

## GA 4101 Advanced Business Statistics and Analytics

Credits: 03

Lectures - 30

### Objectives

- To understand statistical ideas and understand why and when the various methods should be used.
- To build assumptions and limitations of various techniques of correlation, regression and hypothesis testing.
- To develop expertise on use of software like SPSS and Advanced excel for better understanding and interpretation of the results obtained.

### Course Outcomes

On the completion of this course students should be able to:

- Conduct research that addresses real-world problems using quantitative and methodological models
- Build awareness on how to visualize and interpret data.
- Develop basic data literacy and an analytic mindset that will help them make strategic decisions based on data.

### Contents

### No. of Lectures

#### Unit 1 – Data Collection and Descriptive Statistics

06

Data Classification, Arranging of the data in tabular form, Frequency distribution & cumulative frequency distribution, Graphs, charts & diagrams, Measures of central tendency: mean, median and mode and their implications, Measures of Dispersion: range, skewness, standard deviation and mean deviation

Lab Work – Data – Getting started with Excel, Tabular and Graphical Methods in Excel, Descriptive statistics Using Excel

#### Unit 2–Correlation and Regression Analysis

06

Correlation and Regression: Meaning and uses, Various methods of calculation of coefficients and their analysis and implication, Partial Correlation, multiple Correlation & Regression Analysis - two variable and multi variable cases.

Lab Work – Multiple Regression Analysis using Excel, Time Series Analysis using Excel

#### Unit 3 – Sampling and Testing of Hypothesis

12

Sampling & Estimation Theory, Hypothesis Testing : T-Test, Z-Test, F-Distribution, Chi Square and other Non-Parametric Test, Mann-Whitney U Test, Wilcoxon Matched Pairs Test, Kruskal-Wallis Test, ANOVA, Factor Analysis

Lab Work – Performing t-test, Z-test, Chi-Square tests using Excel and other non parametric test using SPSS

#### Unit 4 – Statistical Quality Control

03

Techniques of Statistical Quality Control, In-process quality control Techniques, Control Charts for Variables, Control Charts for attributes : C-Chart, *p*-chart, *np*-Chart

Lab Work – Control Charts using Minitab

#### Unit 5 – Decision Theory and Decision Trees

03

Types of Decision making environments, Decision making under uncertainty, Decision making under risk, Posterior Probabilities and Bayesian Analysis

### Suggested Readings

#### Text books

1. R. Lyman Ott, Micheal T. Longnecker, "An Introduction to Statistical Methods and Data Analysis", 2010, Cengage Learning Inc.
2. Quantitative techniques in management 4 th ed. /by Vohra, N D. – New Delhi: TMH, 2010.

# Formulae Used

1. Summation of  $n$  numbers

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

Simplified expression for the summation of  $n$  numbers

$$\sum x_i = x_1 + x_2 + \dots + x_n$$

2. Sample mean,  $\bar{x} = \frac{\sum x_i}{n}$

$$\text{Population mean, } \mu = \frac{\sum x_i}{N}$$

$$\text{Sample mean for grouped data, } \bar{x} = \frac{\sum f_i m_i}{n}$$

where  $n = \sum f_i$  and  $m_i$  = mid-value of class intervals

3. Weighted mean for a population or a sample,

$$\bar{x}_w \text{ or } \mu_w = \frac{\sum w_i x_i}{\sum w_i}$$

where  $w_i$  = weight for observation  $i$

4. Position of the median in an ordered set of observation

belong to a population or a sample is  $\text{Med} = x_{(n/2) + (1/2)}$

$$\text{Median for grouped data, } \text{Med} = l + \left[ \frac{(n/2) - cf}{f} \right] h$$

5. Quartile for a grouped data

$$Q_i = l + \left[ \frac{i(n/4) - cf}{f} \right] h ; i = 1, 2, 3$$

Decile for a grouped data

$$D_i = l + \left[ \frac{i(n/10) - cf}{f} \right] h ; i = 1, 2, \dots, 9$$

Percentile for a grouped data

$$P_i = l + \left[ \frac{i(n/100) - cf}{f} \right] h ; i = 1, 2, \dots, 99$$

6. Mode for a grouped data

$$M_o = l + \left[ \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \right] h$$

Mode for a multimode frequency distribution

$$M_o = 3 \text{ Median} - 2 \text{ Mean}$$

## Formulae Used

### 1. Range, R

Value of highest observation - Value of lowest observation =  $H - L$

$$\text{Coefficient of range} = \frac{H - L}{H + L}$$

### 2. Interquartile range = $Q_3 - Q_1$

$$\text{Quartile deviation, QD} = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

### 3. Mean average deviation

For ungrouped data

$$(i) \text{ MAD} = \frac{\sum |x - \bar{x}|}{n}, \text{ for sample}$$

$$(ii) \text{ MAD} = \frac{\sum |x - \mu|}{N}, \text{ for population}$$

$$(iii) \text{ MAD} = \frac{\sum |x - \text{Me}|}{n}, \text{ from median}$$

$$\text{For grouped data} \quad \text{MAD} = \frac{\sum f|x - \bar{x}|}{\sum f}$$

$$4. \text{ Coefficient of MAD} = \frac{\text{MAD}}{\bar{x} \text{ or Me}} \times 100$$

### 5. Variance

Ungrouped data

$$\begin{aligned} \sigma^2 &= \frac{\sum (x - \bar{x})^2}{N} = \frac{\sum x^2}{N} - \left( \frac{\sum x}{N} \right)^2 \\ &= \frac{\sum d^2}{N} - \left( \frac{\sum d}{N} \right)^2 \end{aligned}$$

where  $d = x - A$ ; A is any assumed A.M. value

$$\text{Grouped data, } \sigma^2 = \left[ \frac{\sum f d^2}{N} - \left( \frac{\sum f d}{N} \right)^2 \right] h$$

where  $d = (m - A)/h$ ;  $h$  is the class interval and  $m$  is the mid-value of class intervals.

### 6. Standard deviation

$$\text{Ungrouped data, } \sigma = \sqrt{\sigma^2}$$

$$\text{Grouped data, } \sigma = \sqrt{\frac{\sum f d^2}{N} - \left( \frac{\sum f d}{N} \right)^2} \times h$$

$$7. \text{ Coefficient of variation (CV)} = \frac{\sigma}{\bar{x}} \times 100$$

## Formulae Used

### 1. Absolute measure of skewness

$$Sk = \bar{x} - \text{Mode or } Q_3 + Q_1 - 2 \text{ Med}$$

### 2. Coefficient of skewness

Karl Pearson's

$$Sk_p = \frac{\bar{x} - Mo}{\sigma} \text{ or } \frac{3(\bar{x} - \text{Med})}{\sigma}$$

$$\text{Bowley's, } Sk_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$

$$\text{Kelly's, } Sk_k = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}} \text{ or } \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

### 3. Coefficient of skewness based on moments

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}; \beta_2 = \frac{\mu_4}{\mu_2^2}$$

### 4. Moments about the actual mean

$$\mu_r = \frac{1}{n} \sum (x - \bar{x})^r, r = 1, 2, 3, 4$$

About an assumed mean, A

$$\mu'_r = \frac{1}{n} \sum (x - A)^r, r = 1, 2, 3, 4$$

5. Kurtosis  $\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}}$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

6. For a normal curve,  $\beta_2 = 3$  or  $\gamma_2 = 0$ ; for a leptokurtic curve,  $\beta_2 > 3$  or  $\gamma_2 > 0$  and for a platykurtic curve,  $\beta_2 < 3$  or  $\gamma_2 < 0$ .

## Review Self-practice Problems



# Formulae Used

## 1. Counting methods for determining the number of outcomes

- Multiplication method

(i)  $n_1 \times n_2 \times \dots \times n_k$

(ii)  $n_1 \times n_2 \times \dots \times n_k = n^k$

when the event in each trial is the same

- Number of Permutations  ${}^nP_r = \frac{n!}{(n-r)!}$

- Number of Combinations  ${}^nC_r = \frac{n!}{r!(n-r)!}$

## 2. Classical or *a priori* approach of computing probability of an event A

$$P(A) = \frac{\text{Number of favourable cases for A}}{\text{All possible cases}} = \frac{c(n)}{c(s)}$$

## 3. Relative frequency approach of computing probability of an event A in $n$ trials of an experiment

$$P(A) = \lim_{n \rightarrow \infty} \frac{c(A)}{n}$$

## 4. Rule of addition of two events

- When events A and B are mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B)$$

- When events A and B are not mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

## 5. Conditional probability

- For statistically independent events

$$P(A|B) = P(A); P(B|A) = P(B)$$

- For statistically dependent events

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

## 6. Rule of multiplication of two events

- Joint probability of independent events

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

- Joint probability of dependent events

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

$$P(A \text{ and } B) = P(B|A) \times P(A)$$

## 7. Rule of elimination

(i)  $P(B) = \sum P(A_i) P(B|A_i)$

(ii)  $P(A) = \sum P(B_i) P(A|B_i)$

## 8. Baye's rule $P(A_i|B) = \frac{P(A_i) P(B|A_i)}{\sum P(A_i) P(B|A_i)}$

## 9. Basic rules for assigning probabilities

- The probability assigned to each experimental outcome

$$0 \leq P(A_i) \leq 1 \text{ for all } i$$

- Sum of the probabilities for all the experimental outcomes

$$\sum P(A_i) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

Complement of an event,  $P(A) = 1 - P(\bar{A})$

## Formulae Used

1. Expected value of a random variable  $x$

$$E(x) = \sum x \cdot P(x)$$

where  $x$  = value of the random variable

$P(x)$  = probability that the random variable will take on the value  $x$ .

2. Binomial probability distribution

- Probability of  $r$  success in  $n$  Bernoulli trials

$$P(x = r) = {}^nC_r p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

where  $p$  = probability of success

$q$  = probability of failure,  $q = 1 - p$

- Mean and standard deviation of binomial distribution

Mean  $\mu = np$

Standard deviation  $\sigma = \sqrt{npq}$

4. Poisson probability distribution

- Probability of getting exactly  $r$  occurrences of random event

$$P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

where  $\lambda = np$ , mean number of occurrences per interval of time

$e = 2.71828$ , a constant that represents the base of the natural logarithm system

- Mean and standard deviation of Poisson distribution

$$\lambda = np, \sigma = np$$

5. Normal distribution formula:

Number of standard deviations  $\sigma$  a value of random variable  $x$  is away from the mean  $\mu$  of normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

## Formulae Used

1. Standard deviation (or standard error) of sampling distribution of mean,  $\bar{x}$

- Infinite Population:  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Finite Population:  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

where  $n < 0.5N$ ;  $n, N$  = size of sample and population, respectively.

2. Estimate of  $\sigma_{\bar{x}}$  when population standard deviation is not known

- Infinite Population:  $s_{\bar{x}} = \frac{s}{\sqrt{n}}$
- Finite Population:  $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

3. Standard deviation of sampling distribution of sample means

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

4. Standard deviation (or standard error) of sampling distribution of proportion

- Infinite Population:  $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$ ;  $q = 1 - p$
- Finite Population:  $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$

5. Standard deviation of sampling distribution of sample proportions

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}};$$

$$q_1 = 1 - p_1; q_2 = 1 - p_2$$



where  $s = \sqrt{\sum (x - \bar{x})^2 / (n - 1)}$  ;  $df = n - 1$

3. Confidence interval to estimate  $\mu$ : large sample size ( $n > 30$ ) and population standard deviation  $\sigma$  is known

- $\mu = \bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}$ , where  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

- $\bar{x} - z_{\alpha/2} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

4. Confidence interval to estimate  $\mu$  using finite correction factor:

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

5. Confidence interval to estimate population proportion large sample size  $n \geq 30$

$$p = \bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}q}{n}}$$

6. Confidence interval to estimate the difference between the means of two normally distributed populations

- When standard deviations are known

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- When standard deviations are not known

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

7. Sample size when estimating:

- Population mean,  $\mu$

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

- Population proportion,  $p$

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2}; q = 1 - p$$

8. Sample size when estimating finite population mean,  $\mu$  with size  $N$

$$n = \frac{n_0 N}{n_0 + (N - 1)}, \text{ where } n_0 = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

# Formulae Used

## 1. Hypothesis testing for population mean with large sample ( $n > 30$ )

### (a) Test statistic about a population mean $\mu$

- $\sigma$  assumed known,  $z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$

- $\sigma$  is estimated by  $s$ ,  $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

### (b) Test statistic for the difference between means of two populations

- Standard deviation of  $\bar{x}_1 - \bar{x}_2$  when  $\sigma_1$  and  $\sigma_2$  are known

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\text{Test statistic } z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

- Standard deviation of  $\bar{x}_1 - \bar{x}_2$  when  $\sigma_1^2 = \sigma_2^2$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Point estimator of  $\sigma_{\bar{x}_1 - \bar{x}_2}$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Interval estimation for single population mean

$$\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}; \sigma \text{ is known}$$

$$\bar{x} \pm z_{\alpha/2} s_{\bar{x}}; \sigma \text{ is unknown}$$

- Interval estimation for the difference of means of two populations

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sigma_{\bar{x}_1 - \bar{x}_2}; \sigma_1 \text{ and } \sigma_2 \text{ are known}$$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}; \sigma_1 \text{ and } \sigma_2 \text{ are unknown}$$

## 2. Hypothesis testing for population proportion for large sample ( $n > 30$ )

### (a) Test statistic for population proportion $p$

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}}; \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

### (b) Test statistic for the difference between the proportions of two populations

- Standard deviation of  $\bar{p}_1 - \bar{p}_2$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Point estimator of

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

- Interval estimation of the difference between the proportions of two populations

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} s_{\bar{p}_1 - \bar{p}_2}$$

where all  $n_1 p_1$ ,  $n_1(1-p_1)$ ,  $n_2 p_2$  and  $n_2(1-p_2)$  are more than or equal to 5

- Test statistic for hypothesis testing about the difference between proportions of two populations

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}}$$

- Pooled estimator of the population proportion

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$

- Point estimator of  $\sigma_{\bar{p}_1 - \bar{p}_2}$

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

### 3. Hypothesis testing for population mean with small sample ( $n \leq 30$ )

- Test statistic when  $s$  is estimated by  $s$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

- Test statistic for difference between the means of two population proportions

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

where  $s_{\bar{x}_1 - \bar{x}_2}$  is the point estimator of  $\sigma_{\bar{x}_1 - \bar{x}_2}$

when  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  and

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Interval estimation of the difference between means of two populations  $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$

### 4. Hypothesis testing for matched samples (small sample case)

Test statistic for matched samples

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}; s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}}; \bar{d} = \frac{\sum d}{n}$$

### 5. Hypothesis testing for two population variances

$$F = s_1^2/s_2^2; s_1^2 > s_2^2$$

## Formulae Used

1.  $\chi^2$ -test statistic  $\chi^2 = \sum (O - E)^2 / E$
2. Expected frequencies for a contingency table

$$E_{ij} = \frac{\text{Row } i \text{ total} \times \text{Column } j \text{ total}}{\text{Sample size}}$$

3. Degree of freedom for a contingency table  
 $df = (r - 1) (c - 1)$

## Chapter Concepts Quiz

## Formulae Used

### 1. One-way analysis of variance

- Grand sample mean

$$\bar{x} = \frac{\sum_{i=1}^k \sum_{j=1}^r x_{ij}}{n}, n = n_1 + n_2 + \dots + n_r$$

- Correction factor  $CF = \frac{T^2}{n}$
- Total sum of squares

$$SST = \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x})^2 = \sum_i \sum_j x_{ij}^2 - CF$$

- Sum of squares of variations between samples due to treatment

$$SSTR = \sum_{j=1}^r n_j (\bar{x}_j - \bar{x})^2 = \frac{1}{n_j} \sum_{j=1}^r x_j^2 - CF$$

- Sum of squares of variations within samples or error sum of squares

$$SSE = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = SST - SSTR$$

- Mean square between samples due to treatments

$$MSTR = \frac{SSTR}{r - 1}$$

- Mean square within samples due to error

$$MSE = \frac{SSE}{n - r}$$

- Test statistic for equality of  $k$  population means

$$F = \frac{MSTR}{MSE}$$

- Degrees of freedom

$$\text{Total } df (n - 1) = \text{Treatment } df (r - 1) + \text{Random error } df (n - r)$$

### 2. Two-way analysis of variance

- Total sum of squares

$$SST = \sum_{j=1}^r \sum_{i=1}^k x_{ij}^2 - n (\bar{x})^2$$

- Sum of squares of variances between columns due to treatments

$$SSTR = \sum_{j=1}^r (\bar{x}_j)^2 - n (\bar{x})^2$$

- Sum of squares between rows due to blocks

$$SSR = \sum_{i=1}^k (\bar{x}_i)^2 - n (\bar{x})^2$$

- Sum of squares due to error

$$SSE = SST - (SSTR + SSR)$$

- Degrees of freedom

$$df_c = c - 1; df_r = (r - 1)$$

$$df(\text{residual error}) = \text{Blocks } df + \text{Treatments } df = (r - 1)(c - 1)$$

- Mean squares between columns due to treatment

$$MSTR = \frac{SSTR}{c - 1}$$

- Mean square between rows due to blocks

$$MSR = \frac{SSR}{r - 1}$$

- Mean square of residual error

$$MSE = \frac{SSE}{(c - 1)(r - 1)}$$

- Test statistic

$$F_1 = \frac{MSTR}{MSE}; F_2 = \frac{MSR}{MSE}$$

provided numerator is bigger than denominator.



## Formulae Used

### 1. Karl Pearson's correlation coefficient

$$r = \frac{\text{Covariance between } x \text{ and } y}{\sigma_x \sigma_y}$$

- Deviation from actual mean

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

- Deviation from assumed mean

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

$$d_x = x - A, d_y = y - B$$

A, B = constants

- Bivariate frequency distribution

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

- Using actual values of  $x$  and  $y$

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

### 2. Standard error of correlation coefficient, $r$

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

- Probable error of correlation coefficient,  $r$

$$PE_r = 0.6745 \frac{1 - r^2}{\sqrt{n}}$$

### 3. Coefficient of determination

$$r_2 = \frac{\text{Explained variance}}{\text{Total variance}} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

### 4. Spearman's rank correlation coefficient

- Ranks are not equal

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

- Ranks are equal

$$R = 1 - \frac{6 \left[ \sum d^2 + \frac{1}{12} (m_i^3 - m_i) \right]}{n(n^2 - 1)}$$

$$t = 1, 2, \dots$$

### 5. Hypothesis testing

- Population correlation coefficient  $r$  for a small sample

$$t = \frac{r - \rho}{SE_r} = r \sqrt{\frac{n - 2}{1 - r^2}}$$

- Population correlation coefficient for a large sample

$$Z = \frac{z - z_p}{\sigma_z} = \frac{z - z_p}{1/\sqrt{n - 3}}$$

# Formulae Used

1. Simple linear regression model

$$y = \beta_0 + \beta_1 x + e$$

2. Simple linear regression equation based on sample data

$$y = a + bx$$

3. Regression coefficient in sample regression equation

$$b = \hat{y}$$

$$a = \bar{y} - b\bar{x}$$

4. Residual representing the difference between an observed value of dependent variable  $y$  and its fitted value

$$e = y - \hat{y}$$

5. Standard error of estimate based on sample data

- Deviations formula

$$S_{y.x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$$

- Computational formula

$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a \Sigma y - b \Sigma xy}{n - 2}}$$

6. Coefficient of determination based on sample data

- Sums of squares formula

$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

- Computational formula

$$r^2 = \frac{a \Sigma y + b \Sigma xy - n(\bar{y})^2}{\Sigma y^2 - n(\bar{y})^2}$$

7. Regression sum of squares

$$S_{y.x} = S_y \sqrt{1 - r^2}$$

8. Interval estimate based on sample data:  $\hat{y} \pm t_{df} S_{y.x}$

## Formulae Used

1. A multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e$$

2. Estimated multiple regression model

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

3. Sample standard error of the estimate

$$S_{y.12} = \sqrt{\frac{\Sigma y^2 - a \Sigma y - b_1 \Sigma x_1 y - b_2 \Sigma x_2 y}{n - 3}}$$

4. Coefficient of a multiple determination based on sample data for two independent variables

$$R_{y.12}^2 = 1 - \frac{S_{y.12}^2}{S_y^2}$$

5. Partial regression coefficients

$$b_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \left( \frac{s_1}{s_2} \right)$$

$$b_{13.2} = \frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \left( \frac{s_1}{s_3} \right)$$

6. Coefficient of a multiple correlation

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

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{13}^2}}$$

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

7. Partial correlation coefficient

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

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{13}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

## Formulae Used

1. A multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e$$

2. Estimated multiple regression model

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

3. Sample standard error of the estimate

$$S_{y.12} = \sqrt{\frac{\sum y^2 - a \sum y - b_1 \sum x_1 y - b_2 \sum x_2 y}{n - 3}}$$

4. Coefficient of a multiple determination based on sample data for two independent variables

$$R_{y.12}^2 = 1 - \frac{S_{y.12}^2}{S_y^2}$$

5. Partial regression coefficients

$$b_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \left( \frac{s_1}{s_2} \right)$$

$$b_{13.2} = \frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \left( \frac{s_1}{s_3} \right)$$

6. Coefficient of a multiple correlation

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

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{13}^2}}$$

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

7. Partial correlation coefficient

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

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{13}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$



# Formulae Used

## 1. Secular trend line

- Linear trend model

$$y = a + bx$$

$$\text{where } a = \bar{y} - b\bar{x}; b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

- Exponential trend model

$$y = ab^x;$$

$$\log a = \frac{1}{n} \sum \log y; \log b = \frac{\sum x \log y}{\sum x^2}$$

- Parabolic trend model

$$y = a + bx + cx^2$$

$$\text{where } a = \frac{\sum y - c \sum x^2}{n}; b = \frac{\sum xy}{\sum x^2}$$

$$c = \frac{n \sum x^2 y - \sum x^2 \sum y}{n \sum x^4 - (\sum x^2)^2}$$

## 2. Moving average

$$MA_{t+1} = \frac{\sum \{D_t + D_{t-1} + \dots + D_{t-n+1}\}}{n}$$

where  $t$  = current time period

$D$  = actual data value

$n$  = length of time period

## 3. Simple exponential smoothing

$$F_t = F_{t-1} + \alpha(D_{t-1} - F_{t-1})$$

where  $F_t$  = current period forecast

$F_{t-1}$  = previous period forecast

$\alpha$  = a weight ( $0 \leq \alpha \leq 1$ )

$D_{t-1}$  = previous period actual demand

## 4. Adjusted exponential smoothing

$$(F_t)_{\text{adj}} = F_t + \frac{1-\beta}{\beta} T_t$$

Where  $\beta$  = smoothing constant for trend

$T_t$  = exponential smoothed trend factor



## Formulae Used

- Price relatives in period  $n$ ,  $P_{0n} = \frac{p_n}{p_0} \times 100$   
 Quantity relative in period  $n$ ,  $Q_{0n} = \frac{q_n}{q_0} \times 100$   
 Value relative in period  $n$ ,  $V_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_0} \times 100$
- Unweighted aggregate price index in period  $n$

$$P_{0n} = \frac{\sum p_n}{\sum p_0} \times 100$$

Simple average of price relative

$$P_{0n} = \frac{1}{2} \sum \left( \frac{p_n}{p_0} \right) \times 100$$

Simple G.M. of price relative

$$P_{0n} = \text{antilog} \left[ \frac{1}{n} \sum \left( \frac{p_n}{p_0} \right) \times 100 \right]$$

Simple aggregate quantity index

$$Q_{0n} = \frac{\sum q_n}{\sum q_0} \times 100$$

- Weighted aggregate price indexes

(a) Weighted aggregate method in period  $n$

$$P_{0n} = \frac{\sum p_n q}{\sum p_0 q} \times 100$$

$$\text{Laspeyre's index, } I_p(L) = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

$$\text{Paasche's index, } I_p(P) = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

Marshall-Edgeworth's index

$$I_p(M-E) = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$$

Dorbish and Bowley's index

$$I_p(D-B) = \frac{1}{2} (L + P) \times 100$$

$$\text{Fisher's ideal index, } = \sqrt{L \times P} \times 100$$

(b) Weighted average of price relatives in period  $n$

$$P_{0n} = \frac{\sum \left( \frac{p_n}{p_0} \times 100 \right) W}{\sum W}$$

Weighted average of price relatives

$$P_{0n} = \frac{\sum \left( \frac{p_1}{p_0} \times 100 \right) (p_0 q_0)}{\sum p_0 q_0}$$

(base year value as weights)

Weighted average of price relatives

$$P_{0n} = \frac{\sum \left( \frac{p_1}{p_0} \times 100 \right) (p_1 q_1)}{\sum p_1 q_1}$$

(current year value as weights)

- Quantity indexes

(a) Unweighted quantity index in period  $n$

$$Q_{0n} = \frac{\sum q_n}{\sum q_0} \times 100$$

Simple average of quantity relative

$$Q_{0n} = \frac{1}{n} \sum \left( \frac{q_n}{q_0} \times 100 \right)$$

(b) Weighted quantity index in period  $n$

$$Q_{0n} = \frac{\sum q_n W}{\sum q_0 W} \times 100$$

- Tests for adequacy or consistency

Time reversal test:  $P_{0n} \times P_{n0} = 1$

Factor reversal test:  $P_{0n} \times Q_{0n} = \frac{\sum p_n q_n}{\sum p_0 q_0}$

Circular test:  $P_{01} \times P_{12} \times P_{23} \times \dots \times P_{(n-1)n} \times P_{n0} = 1$

- Link relative =  $\frac{\text{Current period price}}{\text{Price of the preceding period}} \times 100$

Chain index =  $\frac{\text{Current period's link relative} \times \text{Preceding period's chain index}}{100}$