UNIT-1

Sample variance $s^2 = \frac{\Sigma[(x_i - \bar{x})^2]}{n-1}$

$\bar{x}$ → mean = $\frac{\text{sum of all obs}}{\text{No of obs}}$

x_i → obs.

n → Total no. of obs

Sample standard Deviation $S = \sqrt{\text{variance}}$

Correlation

+ve ↑↑ | -ve ↑↓ | Perfect ↑↑ (equal prop)

Karl Pearson coeff of correlation

$$r = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$x = x - \bar{x}$

$y = y - \bar{y}$

$$= \frac{\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]}{\frac{1}{n}\left[\Sigma (x_i - \bar{x})^2 * \Sigma (y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

Covariance

$$cov_{xy} = \frac{\Sigma (x_i - \bar{x})\Sigma (y_i - \bar{y})}{N-1}$$

No. of data values

$r \in [-1, 1]$ if $r = 0$ not correlated

if $r = 1$ +vely perfect correlated

if $r = -1$ -vely perfect correlated

Direct method for Corr. coeff

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} * \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

(small x & y)

No. of obs

Regression

Stepping back towards the normal.

Regression line Y on X | Regression line X on Y

Regression line Y on X

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

where $r\frac{\sigma_x}{\sigma_y} = b_{yx}$ is the regression coefficient of y on x

σ_x = standard deviation of x

σ_y = standard deviation of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

$\sqrt{\frac{\Sigma x^2}{n}} = \sigma_x$

$\sqrt{\frac{(x-\bar{x})^2}{n}}$

2) Regression line X on Y

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

where $r\frac{\sigma_x}{\sigma_y} = b_{xy}$ is the regression coefficient of x on y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

σ_x = SD of x

σ_y = SD of y.

Standard Error

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

Coeff of variabelety

$$Cv = \frac{\sigma_x^2}{\sigma_y^2}$$

$(b_{xy})^2 = r^2 \frac{\sigma_x^2}{\sigma_y^2}$

we have $b_{xy} = r\sigma_x/\sigma_y$

$\frac{\sigma_x^2}{\sigma_y^2} = (b_{xy})/r^2$

Rank correlation is measured by spearman rank correlation coeff

$$\rho = 1 - \left[\frac{6 * \left(\sum_{i=1}^{n} d_i^2\right)}{n(n^2-1)}\right] \quad [-1,1]$$

$n \to$ no of obs.
$x_i \to$ Rank of x
$y_i \to$ Rank of y
$d_i = x_i - y_i$

Repetition of Rank Correlation

$$\rho = 1 - \left[\frac{6 * [\Sigma d_i^2 + CF]}{n(n^2-1)}\right]$$

CF = Correlation factor
CF = CF in X series + CF in Y series

$$CF = \frac{\sum m(m^2-1)}{12}$$

$m \to$ no. of times the rank is repeated

Limit of correlation

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} \overbrace{(x_i - \bar{x})}^{a_i} \overbrace{(y_i - \bar{y})}^{b_i}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\Rightarrow r_{xy}^2 = \frac{(1/n)^2\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 * \frac{1}{n}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)} = \frac{\frac{1}{n^2}\left(\sum_{i=1}^{n} a_i b_i\right)^2}{\frac{1}{n^2}\left(\sum_{i=1}^{n} a_i^2 \Sigma b_i^2\right)}$$

$$\Rightarrow \boxed{r_{xy}^2 = \frac{\sum_{i=1}^{n}(a_i b_i)^2}{\Sigma a_i^2 \Sigma b_i^2}} \quad (1)$$

CAUCHY SWARTZ INEQUALITY

$$(\Sigma a_i b_i)^2 \leq (\Sigma a_i)^2 (\Sigma b_i)^2$$

sign of equality holding when $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_n}{b_n}$

Using cauchy swarz inequality in eqn (1) we have

$$r^2(x,y) \leq 1 \Rightarrow r \in [-1,1]$$

Properties of correlation coeff:-

1) $r \in [-1,1]$ 2) Independent of change of origin & Scale

Properties of Regression coeff:-

1) Correlation coeff is Geometric Mean b/w the regression coeff $b_{yx} * b_{xy} = r\frac{\sigma_y}{\sigma_x} * r\frac{\sigma_x}{\sigma_y} = r^2$

$$\boxed{r = \pm\sqrt{b_{yx} * b_{xy}}}$$

(sign of r is same as regression coeff)

2) $r^2 \leq 1 \Rightarrow |b_{xy}| \leq |1/b_{yx}|$

3) $\left|\frac{b_{xy}+b_{yx}}{2}\right| > (r)$ (not less than)

4) regression coefficient independent of change of Origin but not Scale
Origin ✓ Scale ✗

Angle b/w Regression lines

$$\theta = \tan^{-1}\left\{\left(\frac{1-r^2}{|r|}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\}$$

1) $r=0$, $\tan\theta = \infty$, $\theta = \pi/2$
2) $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π

Used to decide whether to include or not an additional independent variable in regression analysis

Partial Correlation

$$r_{12.3} = \frac{Cov(X_{1.3}, X_{2.3})}{\sqrt{Var(X_{1.3})\, Var(X_{2.3})}}$$

$$r_{12.3} = \frac{r_{12} - r_{13} * r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$r_{B.2} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}$$

$$r_{23.1} = \frac{r_{23} - r_{21} * r_{31}}{\sqrt{(1-r_{21}^2)(1-r_{31}^2)}}$$

Correlation ratio (η)

measure of curvilinear relt= b/w 2 variables (Concentration of pt about the curve of best fit)

Note: If regression is linear $\eta = r$, otherwise $|\eta| > |r|$

$$\eta_{yx}^2 = \frac{\sum_{i=1}^{m}\left\{\frac{T_i^2}{n_i}\right\} - \frac{T^2}{N}}{\text{Y on X}}$$

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij}^2) - \left(\frac{T^2}{N}\right)}{}$$

(NEXT PAGE)

η_{xy} = " $\left(\frac{T_j^2}{n_j}\right)$ − " / "
(Same as above) " "

Multiple Correlation

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{(1-r_{23}^2)}$$

$$R_{3.12}^2 = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{(1-r_{12}^2)}$$

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1-r_{13}^2}$$

Properties

1) measures ♡ b/w var & grp
2) When MCC of expected value is calculated by MCC ^ least sq method > MCC linear combination of variables
3) $0 \leq R_{1.23} \leq 1$
4) If $R_{1.23}=0$, Total & Partial Corr involving x_1 is zero
5) If $R_{1.23}=1$ Perfect correlation
6) $R_{1.23} > r_{12}, r_{13}, r_{23}$

Prop. of Correlation Ratio

1) Independent of change in origin & scale
i.e. $\eta^2_{yx} = \eta^2_{xy} + \eta^2_{uv}$

2) $\eta \in [0, 1)$

$N = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}$

$T_j = \sum_{j=1}^{n} f_{ij} y_{ij}$

$T = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} y_{ij} = \sum_{i=1}^{m} T_i$

(Last page (corr. Ratio))

Reference for η

Analysis of Variance (ANOVA)

Anova I way Classification

RSS = Raw Sum of squares

$$RSS = \sum_{i=1}^{k} \sum_{j=1}^{h} x_{ij}^2$$

N = Total no. of observation

G = Grand Total $= \sum_{i=1}^{k} \sum_{j=1}^{h} x_{ij}$

TSS = Total sum of squares

CF = Correction Factor $= \frac{G^2}{N}$

$TSS = RSS - CF$

SST = Sum of squares Treatment $= \sum_{i=1}^{k} \left(\frac{T_i^2}{n_i}\right) - CF$

SSE = Error sum of squares.. (Some of squares due to error)

$SSE = TSS - SST$

Source of variation	df degrees of freedom	Sum of Squares	mean sum of squares	Variance Ratio
Treatment	$k-1$	SST	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	$\sim F_{(k-1),(n-k)}$
TSS	$n-1$			

Stats — Estimation of param.
— Testing of Hypothesis

Estimation — procedure of estimating a population parameter by using sample information or observation.

Parameter — A statistical constant derived from the population values

Anova II way

Rows → k Column → h

Formula:- $TSS = RSS - CF$

$CF = \frac{G^2}{N}$

$SST = \frac{1}{h} \sum T_i^2 - CF$ (h columns)

$SSV = \frac{1}{k} \sum T_j^2 - CF$

$G = \sum_{i=1}^{k} \sum y_{ij}$ $T_i = \sum_{j=1}^{k} y_{ij}$

$T_j = \sum_{j=1}^{h} y_{ij}$

$SSE = TSS - SST - SSV$

Source of Variance	d.f	Sum of Squares	Mean sum of squares	Variance Ratio
Treatment	$k-1$	SST	$MST = \frac{SST}{k-1}$	$F = MST/MSE$
Varieties	$h-1$	SSV	$MSV = \frac{SSV}{h-1}$	$\sim F_{(k-1)/(h-1)(k-1)}$
Error	$(k-1)(h-1)$	SSE	$MSE = \frac{SSE}{(k-1)(h-1)}$	$F = \frac{MSV}{MSE}$ $F_{(h-1)/(h-1)(k-1)}$
Total	$hk-1$			

Anova II way Table

Properties of Estimators

It's good estimator if it is:
(I) unbiased (III) Efficient
(II) Consistent (IV) Sufficient

I) unbiased – A statistic $\hat{\theta}$ is said to be unbiased iff the mean of the sample distribution of the estimator $= \theta$ ie

$$\text{Mean of } \hat{\theta} = E(\hat{\theta}) = \theta$$

* If the estimator is not unbiased, then difference $E(\hat{\theta}) - \theta$ is called bias of the estimator

Positive Bias	Negative Bias
$E(\hat{\theta}) > \theta$	$E(\hat{\theta}) < \theta$

II) Consistency – An estimator $T_n = T(x_1, x_2 \dots x_n)$ based on a random sample of size n is const. if T_n converges to $\gamma(\theta)$ ie if $T_n \xrightarrow{P} \gamma(\theta)$ as $n \to \infty$

OR

In other words, it is consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0$, $\eta > 0$, there exists a the integer $n \geq m(\varepsilon, \eta)$ such that $P\{|T_n - \gamma(\theta)| < \varepsilon\} \to 1$ as $n \to \infty \Rightarrow P\{|T_n - \gamma(\theta)| < \varepsilon\} > 1 - \eta \ \forall n \geq n$ where n a some very large value

$\gamma, \theta \to$ parameter
$T_n \to$ Estimator

III) MVUE (minimum variance unbiased estimator)
OR (most efficient unbiased Estimator)

If $\hat{\theta}_1$ & $\hat{\theta}_2$ are 2 unbiased Estimator of θ, if σ_1^2 & σ_2^2 are variance of their sampling distribution & $\sigma_1^2 < \sigma_2^2$, $\hat{\theta}_1$ is said to be more unbiased of θ

Note – if $\hat{\theta}_1, \hat{\theta}_2 \dots \hat{\theta}_k$ are unbiased estimators of parameter θ $\sigma_1^2, \sigma_2^2, \dots \sigma_k^2$ are variance of there sampling distribution then $\hat{\theta}_1$ is the most efficient Estimator of θ if $\sigma_1^2 < \sigma_2^2 \dots \sigma_k^2$

$$e = \frac{\text{variance } \theta}{\text{variance } \theta_i} \text{ where } i = 2, 3, \dots$$

$$e_1 = \frac{\text{Var } \theta_1}{\text{Var } \theta_2} \qquad e = \text{Var}\theta / \text{Var}\theta_i$$

Sufficiency: An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

If $T = t(x_1, x_2, \ldots x_n)$ is an estimator of a parameter θ based on sample $x_1, x_2 \ldots x_n$ of size n from the population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2 \ldots x_n$ given T, is independent of θ, then t is sufficient estimator of θ

Method of Estimation

$L = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta)$

$= \prod_{i=1}^{n} (f(x_i, \theta))$

1. Maximum likelihood method
2. Method of moments
3. Method of minimum Chi-square
4. Method of min. variance
5. Method of least squares
6. Method of inverse probability
