PREFACE

The principal purpose of writing this book "**CSIR NET Statistics Theory Book**" is to cater to the needs of the candidates aspiring for "**CSIR NET Mathematical Sciences**" exam. Statistics has become an essential tool in modern science, driving research, innovation, and decision-making across fields. The CSIR NET examination stands as a benchmark for aspiring statisticians, requiring a strong grasp of theoretical concepts and practical applications. This book is designed as a comprehensive guide to help students navigate the examination with confidence and clarity.

The content is meticulously aligned with the CSIR NET syllabus. The book includes **15 chapters**. Each chapter builds concepts progressively, supplemented previous years' questions to ensure thorough preparation. Real-world illustrations and detailed explanations are included to enhance understanding of complex topics. This book will help the students not only for CSIR NET exam but also for other competitive exams like IIT-JAM, GATE, etc.

A word or suggestion from your side may add another feather to the cap in the compilation of this book. The author looks forward to the comments, suggestions, and criticism from the students. Constructive suggestions and user feedback would be highly appreciated, gratefully acknowledged, and suitably incorporated. We wish all the best to all aspirants who are preparing for these exams.

ACKNOWLEDGEMENT

Writing this book has been an amazing journey. First of all, we would like to thank brilliant students at Institute for Advanced Studies (IFAS), whose constant demand for such a book has been big motivation behind writing and shaping of this book. We want to extend our sincere appreciation to Mr. Radheshyam Choudhary, Director at IFAS Publications and Dr. Kailash Choudhary, Founder of IFAS. Our special thanks to Mr. Dadasaheb Sondge and Mr. Digambar Jagtap for their valuable support and critical suggestions for bringing out the book in its present form.

We would like to thank **Mr. Tushar Herekar** for his invaluable contribution to this book. His expertise, dedication, and attention to detail have made this work a comprehensive and reliable resource for aspirants. We would also like to thank Mr. Omkar Girkar for reviewing and aligning the content to ensure it meets the highest standards of accuracy and relevance.

This book is a team effort and producing it would be impossible without helping and supportive people at IFAS Publications. We are thankful to the team at IFAS Publications, most notably Ms. Gitanjalee Jadhav for helpful suggestions to improve the quality of the content and regular reminders for timely completion of the project. It has been a pleasure to work with Mr. Vikendra Metha and Mr. Kuldeep Singh who took great care during the copy-editing and production processes of this book.

And finally, our humble greetings to all who put their significant efforts and are unmentioned.

CSIR NET - STATISTICS

Exam Details

Mode of the examination	Computer Based Test
Language	English/Hindi
Duration of exam	3 hours
Number of questions	120
Type of questions	MCQ (Part A and B) and MSQ (Part C)
Total marks	200

Exam Pattern

	Part A	**Part B**	**Part C**	**Total**
Total Questions	20	40	60	120
Maximum questions to be attempted	15	25	20	60
Marks for each correct answer	2	3	4.75	-
Maximum Marks	30	75	95	200
Marks for each incorrect answer	- 0.5	- 0.75	0	-

Syllabus for Statistics Section (Unit IV)

Descriptive statistics, exploratory data analysis.

Sample space, discrete probability, independent events, Bayes theorem. Random variables and distribution functions (univariate and multivariate); expectation and moments. Independent random variables, marginal and conditional distributions. Characteristic functions. Probability inequalities (Tchebyshef, Markov, Jensen). Modes of convergence, weak and strong laws of large numbers, Central Limit theorems (i.i.d. case).

Markov chains with finite and countable state space, classification of states, limiting behaviour of n-step transition probabilities, stationary distribution, Poisson and birth-and-death processes.

Standard discrete and continuous univariate distributions, sampling distributions, standard errors and asymptotic distributions, distribution of order statistics and range.

Methods of estimation, properties of estimators, confidence intervals. Tests of hypotheses: most powerful and uniformly most powerful tests, likelihood ratio tests. Analysis of discrete data and chi-square test of goodness of fit. Large sample tests.

Simple nonparametric tests for one and two sample problems, rank correlation and test for independence. Elementary Bayesian inference.

Gauss-Markov models, estimability of parameters, best linear unbiased estimators, confidence intervals, tests for linear hypotheses. Analysis of variance and covariance. Fixed, random and mixed effects models. Simple and multiple linear regression. Elementary regression diagnostics. Logistic regression.

Multivariate normal distribution, Wishart distribution and their properties. Distribution of quadratic forms. Inference for parameters, partial and multiple correlation coefficients and related tests. Data reduction techniques: Principle component analysis, Discriminant analysis, Cluster analysis, Canonical correlation.

Simple random sampling, stratified sampling and systematic sampling. Probability proportional to size sampling. Ratio and regression methods.

Completely randomized designs, randomized block designs and Latin-square designs. Connectedness and orthogonality of block designs, BIBD. 2K factorial experiments: confounding and construction.

Hazard function and failure rates, censoring and life testing, series and parallel systems.

Linear programming problem, simplex methods, duality. Elementary queuing and inventory models. Steady-state solutions of Markovian queuing models: M/M/1, M/M/1 with limited waiting space, M/M/C, M/M/C with limited waiting space, M/G/1.

CONTENT

CSIR NET – STATISTICS THEORY BOOK

CHAPTER	PARTICULAR	PAGE NO.
1	Descriptive Statistics	1
2	Basic Concepts of Probability	12
3	Theory of Random Variables	19
4	Distribution Theory	31
5	Estimation Theory	49
6	Testing of Hypothesis	57
7	Non Parametric	63
8	Bayesian Inference	68
9	Correlation and Regression Analysis	75
10	Design of Experiments	90
11	Theory of Sampling	140
12	Multivariate Analysis	147
13	Stochastic Processes	153
14	Linear Programming Problem	161
15	Reliability, Survival Analysis and Queuing Theory	165

CHAPTER - 1

Descriptive Statistics

1.1 Measures of Central Tendency

Introduction

After getting the data in our hand it becomes essential to condense the data into a single value for comparison purpose. This value is treated as a representative of data. There are various methods of selecting a single central value.

Objectives

1. To obtain a single representative quantity for the entire data.
2. To facilitate comparison.

Definition: Central Tendency - It is evident from any frequency curve that observations in a data are not uniformly spread. However, most of the observations get clustered in the central part of the data. This property of observations is described as Central Tendency.

Definition: Average or Measure of Central Tendency - As a consequence of the property of Central Tendency, a representative value is chosen from the central part of the data. Such a representative value is known as an Average or Measure of Central Tendency.

Requisites of Good Average:

1. It should be simple to understand and easy to calculate.

2. It should be rigidly defined.

3. It should be based on all observations in the data.

4. It should be capable of further mathematical treatment.

5. It should be least affected by extreme observations.

6. It should possess sampling stability.

Types of Averages

1. **Mathematical Averages**
   - (1) Arithmetic Mean
   - (2) Geometric Mean
   - (3) Harmonic Mean
2. **Non-Mathematical Averages**
   - (1) Median
   - (2) Mode

1.1.1 Arithmetic Mean (A.M.)

Notation

The Arithmetic Mean is denoted by $\overline{x}$.

Definition

$$\text{Arithmetic Mean (A. M.)} = \frac{\text{Sum of Observations}}{\text{Number of Observations}}$$

Procedure

Individual Observations or Ungrouped Data - Suppose $x_1, x_2, \ldots, x_n$ is a set of n observations. Then

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Procedure

Discrete Frequency Distribution - Suppose $x_1, x_2, \ldots, x_k$ are distinct observations with $f_1, f_2, \ldots, f_k$ as the corresponding frequencies. Then

$$\overline{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

Procedure

Continuous Frequency Distribution - Suppose $C_1, C_2, \ldots, C_k$ are k class intervals with $f_1, f_2, \ldots, f_k$ as the corresponding frequencies. Let $x_1, x_2, \ldots, x_k$ be the mid-values of the class intervals. Then

$$\overline{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k}$$

$$= \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{N}$$

Definition: Outlies - Some observations in the data set fall outside the usual pattern of the data. Such observations are called as outlies.

Remark

Computing Arithmetic Mean in the presence of outlies may give faulty results. Thus, the interpretation and conclusion based on such a mean will not be that much reliable.

Definition: Trimmed Mean - Arithmetic Mean obtained by ignoring lowest α % as well as highest α % observations is called as α % Trimmed Mean.

Note: The concept of the Trimmed Mean is useful to overcome the difficulty in computing the Arithmetic Mean in the presence of the outliers.

Property

The algebraic sum of deviations of observations from their Arithmetic Mean is zero. That is

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

Property

The sum of squares of the deviations taken from Arithmetic Mean is minimum. That is

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2 \ , a \in \mathbb{R}$$

Property

Change of Origin and Scale - If

$$u_i = \frac{x_i - b}{k} \Rightarrow \overline{u} = \frac{\overline{x} - b}{k} \ ; \ b, k \in \mathbb{R}$$

Property

Mean of Combined Groups - Let $\overline{x}_1$ be the arithmetic mean of first group of size n_1. Similarly, $\overline{x}_2$ be the arithmetic mean of second group of size n_2. Then the combined mean $\overline{x}_c$ is given by

$$\overline{x}_c = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}$$

Generalization

Let there be k groups with size of i^{th} group as n_i and the Arithmetic Mean as $\overline{x}_i (i = 1,2, \dots, k)$. Then $\overline{x}_c$, the arithmetic Mean of all k groups combined together given by

$$\overline{x}_c = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \cdots + n_k\overline{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i\overline{x}_i}{\sum_{i=1}^{k} n_i}$$

Property

The Arithmetic Mean lies within minimum and maximum of the observations. That is

$$min\{x_1, x_2, \dots, x_n\} \leq \overline{x} \leq max\{x_1, x_2, \dots, x_n\}$$

Limitations

1. It is applicable only to quantitative data.
2. It is unduly affected by extreme observations.
3. It cannot be computed for frequency distribution with open end classes.
4. It cannot be determined graphically.
5. Sometimes Arithmetic Mean may not be an actual observation in a data.

1.1.2 Geometric Mean (G.M.)

Definition

$$\text{Geometric Mean (G. M.)} = (\text{Product of observations})^{\left(\frac{1}{\text{Number of Obsevations}}\right)}$$

Procedure

Individual Observations or Ungrouped Data - Suppose $x_1, x_2, \dots, x_n$ is a set of n observations. Then

$$G.M. = (x_1 \times x_2 \times \dots \times x_n)^{\left(\frac{1}{n}\right)} = \left(\prod_{i=1}^{n} x_i\right)^{\left(\frac{1}{n}\right)}$$

Procedure

Discrete Frequency Distribution - Suppose $x_1, x_2, \dots, x_k$ are distinct observations with $f_1, f_2, \dots, f_k$ as the corresponding frequencies. Then

$$G.M. = \left(x_1^{f_1} \times x_2^{f_2} \times \dots \times x_k^{f_k}\right)^{\left(\frac{1}{N}\right)} = \left(\prod_{i=1}^{k} x_i^{f_i}\right)^{\left(\frac{1}{N}\right)}, \text{where } N = \sum_{i=1}^{k} f_i$$

Procedure

Continuous Frequency Distribution - Suppose $C_1, C_2, \dots, C_k$ are k class intervals with $f_1, f_2, \dots, f_k$ as the corresponding frequencies. Let $x_1, x_2, \dots, x_k$ be the mid-values of the class intervals. Then

$$G.M. = \left(x_1^{f_1} \times x_2^{f_2} \times \dots \times x_k^{f_k}\right)^{\left(\frac{1}{N}\right)} = \left(\prod_{i=1}^{k} x_i^{f_i}\right)^{\left(\frac{1}{N}\right)}, \text{where } N = \sum_{i=1}^{k} f_i$$

Limitations

1. The serious drawback of the G.M. is that it is zero if any of the observations is zero.
2. It is not simple to understand and calculate.
3. It may be imaginary if some observations are negative. Therefore, it is calculated only for the data containing positive values.
4. It is not applicable to qualitative data.
5. It cannot be determined graphically.
6. It cannot be computed if frequency distribution includes open end class.
7. It may not be an actual observation in the data.

Result

If $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$ are the two sets of observations then

$$G.M.\left(\frac{x}{y}\right)=\frac{G.M.(x)}{G.M.(y)}$$

1.1.3 Harmonic Mean (H.M.)

Definition

$$\text{Harmonic Mean (H. M.)}=\frac{1}{\text{A. M. of reciprocal of observations}}$$

Procedure

Individual Observations or Ungrouped Data - Suppose $x_1, x_2, \dots, x_n$ is a set of n observations. Then

$$\text{H. M.}=\frac{1}{(1/n)\times\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}=\frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}=\frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

Procedure

Discrete Frequency Distribution - Suppose $x_1, x_2, \dots, x_k$ are distinct observations with $f_1, f_2, \dots, f_k$ as the corresponding frequencies. Then

$$\text{H. M.}=\frac{1}{(1/\sum_{i=1}^{k}f_i)\times\left(\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_k}{x_k}\right)}=\frac{\sum_{i=1}^{k}f_i}{\left(\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_k}{x_k}\right)}=\frac{\sum_{i=1}^{k}f_i}{\sum_{i=1}^{k}\left(\frac{f_i}{x_i}\right)}=\frac{n}{\sum_{i=1}^{k}\left(\frac{f_i}{x_i}\right)}$$

Procedure

Continuous Frequency Distribution - Suppose $C_1, C_2, \dots, C_k$ are k class intervals with $f_1, f_2, \dots, f_k$ as the corresponding frequencies. Let $x_1, x_2, \dots, x_k$ be the mid-values of the class intervals. Then

$$\text{H. M.}=\frac{1}{(1/\sum_{i=1}^{k}f_i)\times\left(\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_k}{x_k}\right)}=\frac{\sum_{i=1}^{k}f_i}{\left(\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_k}{x_k}\right)}=\frac{\sum_{i=1}^{k}f_i}{\sum_{i=1}^{k}\left(\frac{f_i}{x_i}\right)}=\frac{N}{\sum_{i=1}^{k}\left(\frac{f_i}{x_i}\right)}$$

Limitations

1. The serious drawback of the H.M. is that it cannot be defined if any of the observations is zero.
2. It is not simple to understand and calculate.
3. It is not applicable to qualitative data.
4. It cannot be determined graphically.
5. It cannot be computed if frequency distribution includes open end class.
6. It may not be an actual observation in the data.

Result

Ordering of A.M., G.M., H.M.

A.M., G.M. and H.M are ordered in the following way for any data.

$$A.M.\geq G.M.\geq H.M.$$

Relation between A.M., G.M. and H.M.

A.M., G.M. and H.M are related in the following way for any data.

$$G.M.=\sqrt{A.M.\times H.M.}$$

1.1.4 Median

Definition: Median - Median is the value of the middle most observation in the data when the observations are arranged in either increasing or decreasing order of their values.

Remarks

1. Median is the central observation.
2. Median divides the data into two equal parts. There are equal number of observations above as well as below the Median.
3. It is also called as positional average.
4. It is also a trimmed mean obtained by ignoring $(n-1)/2$ topmost and $(n-1)/2$ lowest observations.

Procedure

Ungrouped Data - Suppose n is the number of observations in the data. Let us arrange these n observations in either increasing or decreasing order. Then

$$\text{Median }(Me)=\begin{cases}\text{Value of }\left(\frac{n+1}{2}\right)^{th}\text{ observation} & \text{, if } n \text{ is odd}\\ \frac{(\text{Value of }(n/2)^{th}\text{ observation})+(\text{Value of }((n/2)+1)^{th}\text{ observation})}{2} & \text{, if } n \text{ is even}\end{cases}$$

Procedure

Continuous Frequency Distribution - Let us consider a continuous variable with corresponding frequency distribution with N as the total frequency. Then regardless of whether the N is odd or even the median is defined as the value of the $(N/2)^{th}$ observation.

Definition: Median Class - It is the class in which median i.e. $(N/2)^{th}$ observation falls. In other words, it is in a class where less than cumulative frequency is equal to or exceeds $N/2$ for the first time.

Steps

1. Obtain the class boundaries.
2. Obtain less than cumulative frequencies.
3. Locate the Median class.
4. Apply the formula and find the median

$$\text{Median } (Me) = l + \frac{(N/2) - c.f.}{f} \times h$$

where

$l =$ Lower boundary of the median class.

$N =$ Total frequency

$c.f. =$ Less than cumulative frequency of the class just preceding the median class.

$f =$ Frequency of median class.

$h =$ Class width

Limitations

1. It is not based on all the observations, hence it is not proper representative.
2. It is not capable of further mathematical treatment.
3. It is not as rigidly defined as the Arithmetic Mean.

1.1.5 Mode

Definition: Mode - The observation with maximum frequency or the most repeated observation is called as Mode.

Procedure

Individual Observations and Discrete Frequency Distribution - In this case we can find the observation with the largest frequency just by inspection. If the largest frequency occurs twice (or more), then we say there are two (or many) modes.

Procedure

Continuous Frequency Distribution -

Definition: Modal Class – A Mode lies in the class with maximum frequency, which is known as the Modal Class.

Steps

1. Obtain the class boundaries.
2. Locate the modal class.
3. Apply the formula and find the mode

$$\text{Mode } (Mo) = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

where

$l =$ Lower boundary of the modal class.

$f_m =$ Frequency of modal class.

$f_1 =$ Frequency of pre-modal class.

$f_2 =$ Frequency of post-modal class.

$h =$ Width of modal class.

Limitations

1. It is not based on all the observations, hence it is not proper representative.
2. It is not capable of further mathematical treatment.
3. It is not as rigidly defined as the Arithmetic Mean.
4. It is indeterminate if the modal class is at the extreme of the distribution.

Note: It is possible to have two modes, such frequency distribution is called as bimodal frequency distribution. Sometimes bimodal frequency distribution is an indication of mixture of two frequency distribution.

1.1.6 Weighted Averages

Definition: Weight - A device of giving due or proper importance to the observation is called as Weight.

Remarks

Weight is a positive number. Further unlike that of the frequency the weight can be fractional.

Procedure

The weighted Averages are computed by simply replacing the weights in place of the frequencies in the formulas of the corresponding averages.

Procedure

Weighted Arithmetic Mean (Wgt. A.M.)

$$\overline{x}_{\text{w}} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$$

Procedure

Geometric Mean (Wgt. G.M.)

$$G.M._w = \left(x_1^{w_1} \times x_2^{w_2} \times \ldots \times x_n^{w_n}\right)^{\left(\frac{1}{\sum_{i=1}^{n} w_i}\right)}$$

$$= \left(\prod_{i=1}^{n} x_i^{w_i}\right)^{\left(\frac{1}{\sum_{i=1}^{n} w_i}\right)}$$

Procedure

Harmonic Mean (Wgt. H.M.)

$$\text{H. M.}_{\text{w}} = \frac{1}{(1/\sum_{i=1}^{n} w_i) \times \left(\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_n}{x_n}\right)}$$

$$= \frac{\sum_{i=1}^{n} w_i}{\left(\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_n}{x_n}\right)} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n}\left(\frac{w_i}{x_i}\right)} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n}\left(\frac{w_i}{x_i}\right)}$$

Remark

If all the weights are same, then the weighted averages reduce to unweighted or simple averages.

Note

The combined Arithmetic Mean formula for k groups given their sizes and Arithmetic Means is an example of Weighted Arithmetic Mean.

$$\overline{x}_c = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \cdots + n_k\overline{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i\overline{x}_i}{\sum_{i=1}^{k} n_i}$$

Here the weights are the group sizes.

1.2 Measures of Dispersion

Average condenses data into a single value. However, Average alone is not sufficient to describe the data completely. There may be two data with same means but data may not be identical. For further study and analysis, it becomes essential to measure the extent of variation.

Definition: Dispersion - The variation in data is nothing but the extent to which the observations in the data are deviated from the Average under consideration. This variation is called as dispersion.

Note

Average remains good representative if dispersion is less. Thus, dispersion decides the reliability of Average.

Definition: Measure of Dispersion - To measure the Dispersion present in data with respect to the particular average a non-negative measure is defined using the data and that particular measure which is known as the Measure of Dispersion.

Note

Measure of Dispersion possess units and hence create difficulty in comparison of dispersion for two or more frequency distributions. Therefore, comparison is not possible until a unit less quantity is available.

Definition: Measure of Dispersion for Comparison - To overcome the difficulty stated in the above note with respect to every such a Measure of Dispersion, Measure of Dispersion for Comparison is defined.

Definition: Coefficient of Measure of Dispersion - Measure of Dispersion for Comparison obtained by dividing the Measure of Dispersion by corresponding Average is known as Coefficient of the respective Measure of Dispersion.

Requisites

Requisites of Good Average –

1. It should be simple to understand and easy to calculate.

2. It should be rigidly defined.

3. It should be based on all observations in the data.

4. It should be capable of further mathematical treatment.

5. It should be least affected by extreme observations.

6. It should possess sampling stability.

1.2.1 Range and Its Coefficient

Definition: Range - If L is the largest observation and S is the smallest observation then Range is the difference between L and S. Thus

$$\text{Range} = L - S$$

Definition: Coefficient of Range - If L is the largest observation and S is the smallest observation. Then

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

Note

Range is a crude measure of dispersion. However, it is the simplest measure and suitable if the extent of variation is small.

Remark

In case of frequency distribution mid-values of first and last class intervals are taken to be the largest and the smallest observations respectively.

Limitations

1. Range is not based on all observations. It uses only two extreme observations. Hence, any change in the in between observations is not going to affect the Range.
2. In many situations extreme observations are widely separated from the remaining observations. In this situation Range will overestimate the Dispersion.

1.2.2 Mean Deviation and Its Coefficient

Definition: Mean Deviation about A Point - The arithmetic mean of absolute deviations taken from any arbitrary real point is called as the Mean Deviation about that point.

Formula

Mean Deviation (M.D.) about Value -

$$\text{M. D.}(a) = \begin{cases} \dfrac{\sum_{i=1}^{n}|x_i - a|}{n} & \text{for individual observations} \\ \dfrac{\sum_{i=1}^{k} f_i\,|x_i - a|}{N = \sum_{i=1}^{k} f_i} & \text{for frequency distribution} \end{cases} ;\ a \in \mathbb{R}$$

Formula

Coefficient of Mean Deviation (M.D.) about Value -

$$\text{Coefficient of M. D. about A Point} = \frac{\text{M. D. about A Point}}{\text{The Point}}$$

Property

Minimality Property of Mean Deviation - Among all mean deviations, Mean Deviation about Median is minimum.

Thus to avoid the effect of choice of point, Mean Deviation about Median is preferred.

Limitations

1. It is not applicable for qualitative data.
2. Since algebraic signs of deviations are ignored, it is not applicable for further mathematical treatment.
3. It cannot be computed for the frequency distribution with open end class.

1.2.3 Mean Square Deviation and Its Coefficient

Definition: Mean Square Deviation about A Point - The arithmetic mean of squared deviations from any arbitrary real point is called as the Mean Square Deviation about that Point.

Formula

Mean Square Deviation (M.S.D.) about A Point -

$$\text{M. S. D.}(a) = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i - a)^2}{n} & \text{for individual observations} \\ \dfrac{\sum_{i=1}^{k} f_i\,(x_i - a)^2}{N = \sum_{i=1}^{k} f_i} & \text{for frequency distribution} \end{cases} ;\ a \in \mathbb{R}$$

Formula

Coefficient of Mean Squared Deviation (M.S.D.) about A Point -

$$\text{Coefficient of M. S. D. about A Point} = \frac{\text{M. S. D. about A Point}}{\text{The Point}}$$

Property

Minimality Property of Mean Square Deviation - Among all mean square deviations, Mean Square Deviation about Mean is minimum.

Thus, to avoid the effect of choice of point, Mean Square Deviation about Mean is preferred.

1.2.4 Variance and Its Coefficient

Notation

The Variance of x is symbolically written as $Var(x)$ or σ^2.

Definition: Variance - The arithmetic mean of squares of deviations taken from arithmetic mean is called as Variance.

$$\text{Variance or } \sigma^2 = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} & \text{for individual observations} \\ \dfrac{\sum_{i=1}^{k} f_i\,(x_i - \overline{x})^2}{N = \sum_{i=1}^{k} f_i} & \text{for frequency distribution} \end{cases}$$

Remark

The units of original items and that of the Variance are not same. The unit of Variance is square of the original unit of the item.

Notation

The Standard Deviation of x is symbolically written as $S.D.(x)$ or σ.

Definition: Standard Deviation (S.D.) - The positive square root of the mean of the deviations taken from the arithmetic mean which is nothing but Variance is called as Least Root Mean Square Deviation or Standard Deviation (S.D.).

$$\text{S. D. or } \sigma = \begin{cases} \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} & \text{for individual observations} \\ \sqrt{\dfrac{\sum_{i=1}^{k} f_i\,(x_i - \overline{x})^2}{N = \sum_{i=1}^{k} f_i}} & \text{for frequency distribution} \end{cases}$$

Computational Formula

Variance -

$$\text{Variance or } \sigma^2 = \begin{cases} \dfrac{\sum_{i=1}^{n} {x_i}^2}{n} - (\overline{x})^2 & \text{for individual observations} \\ \dfrac{\sum_{i=1}^{k} f_i\,{x_i}^2}{N = \sum_{i=1}^{k} f_i} - (\overline{x})^2 & \text{for frequency distribution} \end{cases}$$

Standard Deviation (S.D.) -

$$\text{S. D. or } \sigma = \begin{cases} \sqrt{\dfrac{\sum_{i=1}^{n} {x_i}^2}{n} - (\overline{x})^2} & \text{for individual observations} \\ \sqrt{\dfrac{\sum_{i=1}^{k} f_i\,{x_i}^2}{N = \sum_{i=1}^{k} f_i} - (\overline{x})^2} & \text{for frequency distribution} \end{cases}$$

Coefficient of Variation (C.V.) -

$$C.V. = \frac{\text{S. D.}}{|\text{ A. M. }|} \times 100 = \frac{\sigma}{|\overline{x}|} \times 100$$

Note

Coefficient of Variation (C.V.) is always expressed in percentage. That is it is free from any unit.

Property

Effect of Change of Origin and Scale - If

$$u_i = \frac{x_i - b}{k} \Rightarrow Var(u_i) = \frac{Var(x_i)}{k^2} \Rightarrow S.D.(u_i) = \frac{S.D.(x_i)}{k} \; ; \; b, k \in \mathbb{R}$$

Property

Combined Variance and S.D. - Suppose there are two groups. First is of size n_1 with Arithmetic Mean $\overline{x}_1$ and Variance σ_1^2. Second group is of size n_2 with Arithmetic Mean $\overline{x}_2$ and Variance σ_2^2. Then the Variance of combined group of size $n_1 + n_2$ is given by

$$\sigma_c^2 = \frac{n_1(\sigma_1^2 + (\overline{x}_1 - \overline{x}_c)^2) + n_2(\sigma_2^2 + (\overline{x}_2 - \overline{x}_c)^2)}{n_1 + n_2}$$

where $\overline{x}_c$ is combined arithmetic mean.

Property

Generalization - Let there be $k(k \geq 2)$ groups with size of i^{th} group as n_i, Arithmetic Mean $\overline{x}_i$ and Variance $\sigma_i^2, i = 1,2,3, \ldots, k$. The combined variance of k groups is given by

$$\sigma_c^2 = \frac{\sum_{i=1}^{k} n_i(\sigma_i^2 + (\overline{x}_i - \overline{x}_c)^2)}{\sum_{i=1}^{k} n_i}$$

where $\overline{x}_c$ is combined arithmetic mean.

Property

$$\text{S. D.} \geq \text{M. D. about Arithmetic Mean}$$

Limitations

1. It is difficult to understand and calculate.
2. It cannot be computed for a frequency distribution with open-end class.
3. It is unduly affected due to extreme deviations.
4. It cannot be calculated for qualitative data.

1.2.5 Quantiles or Partition Values

Definition: Percentiles - The observations $P_1, P_2, \ldots, P_{99}$ which divide the total number of observations into 100 equal parts are called Percentiles.

Definition: Deciles - The observations $D_1, D_2, \ldots, D_9$ which divide the total number of observations into 10 equal parts are called Deciles.

Definition: Quartiles - The observations Q_1, Q_2, Q_3 which divide the total number of observations into 4 equal parts are called Quartiles.

Note

$$Q_1 = P_{25} \; ; Q_2 = D_5 = P_{50} = \text{Median} \; ; Q_3 = P_{75}$$

Procedure

Individual Observations or Ungrouped Data - To obtain the partition values of series of individual observations, many calculations or formulae are not required. Many of the times rearrangement of data in either increasing or decreasing order and simple calculations are sufficient to get these values.

Procedure

Continuous Frequency Distribution - Let us consider a continuous variable with corresponding frequency distribution with N as the total frequency. To get the partition values for this continuous frequency distribution, corresponding formula of Median can be suitably modified. Some common steps to obtain the partition values are as follows.

Steps

Percentile

1. Obtain the class boundaries.

2. Obtain less than cumulative frequencies.

3. Locate the class in which the partition value lies.

4. Apply the formula and find the partition value.

The different formulas for the different partition values and common notations used are as follows.

$$k^{th} \text{ Percentile} = l + \frac{(kN/100) - c.f.}{f} \times h \, , k = 1,2, \ldots, 99$$

$$k^{th} \text{ Decile} = l + \frac{(kN/10) - c.f.}{f} \times h \, , k = 1,2, \ldots, 9$$

$$k^{th} \text{ Quartile} = l + \frac{(kN/4) - c.f.}{f} \times h \, , k = 1,2,3$$

where

$l =$ Lower boundary of the class in which the k^{th} partition value lies.

$N =$ Total frequency

$c.f. =$ Less than cumulative frequency of the class just preceding the class in which the k^{th} partition value lies.

$f =$ Frequency of class in which the k^{th} partition value lies.

$h =$ Class width.

(CSIR NET JUNE 2012)

Q. Suppose that we have a data set consisting of 25 observations, where each value is either 0 or 1. Then which of the following is/are correct?

(a) The mean of the data cannot be larger than the variance.

(b) The mean of the data cannot be smaller than the variance.

(c) The mean being same as the variance implies that the mean is zero.

(d) The variance will be 0 if and only if the mean is either 1 or 0.

Explanation:

The correct answers are **(b)** and **(d)**. Here's the explanation:

(b) The mean of the data cannot be smaller than the variance.
In a binary dataset (values 0 or 1), the variance is given by Var $= \mu(1-\mu)$, where μ is the mean. Since $\mu(1-\mu) \leq \mu$, the variance cannot exceed the mean, meaning the mean cannot be smaller than the variance.

(d) The variance will be 0 if and only if the mean is either 1 or 0.
If all the observations are the same (either all 0s or all 1s), there is no variation in the data, and hence the variance is 0. In such cases, the mean is either 0 (all observations are 0) or 1 (all observations are 1).

(a) is incorrect: The mean can be larger than the variance; for example, if $\mu = 0.8$, the variance will be $0.8 \times 0.2 = 0.16 < 0.8$.

(c) is incorrect: If the mean equals the variance ($\mu = \mu(1-\mu)$), solving gives $\mu = 0$ or $\mu = \frac{1}{2}$, not necessarily 0.

(CSIR NET DEC. 2015)

Q. A set of N observations resulted in k distinct values $x_1, x_2, \ldots, x_k$ with respective frequencies $f_1, f_2, \ldots, f_k$, so that $\sum_{i=1}^{k} f_i = N$. Another k observations resulted in observations $x_1, x_2, \ldots, x_k$ once each, so that the modified (new) sample of $N + k$ has observation x_i with frequency $f_i + 1$.

(a) The new mean is necessarily less than or equal to the original mean

(b) The new median is necessarily more than or equal to the original median

(c) The new variance is necessarily less than or equal to the original variance

(d) The new mode will be same as the original mode.

Explanation:

We analyse the given question step-by-step to evaluate the statements regarding the new mean, median, variance, and mode.

Observations:

1. Initially, the data set consists of N observations: $x_1, x_2, \ldots, x_k$, with frequencies $f_1, f_2, \ldots, f_k$, such that $\sum_{i=1}^{k} f_i = N$.
2. In the modified data set, k additional observations are added, each corresponding to $x_1, x_2, \ldots, x_k$, with frequency incremented by 1 for each value. The total number of observations is now $N + k$, and the new frequencies are $f_i + 1$.

Statement Analysis:

(a) The new mean is necessarily less than or equal to the original mean.

- The new total sum of observations is: New sum $=$ Old sum $+ \sum_{i=1}^{k} x_i$.
- The new total number of observations is $N + k$. Hence, the new mean is: New mean $= \frac{\text{Old sum} + \sum_{i=1}^{k} x_i}{N+k}$.
- Compared to the original mean, whether the new mean is less than or equal depends on the distribution of x_i values relative to the original mean. This is **not necessarily true** in all cases.

(b) The new median is necessarily more than or equal to the original median.

- Adding k new observations does not alter the order of the data values, but it increases the frequency of each x_i. As such, the median depends on the relative positioning of the frequencies. However, this does **not guarantee** that the new median is greater than or equal to the original median in all cases.

(c) The new variance is necessarily less than or equal to the original variance.

- Variance measures the spread of the data. When adding k observations with the same values as existing ones, the spread decreases because the data becomes more concentrated around its mean. Hence, the **new variance is necessarily less than or equal to the original variance.**

(d) The new mode will be the same as the original mode.

- The mode is the value with the highest frequency. Since the relative ordering of frequencies does not change (all frequencies increase by 1), the mode remains the same. Hence, the **new mode will be the same as the original mode.**

  Therefore, options (c) and (d) are correct.

1.3 Moments

Definition: Moment about 'a' - Suppose 'a' is any arbitrary constant, then r^{th} Moment about a is denoted by $\mu_r(a)$ and it is defined as follows

$$\mu_r(a) = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i - a)^r}{n} & \text{for individual observations} \\ \dfrac{\sum_{i=1}^{k} f_i\,(x_i - a)^r}{N = \sum_{i=1}^{k} f_i} & \text{for frequency distribution} \end{cases} ;\ a \in \mathbb{R}$$

Definition: Moment about Zero or Raw Moment - The r^{th} order Raw Moment is denoted by μ_r' and it is defined as follows

$$\mu_r' = \begin{cases} \dfrac{\sum_{i=1}^{n} {x_i}^r}{n} & \text{for individual observations} \\ \dfrac{\sum_{i=1}^{k} f_i\,{x_i}^r}{N = \sum_{i=1}^{k} f_i} & \text{for frequency distribution} \end{cases}$$

Definition: Moment about '$\overline{x}$' or Central Moment - The r^{th} order Central Moment is denoted by μ_r and it is defined as follows

$$\mu_r = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^r}{n} & \text{for individual observations} \\ \dfrac{\sum_{i=1}^{k} f_i\,(x_i - \overline{x})^r}{N = \sum_{i=1}^{k} f_i} & \text{for frequency distribution} \end{cases}$$

Property

Effect of Change of Origin and Scale - If

$$u_i = \frac{x_i - b}{k} \Rightarrow \mu_r \text{ of } u_i = \frac{\mu_r \text{ of } x_i}{k^r} ;\ b, k \in \mathbb{R}$$

Relation between Raw and Central Moments -

$$\mu_r = \mu_r' - \binom{r}{1}\mu_{r-1}'\mu_1' + \binom{r}{2}\mu_{r-2}'{\mu_1'}^2 - \binom{r}{3}\mu_{r-3}'{\mu_1'}^3 + \cdots + (-1)^r{\mu_1'}^r$$

Skewness and Kurtosis

1.3.1 Skewness

Symmetry

A frequency distribution is symmetric about a value 'a', if the corresponding frequency curve is symmetric about 'a'.

Properties

1. In case of bell shaped unimodal symmetric frequency distributions, arithmetic mean, mode, median coincide.
2. The quartiles of symmetric distributions are equispaced. By that we mean $Q_3 - Q_2 = Q_2 - Q_1$.
3. The odd ordered central moments of symmetric distribution are zero. That is
$$\mu_1 = \mu_3 = \mu_5 = \cdots = 0$$
4. Frequencies of classes equidistant from central class on either side are same.

Definition: Skewness - Skewness is the lack of Symmetry or Departure from Symmetry.

If the distribution is skew the corresponding frequency curve is elongated on either side.

Definition: Positive Skewness - If the frequency curve is elongated towards right side, then the distribution is said to possess Positive Skewness. That is in this case frequency increases rapidly to reach the maximum and further decreases slowly.

Remark

In case of positively skew distribution we observe that

$$\text{Mode} < \text{Median} < \text{Arithmetic Mean}$$

Definition: Negative Skewness - If the frequency curve is elongated towards left side, then the distribution is said to possess Negative Skewness. That is in this case frequency increases slowly to reach the maximum and further decreases rapidly.

Remark

In case of negatively skew distribution we observe that

$$\text{Arithmetic Mean} < \text{Median} < \text{Mode}$$

Karl Pearson's Coefficient of Skewness

Notation

Karl Pearson's Coefficient of Skewness is denoted by S_K.

Definition

$$S_K = \frac{A.M. - Mode}{\sigma}$$

Conclusion

If

$S_K < 0$, the distribution is Negatively Skew.

$S_K = 0$, the distribution is Symmetric.

$S_K > 0$, the distribution is Positively Skew.