VARIANCE

$$\sigma^2_{population} = \frac{\sum (x-\mu)^2}{N} \quad S^2_{sample} = \frac{\sum (x-\bar{x})^2}{n-1}$$

Sum of squared deviations of every observation from its mean over total number of observations. Units are square units of the original variable.

Computational Formulae :

$$\sigma^2_{population} = \frac{\sum_1^n (x_i^2) - N\mu^2}{N} \quad S^2_{sample} = \frac{\sum_1^n (x_i^2) - n\bar{x}^2}{n-1}$$

$$\begin{aligned} numerator &= \sum (x-\mu)^2 \\ &= \sum (x^2 - 2x\mu + \mu^2) \\ &= \sum (x^2) - 2\mu \sum x + \sum \mu^2 \\ &= \sum (x^2) - 2\mu \,.\, n \,.\, \mu + n\mu^2 \\ &= \sum_1^n (x_i^2) - n\mu^2 \end{aligned}$$

<= useful for manual calculations

STANDARD DEVIATION

Square root of Variance. Units are associated with numerical measures. σ or s

PERCENTILES, QUARTILES, IQR,Box Plot

100p percent of data $\leq$ percentile $\leq$ 100(1-p) percent of data.
Compute : Arrange ascending, decimal np? next pos : integer np? avg of value at np&np+1.
Percentile need not be part of data set. Q1 = 25th p, Q2= median = 50th p, Q3= 75th p.
5 number summary : Min,Q1,Q2,Q3,Max
IQR = Q3-Q1 ; Range = Max-Min

OUTLIERS

$$outlier < Q_1 - 1.5 * IQR; outlier > Q_3 + 1.5 * IQR$$

TWO WAY CONTINGENCY TABLE

Association between 2 categorical variables - relative frequencies. Bivariate Categorical Data. Row Total and Column Total. Row Relative Frequencies (cell frequency/row total) & Column Relative Frequencies (cell frequency/column total). If the row or column relative frequencies are the same for all rows/columns then the 2 variables are not associated.

STACKED/SEGMENTED BAR CHART

Represents counts of a particular category and segments representing frequency of category. 100 percent stacked bar chart is useful for point to whole relationships. (e.g. Gender vs Phone ownership)

SCATTER PLOT

y-axis : response variable ; x-axis : explanatory variable
e.g. x-axis : age and y-axis : height

COVARIANCE

$$cov(x,y)_{population} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum_1^n (x_i \,.\, y_i) - \frac{\sum_1^n x_i \,.\, \sum_1^n y_i}{N}}{N}$$

$$\begin{aligned}
numerator &= \sum_1^n (x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum_1^n (x_i \,.\, y_i - x_i \,.\, \bar{y} - \bar{x} \,.\, y_i + \bar{x} \,.\, \bar{y}) \\
&= \sum_1^n x_i \,.\, y_i - \sum_1^n x_i \,.\, \bar{y} - \sum_1^n \bar{x} \,.\, y_i + \sum_1^n \bar{x} \,.\, \bar{y} \\
&= \sum_1^n x_i \,.\, y_i - \bar{y} \,.\, \sum_1^n x_i - \bar{x} \,.\, \sum_1^n y_i + \sum_1^n \bar{x} \,.\, \bar{y} \\
&= \sum_1^n x_i \,.\, y_i - \bar{y} \,.\, N \,.\, \bar{x} - \bar{x} \,.\, N \,.\, \bar{y} + N \,.\, \bar{x} \,.\, \bar{y} \\
&= \sum_1^n x_i \,.\, y_i - N \,.\, \bar{x} \,.\, \bar{y} \\
&= \sum_1^n x_i \,.\, y_i - \frac{\sum_1^n x_i \,.\, \sum_1^n y_i}{N}
\end{aligned}$$

$$cov(x,y)_{population} = \frac{\sum_1^n x_i \,.\, y_i - \frac{\sum_1^n x_i \,.\, \sum_1^n y_i}{N}}{N} \qquad cov(x,y)_{sample} = \frac{\sum_1^n x_i \,.\, y_i - \frac{\sum_1^n x_i \,.\, \sum_1^n y_i}{n}}{n-1}$$

Covariance quantifies the strength of the linear association between 2 numerical variables. Units of Covariance: x-variable X y-variable (e.g. years.cm or years.Rs); hence difficult to interpret. When both variables are moving in the same direction covariance is positive.

CORRELATION COEFFICIENT

$$r = \frac{cov(x,y)}{S_x \,.\, S_y} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2} \,.\, \sqrt{\sum_1^n (y_i - \bar{y})^2}} = \frac{\sum_1^n x_i \,.\, y_i - \frac{\sum_1^n x_i \,.\, \sum_1^n y_i}{n}}{\sqrt{(\sum x^2 - n \,.\, \bar{x}^2)} \,.\, \sqrt{(\sum y^2 - n \,.\, \bar{y}^2)}}$$

$$-1 \leq r \leq +1$$

The Pearson correlation coefficient is derived from covariance. Divide covariance of x and y by product of standard deviations of x and y. Units of standard deviations cancel out the units of covariance.

POINT BI-SERIAL CORRELATION COEFFICIENT

For a dichotomous categorical variable (2 categories) . X is a numerical variable and Y is a categorical variable.

e.g.: Gender(Y0 and Y1) and Marks(X)

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{S_x}\right) \cdot \sqrt{p_0 \cdot p_1} \qquad p_0 = \frac{n_0}{n}; p_1 = \frac{n_1}{n}$$

GOODNESS OF FIT

$$R^2 = r^2$$

R^2 is a measure of the proportion of variance in the data-set explained by the explanatory variable. The measure is closer to 1 when r is closer to -1 or +1.

EFFECT OF MANIPULATING DATA WITH CONSTANT

	Add Constant (+C)	Multiply Constant (xC)	Outliers
Mean	$+C$	$*\,C$	Affected
Median	$+C$	$*\,C$	Not Affected
Mode	$+C$	$*\,C$	Not Affected
Variance	no effect	$*\,C^2$	Affected
Standard Deviation	no effect	$*\,C$	Affected
Covariance	?	?	?
Correlation Coefficient	?	?	?

Calculation Tables for Computational Formulae (Shortcuts)

Variance (sample)	x	x^2					$S^2_{sample} = \dfrac{\sum_1^n (x_i^2) - n\,\bar{x}^2}{n-1}$
Covariance (sample)	x	y	$x\,y$				$= \dfrac{\sum_1^n x_i \,.\, y_i - \dfrac{\sum_1^n x_i \cdot \sum_1^n y_i}{n}}{n-1}$
Correlation Coefficient	x	y	$x\,y$	x^2	y^2		$= \dfrac{\sum_1^n x_i \,.\, y_i - \dfrac{\sum_1^n x_i \cdot \sum_1^n y_i}{n}}{\sqrt{(\sum x^2 - n\,.\,\bar{x}^2)}\,.\,\sqrt{(\sum y^2 - n\,.\,\bar{y}^2)}}$