}$ *0.8 =(2/15)*0.8= **8/75**

264

Calculate byx and bxy

byx

$$b_{yx} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

$$a = \overline{Y} - b\overline{X}$$

bxy

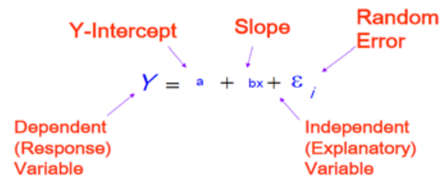
$$b_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}}$$

$$a = \overline{X} - b\overline{Y}$$

265

Population linear regression model

- Relationship between variables is described by a liner function
- Change in independent variable causes change in dependent variable



266

- Using ordinary least squares, we can find values of a and b that minimize the sum of squared residuals

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

267

Example

- In a partially destroyed record, following data are available
- Variance of x=25
- Regression equation of X on y is 5X-Y=22
- Regression equation of Y on X is 64X-45Y=24
- Find
  - Mean values of X and Y
  - Coefficient of correlation between X and Y
  - Standard deviation of Y

268

- Mean values of X and Y lies on regression lines and are obtained by solving following equations
- 5X-Y=22 and 64X-45Y=24
- X=6 and Y=8
- Regression equation y on x is 64X-45Y=24
- i.e. Y= (-24/25)+(64/45)X
- byx=64/45

269

- Regression equation X on Y is 5X-Y=22
- X=(22/5)+(1/5)Y
- bxy=1/5
- r=$\pm$(byx*bxy)$^{1/2}$
- r=+[(4/45)*(1/5)]$^{1/2}$= 8/15
- Since both regression coefficients are +ve, r is taken as +ve

270

Example

- For the variables x and y, the regression equation is given as 7x-3y-18=0 and 4x-y-11=0
- Find arithmetic mean of x and y
- Identify regression equations
- Regression equations
- Assume 7x-3y-18=0 as y on x
- And 4x-y-11=0 as x on y
- y=(-18/3)+(7/3)x
- x=(11/4)+(1/4)y

271

- Mean values of X and Y lies on regression lines and are obtained by solving following equations 7x-3y-18=0 and 4x-y-11=0
- bxy=1/4 and byx=7/3
- $r=[(7/3)*(1/4)]^{1/2}$
- $r=(7/12)^{1/2}$

272

- For the following observations find regression coefficients byx and bxy and hence find correlation coefficient between x and y
- X 4 2 3 4 2
- Y 2 3 2 4 4

273

X	Y	X^2	XY	Y^2		
4	2	16	8	4		
2	3	4	6	9		
3	2	9	6	4		
4	4	16	16	16		
2	4	4	8	16		

Regression equations
$\sum y=na+b\sum x$ and $\sum xy=a\sum x+b\sum x^2$
15=5a+15b and 44=15a+49b
solving we get byx= (-1/4)

similarly $\sum x=na+b\sum y$ and $\sum xy=a\sum y+b\sum y^2$
15=5a+15b and 44=15a+49b
solving we get bxy= (-1/4)
$r=(bxy*byx)^{1/2}$= ¼= 0.25

274

- For 100 students of a class, the regression equation of marks in statistics (X) on marks in commerce(Y) is 3Y-5X+180=0. The mean marks in commerce is 50 and variance in statistics is 4/9 of variance in commerce. Find mean marks in statistics and coefficient of correlation

275

- From the following data estimate the most likely height of brother whose sisters height is 70cms.
- Brothers mean height is 67cms. With S.D. of 3.5cm
- Sisters mean height is 65cms with SD of 2.5cms
- Correlation coefficient is +0.8

276

Probability

CA Sumit L. Sarda

Definition

- **Probability** is ordinarily used to describe an attitude of mind towards some proposition of whose truth we are not certain.
- Probability of a given event is an experiment of likelihood of occurrence of an event
- It is the ratio of number of favorable cases to total no. of equally likely cases-Laplace
- It is denoted by P(A)

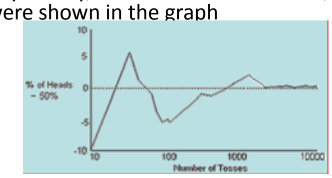
$$P(A) = \frac{\text{number of favourable cases}}{\text{Total number of equally likely cases}}$$

Questions

- Probability of getting an even no. in the dice
- A bag contains 6 black balls and 9 white balls. What is the probability of drawing a black ball
- What is the probability that if a card is drawn at random from an ordinary pack of cards, it is(i) a red card (ii) a club (iii) one of the court cards
- What is the probability that a leap year, selected at random, will have 53 Sundays

10,000 coins tosses

- John Kerrich, a south American mathematician was visiting Copenhagen when World War II broke out. Two days before he was scheduled to fly to England, Germans invaded Denmark. Kerrich spent rest of the war interned at a camp and to pass time he carried out series of experiment in probability theory, in one he tossed coin 10,000 times . His results were shown in the graph



The horizontal axis is on a log scale

Experiments and events

- Simple experiment
  - Some well defined act or process that leads to a single well defined outcome
- **Examples:**

1. Flip a coin
2. Flip a coin 3 times
3. Roll a die

- Typically experiments have following characteristics
  - Set of possible outcomes is known before the experiment
  - Outcome of the experiment is not known beforehand

Properties of probability function

- P(x) is positive for all values of x
- Total of all values of p(x) for different x is always 1

Term	Meaning	Example
Event	Outcome of experiment	Occurrence of 1,2,3,4,5,6 in case of throwing of a dice
Equally likely event	Each outcome has same chance of appearing as any other	Occurrence of 1,2,3,4,5,6 are equally likely
Exhaustive events	When their totality includes all possible outcomes of a random variable	Tossing a coin, possible outcome is head or tail
Simple event	Occurrence of a single event is considered	Tossing of coin, occurrence of head is considered
Compound event	Joint occurrence of two or more events is considered	Tossing of two unbiased coins occurrence of one or more heads
Mutually exclusive events	Cannot occur simultaneously in same trail(occurrence of one prevents occurrence of other)	Tossing a coin, head and tail

Events as set of possibilities

- Symbol S or U are used to represent sample space or set of elementary events for a simple experiment
- Sample space= S or U
  - A coin is tossed and outcome is observed
    S={H,T}
  - An individual is randomly sampled and there sex is noted
    U={M,F}

- Capital letters A , B and so on represent events, each of which is a sub set or sub grouping of elementary events in S
- Suppose we have two events A and B in S. Then we consider and experimental outcome that is a member of both A and B- event A and B has occurred, This is symbolized as
- (A and B)=(A ∩)
- Similarly we can imagine an outcome that is either event A or B. Known as union of A and B
- (A and B)=(A ∪)
- By definition (A ∩) is also(A E∪, but reverse is not true

The word ***and*** in probability means the intersection of two events.

What is the probability that you roll an even number ***and*** a number greater than 3?

E = rolling an even number *F* = rolling a number greater than 3

$$P(E \cap F) = \frac{2}{6} = \frac{1}{3}$$

How can E occur? {2, 4, 6}

How can F occur? {4, 5, 6}

$$E \cap F = \{2,4,6\} \cap \{4,5,6\} = \{4,6\}$$

The word ***or*** in probability means the union of two events.

What is the probability that you roll an even number ***or*** a number greater than 3?

$$P(E \cup F) = \frac{4}{6} = \frac{2}{3} \qquad E \cup F = \{2,4,6\} \cup \{4,5,6\} = \{2,4,5,6\}$$

$\overline{E}$ This is read "*E* complement" and is the set of all elements in the sample space that are not in *E*

Remembering our second property of probability, "The sum of all the probabilities equals 1" we can determine that:

$$P(\overline{E}) + P(E) = 1$$

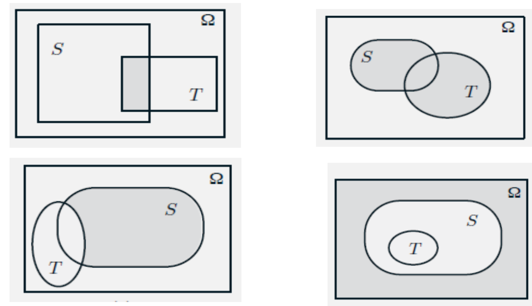
This is more often used in the form

$$P(\overline{E}) = 1 - P(E)$$

If we know the probability of rain is 20% or 0.2 then the probability of the complement (no rain) is 1 - 0.2 = 0.8 or 80%

- A card is drawn from an ordinary pack of playing cards and a person bets that it is a spade or an ace. What are the odds against him?
- Find the chance of getting at least one up in a single throw with three dices
- 4 coins are tossed, find the probability of getting at least one head turn up
- A problem of mathematics is given to three students A,B and C chances of solving are ½,1/3,1/4. what is the probability that the problem is solved
- a man and his wife appear for an interview. Probability of husband selected is 1/7 and wife selected is 1/5. what is the probability that only one of them is selected

- Salesman has 60% chance of making a sale, two customers A and B enter, what is the probability that customer will make a sale



- Probability is a function P that assigns to each event A in the sample space S a number P(A), such that the following are true
- $1 \geq P(A) \geq 0$
- $P(S)=1$
- If there is a set of countable events and of the events are mutually exclusive, then

$$p(A_1 \cup A_2 \cup \ldots \cup A_n) = p(A_1) + p(A_2) + \ldots p(A_n)$$

Rules

- *Commutative Laws:*
  - $A \cup B = B \cup A,\ A \cap B = B \cap A$
- *Associative Laws:*
  - $(A \cup B) \cup C = A \cup (B \cup C)$
  - $(A \cap B) \cap C = A \cap (B \cap C)$.
- *Distributive Laws:*
  - $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
  - $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

292

Independent event

- Two or more events are said to be independent if occurrence of one does not effect the occurrence of another
- Probability of drawing a spade card from a pack of playing cards is 13/52 and that of another spade card after replacing the first one is also 13/52
- Also known as multiplication rule

$$P(A \text{ and } B) = P(A \cap B) = P(A) \cdot P(B)$$

Dependent event

- If occurrence of one effects occurrence of other
- Probability of drawing a spade card from a pack of playing cards is 13/52 and that of another spade card without replacing the first one is also 12/51
- Also known as multiplication rule

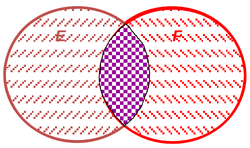
$$P(A \text{ and } B) = P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$$

ADDITION RULE

For any two events *E* and *F*,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Let's look at a Venn Diagram to see why this is true:



If we count *E*

and then count *F*,

we've counted the things in both twice so we subtract off the intersection (things in both).

ADDITION RULE for Mutually Exclusive Events

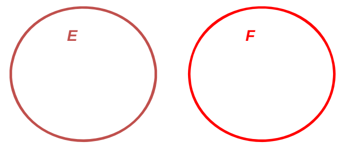
If *E* and *F* are mutually exclusive events,

$$P(E \cup F) = P(E) + P(F)$$

Mutually exclusive means the events are disjoint.

This means $E \cap F = \varnothing$

Let's look at a Venn Diagram to see why this is true:



You can see that since there are not outcomes in common, we won't be counting anything twice.

- A dice is rolled. What is the probability that a number 1 or 6 may appear on the upper face
- Probability of horse A winning the race is 1/5 and Horse B is 1/6, what is the probability that one of the horse will win the race
- Two students solve a problem, probability of x solving a problem is (3/4) and Y solving it is 2/3, what is the probability that the problem is solved
- Probability that a student passes Physics is 2/3 and the probability that he passes both Physics and English is 14/45, probability of passing one test is 4/5. probability of passing English is?
- Two dice are tossed, what is the probability that total is divisible by 3 or 4?

Definition of odds in favor and odds in against

- Find the probability of event A if

(a) Odds in favor are 3:2

(b) Odds against it are 1:4

- Odds in favor of A solving statistics problem are 6 to 9 and against B solving problem are 12 to 10. find probability of problem being solved
- Suppose that it is 11 to 5 against a person who is 38yrs of age living till he is 73 and 5 to 3 against B now 43 living till he is 78 yrs. Find the chance that at least one of the persons will be alive 35yrs from now

Conditional probability

- If two events A and B are said to be dependent where B can occur only when A is known to have occurred. Then the probability of occurrence of such event is called conditional probability

---

- If two events A and B are said to be dependent, then the conditional probability B given A is

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

- And A given B is

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

- Let A and B be the event with P(A)=1/3, P(B)= ¼, P(A⌒B)=1/12
- Find P(A/B), P(B/A), P(B/A^c), P(A⌒B^c)

Bayes Theorem

$$P[B_j \mid A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A \mid B_j]P[B_j]}{\sum_{k=1}^{n} P[A \mid B_k]P[B_k]}$$

Permutation related question

- There are two bags, one contains 4 white balls and 2 black balls, and the other contains 5 white and 4 black balls. Two balls are transferred from one bag to another and then one ball is drawn from second bag
- Four persons are chosen at random from a group containing 3 men, 2 women and four children. What is the probability that 2 of them will be children
- Four cards are drawn from a pack of cards, find probability that 2 are spades and 2 are hearts

- A bag contains 5 white and 8 red balls. Two drawings of 3 balls are made such that balls are replaced before the second trial and(b) balls are not replaced for second trial. Find the probability that first drawing will give 3 white and second three red balls
- A speaks truth in 75% cases and B in 80% cases. What percentage of cases are they likely to contradict each other

Random Variable

- A random variable *x* takes on a defined set of values with different probabilities.
  - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
  - For example, if you poll people about their voting preferences, the percentage of the sample that responds "Yes on Proposition 100" is a also a random variable (the percentage will be slightly differently every time you poll).
- Roughly, <u>probability</u> is how frequently we expect different outcomes to occur if we repeat the experiment over and over ("frequentist" view)

Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - <u>Examples</u>: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - <u>Examples:</u> blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Expected values

- Sum of multiplication probable value of each variable by its corresponding probability
- If the possible outcomes for some experiment are a1, a2, . . ., an, and if the probabilities of these outcomes are p1, p2, . . ., pn, then the expected value is
- E = a1 p1 + a2 p2 + . . . an pn

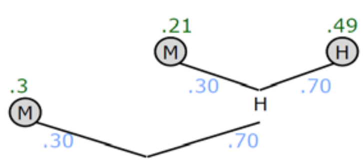
Example

- Suppose Charley goes to 2 movies 10% of all weekends, he goes to 1 movie 40% of the time, and he goes to no movies 50% of the time. What is the expected value for the number of movies he goes to during a weekend?
- Solution: E = 0(.50) + 1(.40) + 2(.10) = .6

Face about independent trials

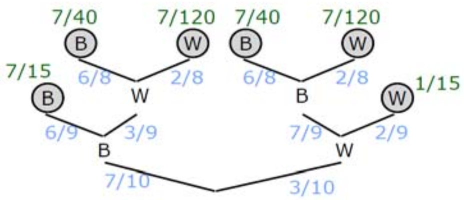
- When *n independent trials are conducted of* an experiment where *p is the probability of* success on each trial, then the expected value for the number of successes in the *n* trials is
- E = *np*
- Example: You throw 5 darts at a target. Each throw has a 45% probability of hitting the target. What is the expected value for the number of hits?
- Solution: *n = 5, p = .45 E = np = 5 * .45 = 2.25*

- Mark is a 70% free throw shooter. If he goes to the foul line to shoot a “one-and-one”,
- (a) what is the expected number of points he will score?
- (b) what is the expected number of shots he will take?



- (a) what is the expected number of points he will score?
- Solution: (a) E = 0(.3) + 1(.21) + 2(.49) = 1.19
- (b) what is the expected number of shots he will take?
- Solution: (b) E = 1(.3) + 2(.21 + .49) = .3 + 1.4 = 1.7

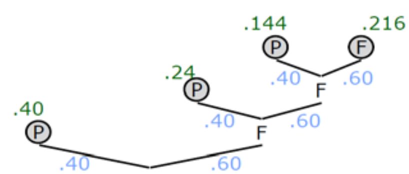
- Bill’s dresser drawer contains 7 blue socks and 3 white ones. He pulls out socks one at a time until he has a matching pair.
- What is the probability he gets a pair of white socks?
- What is the probability he gets a match with the first two socks?
- What is the expected number of socks he will pull from the drawer?



- What is the probability he gets a pair of white socks?
- Solution: 7/120 + 7/120 + 1/15 = 11/60
- What is the probability he gets a match with the first two socks?
- Solution: 7/15 + 1/15 = 8/15
- What is the expected number of socks he will pull from the drawer?
- Solution: E = 2(8/15) + 3(7/15) = 37/15 = 2.4667

- A basketball player hits 70% of his free throws. If his attempts are considered independent trials and he makes 5 attempts in a game, what is the expected value for the number of shots he hits?
- Solution: *E = np = 5 .7 = 3.5*

- Tanya wants to pass a qualifying exam, and she is allowed up to 3 attempts at passing the exam. Each time she takes the exam there is a 40% chance she will pass. What is the expected number of times she will take the exam?



- Solution:
- E = 1(.40) + 2(.24) + 3(.144 + .216)
  = .40 + .48 + 1.08
  = 1.96

- The variance of random variable X, denoted Var(X) or S.D.²
- x ,
- is the long run average of its squared deviation scores, i.e.
- Var(X) = E(X - E(X))²

Properties of Expected value

1. $E(a) = a$

   $$E(a) = \frac{\sum a}{N} = \frac{(a + a + \ldots + a)}{N} = \frac{Na}{N} = a$$

2. $E(X+Y) = E(X) + E(Y)$

   $$E(X+Y) = \frac{\sum (X_i + Y_i)}{N} = \frac{\sum X_i + \sum Y_i}{N} = \frac{\sum X_i}{N} + \frac{\sum Y_i}{N} = E(X) + E(Y)$$

3. $E(XY) = E(X)E(Y)$ if X and Y are independent

   $$E(XY) = \frac{\sum X_i Y_i}{N} = \left(\frac{\sum X_i}{N}\right)\left(\frac{\sum Y_i}{N}\right)$$ (Note: this step will not hold if X and Y are dependent)

   $$= E(X)E(Y)$$

4. $E\left(\frac{X}{Y}\right)=\frac{E(X)}{E(Y)}$ if X and Y are independent

$$E\left(\frac{X}{Y}\right)=\frac{\sum\left(X_i/Y_i\right)}{N}=\frac{\sum X_i/N}{\sum Y_i/N} \text{ (Note: this step will not hold if } X \text{ and } Y \text{ are dependent)}$$

$$=\frac{E(X)}{E(Y)}$$

5. $E(aX)=aE(X)$

$$E(aX)=\frac{\sum aX_i}{N}=\frac{(aX_1+aX_2+\ldots+aX_N)}{N}=\frac{a\sum X_i}{N}=aE(X)$$

6. $E(aX+b)=aE(X)+b$

$$E(aX+b)=\frac{\sum(aX_i+b)}{N}=\frac{\sum aX_i+\sum b}{N}=\frac{(aX_1+\ldots aX_N)}{N}+\frac{Nb}{N}=aE(X)+b$$

Examples

- Probability distribution of a random variable x is given below. Find
- E(x), V(x), E(2x-3)
- X   -2   -1   0   1   2
- P(x)   .2   .1   .3   .3   .1

X	P(x)	Xp(x)	X^2	$X^2p(x)$
-2	.2	-.4	4	.8
-1	.1	-.1	1	.1
0	.3	0	0	0
1	.3	.3	1	.3
2	.1	.2	4	.4
		∑=0		∑=1.6

- E(x)=0   V(x)= $E(x)^2$- $(E(x))^2$ =1.6-0=1.6
- E(2x-3)= 2E(x)-3=-3

- Mean and S.D. of of a random variable is 4 and 4 respectively
- Find $E(x^2)$
- V(x)= $E(x^2)$-$(E(x))^2$=16
- $E(x^2)$=16+25=41