STATISTICS FORMULAE

By CA Anand V Kaku

No. of class interval × class lengths = Range
LCB = LCL –D/2
UCB=UCL +D/2
Mid-point =LCB+UCB/2 or LCL+UCL/2
Frequency density of a class interval = frequency of that class interval/corresponding class length
Relative frequency and percentage frequency of a class interval=class frequency/ total frequency

	ARITHMETIC MEAN	GEOMETRIC MEAN	HARMONIC MEAN	MODE
Discrete Series	$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$ $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	GM= $(x_1 \times x_2 \times x_3 \ldots . . \times x_n)^{\frac{1}{n}}$ Logarithm of G for a set of observations is the AM of the logarithm of the observations; i.e. $\log GM = \frac{\sum \log x}{n}$ $G.M = Antilog\ \frac{\sum \log x}{n}$	$H.M = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$	the value that occurs the maximum number of times
Frequency Distribution	$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{f_1 + f_2 + f_3 + \ldots + f_n}$	GM= $\left({x_1}^{f_1} \times {x_2}^{f_2} \times {x_3}^{f_3} \ldots . . \times {x_n}^{f_n}\right)^{\frac{1}{N}}$	$H.M = \frac{N}{\sum\left(\frac{f_i}{x_i}\right)}$	$Mode\ = l_1 + \left(\frac{f_0 - f_{-1}}{2f_0 - f_1 - f_{-1}}\right) X\ C$ where, l_1 = LCB of the modal class. i.e. the class containing mode. f_0= frequency of the modal class f_{-1} = frequency of the pre – modal class f_1= frequency of the post modal class C = class length of the modal class

STATISTICS FORMULAE

By CA Anand V Kaku

	ARITHMETIC MEAN	GEOMETRIC MEAN	HARMONIC MEAN	MODE

Relationship variables	$\bar{y} = a + b\bar{x}$	if z = xy, then
		GM of z =(GM of x)×(GM of y)
		
		if z = x/y then
		GM of z =(GM of x)/(GM of y)
			-	$y_{mo}=a+bx_{mo}$

Weighted Mean	Weighted A.M= $\frac{\sum x_i w_i}{\sum w_i}$	$Weighted\ G.M = Antilog\ \frac{\sum w_i \log x_i}{\sum w_i}$	$Weighted\ H.M = \frac{\sum w_i}{\sum \left(\frac{w_i}{x_i}\right)}$	-
Combined Mean	$Combined\ A.M\ \bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$	-	$Combined\ H.M = \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}}$	-
Important points	the algebraic sum of deviations of a set of observations from their AM is zero $\sum(x_i - \bar{x}) = 0$	-	-	-
	AM is affected due to a change of origin and/or scale which implies that if the original variable x is changed to another variable y by effecting a change of origin,	-	-	-
Relation among Averages	For Given two positive numbers **(A.M) x (H.M) = (G.M)²**			
	AM≥ GM ≥ HM The equality sign occurs, as we have already seen, when all the observations are equal.			
	Mode = 3 Median – 2 Mean			

STATISTICS FORMULAE

By CA Anand V Kaku

	MEDIAN	QUARTILES (Q_1,Q_2 & Q_3)	DECILES (D_1,D_2 ,D_3,...... D_9)	PERCENTILES (P_1,P_2,P_3,...., P_{99})
Discrete Series /Unclassified Data	Median $=$ Size of $\left(\frac{N+1}{2}\right)^{th}$ *item*	Q_1 quartile is given by the $\frac{1}{4}(N+1)th$ value the Q_n quartile is given by the $\frac{n}{4}(N+1)th$ value	the **D_1** Decile is given by the $\frac{1}{10}(N+1)th$ value **D_n** Decile is given by the $\frac{n}{10}(N+1)th$ value	the P_1 Percentile is given by the $\frac{1}{100}(N+1)th$ value P_n Percentile is given by the $\frac{n}{100}(N+1)th$ value
Group Frequency Distribution	$Median = l_1 + \left(\frac{\frac{N}{2} - N_l}{N_u - N_l}\right) X\, C$ l_1= lower class boundary of the median class i.e. the class containing median. N = total frequency. N_l= less than cumulative frequency corresponding to $l1$. (Pre median class) N_u= less than cumulative frequency corresponding to $l2$. (Post median class) l_2= being the upper class boundary of the median class. C = l_2-l_1= length of the median class. y_{me} = a + b x_{me}	$Q_n = l_1 + \left(\frac{N.p - N_l}{N_u - N_l}\right) X\, C$ l_1= lower class boundary of the Quartile class i.e. the class containing Quartile. N = total frequency. p= $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ for Q_1, Q_2 ,Q_3 respectively N_l= less than cumulative frequency corresponding to $l1$. (Pre Quartile class) N_u= less than cumulative frequency corresponding to $l2$. (Post Quartile class) l_2= being the upper class boundary of the Quartile class. C = l_2-l_1= length of the Quartile class.	$D_n = l_1 + \left(\frac{N.p - N_l}{N_u - N_l}\right) X\, C$ l_1= lower class boundary of the Decile class i.e. the class containing Decile. N = total frequency. p= $\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \ldots, \frac{9}{10}$ for D_1, D_2 ,D_3 ,...., D_9 respectively N_l= less than cumulative frequency corresponding to l_1. (Pre Decile class) N_u= less than cumulative frequency corresponding to l_2. (Post Decile class) l_2= being the upper class boundary of the Decile class. C = l_2-l_1= length of the Decile class.	$P_n = l_1 + \left(\frac{N.p - N_l}{N_u - N_l}\right) X\, C$ l_1= lower class boundary of the Percentile class i.e. the class containing Percentile. N = total frequency. p= $\frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \ldots, \frac{99}{100}$ for P_1, P_2 ,P_3,..., P_{99} respectively N_l= less than cumulative frequency corresponding to l_1 . (Pre Percentile class) N_u= less than cumulative frequency corresponding to l_2. (Post Percentile class) l_2= being the upper class boundary of the Percentile class. C = l_2-l_1= length of the Percentile class.
	$\sum(x_i - A)$ is minimum if we choose A as the median.			

STATISTICS FORMULAE

By CA Anand V Kaku

DISPERSIONS

	Absolute	Relative	If y = a+bx
RANGE (R)	Range= Laregest (L) – Smallest (S)	Co efficient of Range = $\frac{L-S}{L+S} \times 100$	$R_y = \lvert b \rvert \times R_x$
MEAN DEVIATION (M.D) about A	$M.D_A = \frac{1}{n}\sum\lvert x - A\rvert$	Co efficient of M.D from A = $\frac{\text{M.D about A}}{A} \times 100$	$M.D_y = \lvert b \rvert \times M.D_x$
MEAN DEVIATION (M.D) about A.M ($\bar{X}$)	M.D about Mean $= \frac{1}{n}\sum\lvert x_i - \bar{X}\rvert$	Co efficient of M.D from A.M = $\frac{\text{M.D about } \bar{X}}{\bar{X}} \times 100$	$M.D_y = \lvert b \rvert \times M.D_x$
MEAN DEVIATION (M.D) about Median	M.D about Median $= \frac{1}{n}\sum\lvert x_i - Median\rvert$	Co efficient of M.D from Median $= \frac{\text{M.D about A}}{A} \times 100$	$M.D_y = \lvert b \rvert \times M.D_x$
STANDARD DEVIATION (σ)	$\sigma = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n}}$ $\sigma = \sqrt{\frac{\sum x_i^2}{n} - \bar{X}^2}$	Co efficient of Variation = $\frac{\sigma}{\bar{X}} \times 100$	$\sigma_y = \lvert b \rvert \times \sigma_x$
	Standard Deviation for Two numbers, $\sigma = \frac{\lvert a-b\rvert}{2}$ Standard Deviation for First n Natural numbers , $\sigma = \sqrt{\frac{n^2-1}{12}}$	Combined Standard Deviation, $\sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1+n_2}}$ Where $d_1 = \bar{x}_1 - \bar{x}_{12}$, $d_2 = \bar{x}_2 - \bar{x}_{12}$	
QUARTILE DEVIATION (Q_d)	$Q_d = \frac{Q_3 - Q_1}{2}$	Co-efficient of Q.D $= \frac{Q_3-Q_1}{Q_3+Q_1} \times 100$ Or Co-efficient of Q.D = $\frac{Q.D}{Median} \times 100$	
VARIANCE (σ^2)	Variance means Square of Standard Deviation		

Downloaded From www.castudynotes.com

STATISTICS FORMULAE

By CA Anand V Kaku

CORRELATION

KARL PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT	SPEARMAN'S RANK CORRELATION COEFFICIENT	COEFFICIENT OF CONCURRENT DEVIATIONS
$r = r_{xy} = \frac{Covariance\ Between\ x\ \&\ y}{\sigma_x \times \sigma_y}$ Where, $Covariance\ Between\ x\ \&\ y = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$ $or\ = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y}$ $\sigma_x = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n}}$ $\sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}}$ $r = \frac{n\sum x_i y_i - \sum x_i \times \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2] \times [n\sum y_i^2 - (\sum y_i)^2]}}$ $r_{xy} = \frac{bd}{\lvert b\rvert\lvert d\rvert} r_{uv}$ where $u = \frac{x-a}{b}$ $and\ v = \frac{y-c}{d}$	$r_R = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ For tied ranking, $r_R = 1 - \frac{6\left[\sum d_i^2 + \sum\frac{(t^3-t)}{12}\right]}{n(n^2-1)}$	$r_C = \pm\sqrt{\pm\frac{(2c-m)}{m}}$ where c is concurrent deviation, m is one less than number of pairs of observations

STATISTICS FORMULAE

By CA Anand V Kaku

REGRESSION ANALYSIS

	Y depends on X	X depends on Y	
Simple Regression Equation	$y = a + b_{yx}\, x$	$x = a + b_{xy}\, y$	Intersection point of these two lines is $\bar{x}, \bar{y}$
Normal Equations	$\sum y_i = na + b_{yx} \sum x_i$ $\sum x_i y_i = a \sum x_i + b_{yx} \sum x_i^2$	$\sum x_i = na + b_{xy} \sum y_i$ $\sum x_i y_i = a \sum y_i + b_{xy} \sum y_i^2$	
Regression Co efficient	$b_{yx} = \frac{Cov(x,y)}{\sigma_x^2}$ $b_{yx} = \frac{r\,\sigma_y}{\sigma_x}$	$b_{xy} = \frac{Cov(x,y)}{\sigma_y^2}$ $b_{xy} = \frac{r\,\sigma_x}{\sigma_y}$	
	$b_{yx} = \frac{n\sum x_i y_i - \sum x_i \times \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$	$b_{xy} = \frac{n\sum x_i y_i - \sum x_i \times \sum y_i}{n\sum y_i^2 - (\sum y_i)^2}$	

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

$$b_{yx} = \frac{q}{p} \times b_{vu} \text{ where } u = \frac{x-a}{p} \text{ and } v = \frac{y-c}{q}$$

$$Coefficient\ of\ Determination = r^2$$

$$Coefficient\ of\ Non-Determination = 1 - r^2$$

STATISTICS FORMULAE

By CA Anand V Kaku

NAME	CONDITION	PROBABILITY MASS FUNCTION	Notation	MEAN	VARINACE	MODE	Remarks
Binomial Distribution	Trials are independent and each trail has only two outcomes Success & Failure.	$P(X=x) = {}^nC_x\, p^x\, q^{n-x}$ For x = 0,1,2,..,n	X ~ B (n,p)	$\mu=np$	$\sigma^2=npq$	If (n+1)p is integer then Mode, $\mu_0 = (n+1)p-1$ If (n+1)p is non-integer then Mode, μ_0 =Highest Integer in (n+1)p	p+q=1
Poisson Distribution	Trials are independent and probability of occurrence is very small in given time.	$P(X = x) = \frac{e^{-m}.m^x}{x!}$ For x = 0,1,2,..,n	X ~ P (m)	$\mu=m$	$\sigma^2=m$	If m is Integer Mode, $\mu_0 = m-1$ If m is non- integer Mode, μ_0 = Highest Integer in m	e = 2.71828
Normal or Gaussian Distribution	When distribution is symmetric	$P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\bar{x}-\mu)^2/2\sigma^2}$ For $-\infty < x < +\infty$	$X \sim N(\mu,\sigma^2)$ $Z = \frac{x-\mu}{\sigma}$	Mean = Median = Mode = μ	σ^2	μ	**Mean Deviation** = 0.8 σ **First Quartile** = μ - 0.675σ **Third Quartile** = μ + 0.675σ **Quartile Deviation**=0.675σ **Point of Inflexion** x=μ - σ and x=μ + σ