DegreeLive

B.COM HONOURS	SEMESTER 6	CALICUT UNIVERSITY	COURSE CODE: COM6FS113(3)

DATA ANALYSIS USING STATISTICAL SOFTWARE

EXAM REVISION NOTES

JOIN WHATSAPP CHANNEL →

MODULE I

INTRODUCTION TO STATISTICAL SOFTWARE

Research Planning Measurement Foundations: Quantitative analysis does not commence with software execution; it originates in disciplined conceptual planning.

Empirical Research Planning Sampling Design: Scientific inquiry begins with precise methodological planning. The investigator must establish the research philosophy, temporal horizon, and sampling framework before collecting data:

- **Cross-Sectional vs. Longitudinal Designs:** , assessing investor sentiment during a specific market correction).
- **Probability Sampling Architectures:** Every population element possesses a known, non-zero mathematical probability of selection: Simple Random Sampling: Each unit in the sampling frame has an equal probability of selection using random number generators.
- **Stratified Random Sampling:** , tier-1 cities vs.
- **Cluster Sampling:** , bank branches or college campuses); entire clusters are randomly sampled.
- **Non-Probability Sampling Techniques:** , active derivative traders executing at least 10 trades per month).

Scaling Techniques in Business Research

- **Semantic Differential Scales:** Developed by C.

Psychometric Evaluation: Reliability and Validity: To produce scientifically generalizable findings, empirical instruments must demonstrate psychometric robustness across two dimensions: Reliability (consistency of measurement) and Validity (accuracy of measurement).

Forms of Instrument Validity

- **Content / Face Validity:** The qualitative judgment of whether the questionnaire items adequately cover all representative facets of the conceptual domain.
- **Construct Validity:** Confirms whether the operationalized questionnaire items genuinely reflect the theoretical latent construct.
- **Discriminant Validity:** The degree to which measures of distinct constructs are empirically unique and show low correlation with each other.
- **Criterion-Related Validity:** The degree to which an instrument correlates with an established external outcome criterion.

Questionnaire Engineering Wording Rules: Designing a structured survey questionnaire requires translating research objectives into clear, objective, and unambiguous inquiry items.

Best-Practice Principles for Question Wording

- **Double-Barreled Questions:** ; Split into two discrete items: (1) "Rate the processing speed of the app" and (2) "Rate the ease of navigation of the app".
- **Leading / Biased Wording:** "
- **Ambiguous Vague Terms:** "
- **Jargon Acronym Overuse:** "; Use accessible language or spell out technical terms clearly for retail participants.

Preparing a Codebook for Data Analysis: A Codebook is a comprehensive technical document that translates raw survey questionnaires into a structured numerical layout ready for statistical processing.

Comprehensive Research Codebook Architecture

- **Respondent_ID:** Unique Identification Number; 1 to 500 (Sequential integers)
- **Gender:** Biological Gender of Investor; 1 = Male2 = Female3 = Non-Binary / Other; 99 = No Response
- **Age_Years:** , 24, 45, 62); 999 = Unreported
- **Edu_Level:** Highest Academic Qualification; 1 = Higher Secondary2 = Bachelor Degree3 = Master Degree4 = Doctorate; 9 = Not Stated

Coding Open-Ended Responses: "), researchers apply a rigorous post-coding process:

- **Step 1:** Thematic Scanning: The researcher reviews a representative sample of 50 to 100 responses to identify recurrent conceptual patterns.
- **Step 2:** , 1 = Fear of Financial Ruin, 2 = Lack of Technical Knowledge, 3 = High Capital / Margin Requirements, 4 = Regulatory Restrictions, 5 = Other).
- **Step 3:** Numerical Assignment: Each narrative text response is reviewed and converted into its corresponding numerical code for data entry in SPSS.

The Tripartite Window Architecture of SPSS

Data View vs. Variable View in Data Editor

- **Data View:** , Row 14 contains all data for Respondent #14).
- , Column 1 is ID, Column 2 is Gender, Column 3 is Age).

The Syntax Workflow and Dialogue Box "Paste" Button: In almost every statistical dialogue box in SPSS, beside the standard OK, Cancel, and Help buttons, there exists a vital command button: Paste.

- **The Power of the Paste Command:** Clicking 'Paste' instead of 'OK' does not execute the analysis immediately.
- **Audit Trails Reproducibility:** In empirical commerce research, journal peer-reviewers often require exact computational replication.
- **Batch Execution:** Highlight the syntax block and click the green 'Play' button (or press Ctrl + R) to re-run complex models across updated datasets in milliseconds.

Summary Synthesis

- **Measurement Scales:** ; Determines the mathematical validity of descriptive and inferential parametric tests.
- **Reliability:** ; Ensures survey scale items are statistically coherent and free from random error.
- **Validity:** ; Guarantees that measurement items measure the true theoretical construct accurately.
- **Questionnaire Design:** ; Eliminates structural bias and respondent confusion at the data collection stage.

Empirical Casework: Forensic Analysis, Audit Trail Verification, and Internal Controls: Empirical accounting investigations across modern industrial enterprises illustrate that financial distortion, fraud, and misstatement rarely occur as isolated incidents; rather, they stem from structural vulnerabilities within accounting workflows, segregation-of-duties lapses, or inadequate internal audit scrutiny.

- **Ledger amp; Voucher Integrity:** ; Companies Act (Sec.
- **Revenue amp; Cost Allocation:** ; Ind AS 115 (Revenue from Contracts with Customers) amp; CAS 1 to 24 Cost Accounting Standards.
- **Fiscal Risk amp; Tax Compliance:** ; GST Law Section 16/36 Rules amp; Income Tax Act Transfer Pricing Provisions.
- **Working Capital Governance:** ; Tandon amp; Chore Committee Norms for Bank Working Capital Credit Appraisal.

Comparative Regulatory and Practical Implementation Standards: A rigorous examination of corporate commercial practice highlights the indispensable role of standardized accounting principles.

DegreeLive

MODULE II

PREPARING THE DATA FILE

Data Architecture Variable Definition in SPSS: sav) do not merely hold raw numerical values; they embed an intricate metadata dictionary that defines the structural characteristics, display formatting, and analytical roles of each measurement column.

Comprehensive Analysis of the Eleven Variable Properties

- **1. Name:** The unique syntax variable identifier.
- **2. Type:** Specifies the data storage format.
- **String:** , respondent names, open-ended qualitative comments).
- **Date:** , dd-mmm-yyyy, mm/dd/yyyy).
- **Custom Currency:** Numeric variables formatted with currency symbols and comma thousand separators.

Operational Shortcuts: Copying Variable Properties: Manually entering identical Value Labels and Missing codes across 40 or 50 Likert scale items is inefficient and error-prone.

- **Standard Copy-Paste Method:** , Sat_1).
- **Define Variable Properties Utility:** Located at Data -> Define Variable Properties.
- **Apply Data Dictionary Wizard:** sav) into an unformatted new dataset.

Structural File Modifications

- **Inserting Deleting Cases (Rows):** To insert an omitted respondent between existing records, right-click on the row number header and select Insert Cases.
- **Inserting Deleting Variables (Columns):** To introduce an omitted survey item, right-click the column header and select Insert Variable.
- **Sorting Cases:** Navigating to Data -> Sort Cases allows the researcher to sort observations based on one or multiple criteria.

Splitting Datasets: Data -> Split File

Filtering Datasets: Data -> Select Cases

Transforming Variables: Recoding into Different Variables: Recoding involves transforming existing data values into new categories or numeric groupings. In empirical commerce research, researchers routinely convert continuous age into generational cohorts, or invert reverse-worded Likert statements.

Deriving Composite Scales: The Compute Variable Command: Psychometric research rarely relies on single individual survey questions. Instead, multiple Likert items measuring the same theoretical construct are aggregated into a single composite Sum of mated score using Transform -> Compute Variable.

- **Mean Score vs. Summated Total Score:** Researchers can compute either the arithmetic mean or the raw Sum of across construct items.
- **Handling Missing Data during Computation:**) to the entire composite score.
- **Using Statistical Functions (MEAN.n / SUM.n):** 3(Item1, Item2, Item3, Item4) instructs SPSS to compute the mean provided the respondent answered at least 3 out of the 4 items, mitigating unnecessary case attrition.

Importing Microsoft Excel Files into SPSS: xlsx) launches the Excel Import Wizard. Successful importation requires observing critical structural conditions:

- **First Row Variable Names:** Check the box "Read variable names from first row of data".
- **Worksheet Selection:** If the Excel workbook contains multiple tabs, select the exact target sheet from the dropdown menu.
- **Range Specification:** , A1:AZ500).
- **Data Type Homogeneity:** Ensure every column in Excel contains homogeneous data types.

Merging Datasets: Adding Cases vs. Adding Variables: Large-scale research projects frequently necessitate merging disparate empirical files. SPSS provides two fundamentally distinct merging mechanisms under the Data -> Merge Files menu:

Merging Files: Add Variables (Column Joining)

Data Screening, Quality Assurance Data Cleaning: Data cleaning is the rigorous quality assurance process executed prior to conducting any formal hypothesis testing.

Screening Continuous Variables: Descriptive Statistics Explore: For continuous scale variables (Age, Annual Income, Monthly Portfolio Turnover), frequency tables are inefficient because every unique number produces a separate row.

- **Procedure:** Navigate to Analyze -> Descriptive Statistics -> Descriptives (or the powerful Analyze -> Descriptive Statistics -> Explore command).
- **Minimum and Maximum Scrutiny:** In the descriptive output table, carefully inspect the Minimum and Maximum columns: If Age_Years exhibits a Minimum of 2 or a Maximum of 450, gross data entry errors exist.
- If 5-point Likert items display values of 0, 6, or 55, out-of-bounds errors must be flagged.
- **Mean vs. 5% Trimmed Mean:** The Explore procedure calculates both the standard Arithmetic Mean and the 5% Trimmed Mean (which removes the top 5% and bottom 5% extreme values).

Detecting Duplicate Records: When collecting data via online electronic forms, respondents occasionally submit their survey multiple times due to slow internet connections or accidental double-clicking.

- **Procedure:** Navigate to Data -> Identify Duplicate Cases.
- **Output Indicator:** SPSS creates a new binary system variable named PrimaryLast or DuplicateIndicator (where 0 = Duplicate Case and 1 = Primary Unique Case).

Univariate Outlier Detection via Standardized Z-Scores: An outlier is an observation that deviates markedly from the overall distribution of data. In continuous financial and behavioral measures, undetected outliers distort the sample mean and inflate standard deviations, creating false rejections or acceptances of hypotheses:

- **Standardized Z-Score Calculation:** , ZAge, ZIncome) to the active dataset.
- **Z-Score Evaluation Benchmarks:** 01) indicate potential univariate outliers.
- **For large sample sizes (N > 100):** 29 (p are flagged as severe univariate outliers.
- **Treatment Protocols:** Analysts must decide whether to (1) verify the original source for typographical entry error, (2) delete the outlier case if it represents an unrepresentative population element, or (3) apply Winsorization (re-coding the extreme value to the next highest non-outlying score at the 99th percentile).

Missing Data Mechanism Classification: Little's MCAR Test: Understanding the underlying statistical mechanism driving missing data determines which treatment protocol is methodologically valid:

- **Missing Completely at Random (MCAR):** , a questionnaire page was accidentally dropped by a scanner).
- **Missing at Random (MAR):** , older investors are less likely to report online trading frequency, but conditional on age, missingness is random).
- **Missing Not at Random (MNAR):** , ultra-high-income individuals refusing to disclose income due to tax avoidance sensitivity).

MODULE III

PRELIMINARY ANALYSIS

Descriptive Statistics: Categorical Continuous Exploration: Descriptive statistics synthesize large volumes of raw empirical data into concise, interpretable numerical Sum of maries.

Analyzing Categorical Variables: The Frequencies Procedure: When analyzing categorical variables such as Gender, Investment Platform Preference, Educational Level, or Geographic Sector, researchers generate frequency tables in SPSS.

- **Frequency:** The absolute count of respondents falling within each discrete response category.
- **Percent:** The raw percentage calculated against the Total Sample Size (N), including respondents who skipped the question or had missing data.
- **Valid Percent:** The percentage calculated strictly against the number of Valid Cases (excluding user-missing codes and system-missing blank cells).
- **Cumulative Percent:** The progressive running total of valid percentages Sum of med from top to bottom.
- **Statistics Sub-dialog:** While primarily intended for continuous variables, researchers can request Quartiles, Cut points for equal groups, Percentiles, Mean, Median, Mode, Standard Deviation, Variance, Skewness, and Kurtosis directly within the Frequencies run.

The Descriptives Procedure and Automatic Standardized Z-Scores: The Descriptives procedure (Analyze -> Descriptive Statistics -> Descriptives) is optimized for quickly generating Sum of mary tables across multiple continuous variables simultaneously.

Assessing Distributional Normality Outliers: Most premier inferential statistical procedures—including Independent Samples T-Tests, Paired T-Tests, One-Way Analysis of Variance (ANOVA), Pearson Correlation, and Linear Regression—are parametric tests that rest on the critical assumption that continuous variables follow an approximate Normal Distribution (Gaussian bell-shaped curve).

Statistical Hypothesis Tests of Normality: Kolmogorov-Smirnov Shapiro-Wilk: Navigating to Analyze -> Descriptive Statistics -> Explore -> Plots and checking Normality plots with tests generates two rigorous formal hypothesis tests of normality: the Kolmogorov-Smirnov Test and the Shapiro-Wilk Test.

- Kolmogorov-Smirnov (K-S) Test with Lilliefors Significance Correction: Suitable for large samples (N > 50).
- **Shapiro-Wilk (S-W) Test:** Widely recognized by biometricians and econometricians as the most statistically powerful normality test, particularly for small-to-moderate sample sizes (N < 50), calculated from order statistics and regression of order statistics on normal quantiles.

Comprehensive Outlier Diagnostics and Treatment: An outlier is an empirical observation that lies an abnormal distance from other values in a random sample. Outliers distort arithmetic means, artificially inflate standard deviations, compromise linear regression coefficients, and produce Type I or Type II inferential errors.

Multivariate Outlier Detection: Mahalanobis Distance: , an individual reporting an Age of 18 with an Annual Income of INR 50,00,000 and 25 years of investment experience).

Graphical Data Representation The SPSS Chart Editor: Data visualization is an essential analytical skill that communicates complex empirical trends and relationships transparently.

Bar Charts (Categorical Distributions)

Boxplots (Tukey's Five-Number Summary)

Scatter Plots (Bivariate Relationships)

Exporting Outputs into Word Documents: Transferring output tables and charts into Microsoft Word for academic publication can be achieved through multiple methods:

- **Copy as Image / Enhanced Metafile (EMF):** Right-click the chart or table in the Output Viewer navigation tree, select Copy Special, and choose Enhanced Metafile or Image.
- **Copy as Editable Word Table:** Right-click and select Copy.
- **Automated Batch Export:** Navigate to File -> Export in the Output Viewer.
- **Syntax-Driven Exporting (OUTPUT EXPORT):** Researchers can automate document generation directly via syntax, specifying document type, margins, and destination path.

The Theoretical Foundation of Classical Test Theory (CTT): Scale reliability is grounded in Classical Test Theory, which asserts that any observed score (X) obtained from an empirical measurement is composed of two additive components: the true score (T) and random measurement error (E):

- **Test-Retest Reliability:** Administering the same test to the same respondents across two time intervals to evaluate temporal stability over time.
- **Parallel-Forms Reliability:** Administering two equivalent versions of a questionnaire to assess consistency across alternate forms.
- **Split-Half Reliability:** , odd versus even items) and calculating the correlation between halves, adjusted via the Spearman-Brown prophecy formula.
- **Internal Consistency Reliability:** Evaluating the extent to which all items comprising a scale correlate with one another and measure a single unified construct.

Step-by-Step SPSS Execution Protocol

- **Step 1:** Launch Dialogue: Navigate to Analyze -> Scale -> Reliability Analysis.
- **Step 2:** Item Selection: Select all questionnaire items belonging to a single specific construct and transfer them into the Items box.
- **Step 3:** Model Selection: Ensure Alpha is selected in the Model dropdown (the default).
- **Step 4:** Configure Statistical Output: Click the Statistics button.
- **Inter-Item:** Check Correlations.

Resolving Low Alpha: The Role of Reverse-Scoring and Multidimensionality: Researchers frequently encounter unexpectedly low or even negative Cronbach's Alpha values. Two primary culprits explain this phenomenon:

- **Multidimensionality:** Cronbach's Alpha assumes that all items measure a single unidimensional construct.

Academic Reporting: APA Style Formatting Guidelines: To meet high scholarly publication standards, statistical findings from preliminary analyses must be reported in accordance with the American Psychological Association (APA 7th Edition) guidelines:

MODULE IV

OVERVIEW OF DIFFERENT STATISTICAL TECHNIQUES

Correlation Analysis: Bivariate and Partial Association: Correlation analysis evaluates the direction, strength, and statistical significance of the association between two or more metric variables without implying a cause-and-effect relationship.

Pearson Product-Moment Correlation: 00:

- , Advertising Expenditure and Sales Revenue).
- , Interest Rates and Bond Prices).

- **Zero Correlation (r = 0.00):** Indicates the complete absence of any linear association between the two variables.

Partial Correlation: Controlling for Confounding Variables: In business research, a significant bivariate correlation between two variables (X and Y) may be spurious, driven entirely by a common third variable (Z).

Regression Model Building Strategies: SPSS supports multiple distinct entry methods for incorporating predictor variables into the regression equation:

- Standard / Simultaneous Regression (Enter Method): All independent variables are entered into the equation simultaneously.

SPSS Execution Protocol

- **Step 1:** Navigate to Analyze -> Regression -> Linear.
- **Step 2:** Assign the continuous outcome variable into the Dependent box.
- **Step 3:** Assign the predictor variables into the Independent(s) box.
- **Step 4 (Statistics):** Click Statistics.
- **Step 5 (Plots):** Click Plots.

Detailed Empirical Regression Case Illustration: Consider an empirical study examining corporate financial performance. A financial econometrician models Return on Equity (ROE) using three predictors: Financial Leverage (Debt-to-Equity Ratio), Firm Size (Natural Log of Total Assets), and Liquidity (Current Ratio) across N = 200 publicly listed commercial firms:

Exploratory Factor Analysis (EFA) Principal Component Analysis (PCA): Exploratory Factor Analysis (EFA) is a multivariate data-reduction and dimension-exploration technique. In commerce, marketing, and organizational research, researchers often administer survey instruments containing dozens of individual items.

Extraction Methods: PCA vs Common Factor Analysis: Researchers must select the appropriate mathematical extraction method in the Extraction sub-dialog:

- **Principal Component Analysis (PCA):** Analyzes the total variance (common variance + specific variance + error variance) present in the variables.
- 0s on the diagonal of the correlation matrix with communality estimates (squared multiple correlations).

Factor Rotation Matrix Interpretation: Unrotated factor matrices are mathematically optimal but conceptually difficult to interpret because the first factor typically loads heavily on nearly all items.

- **Orthogonal Rotation:** ; Varimax: Maximizes variance of squared loadings, sharpening factor distinctions.
- **Oblique Rotation:** ; Direct Oblimin Promax: Allows factor axes to rotate freely at angles other than 90 degrees.

Comparison of Means: T-Tests and Analysis of Variance: When business researchers wish to investigate whether mean differences across experimental conditions or demographic cohorts are statistically significant, they employ mean comparison tests.

Independent Samples T-Test: , Comparing mean trading returns between Male and Female investors, or comparing job satisfaction between Remote and In-Office employees).

- **Critical Parametric Assumption:** Homogeneity of Variances (the assumption that the variance of the dependent variable is equal across both groups).
- **Levene's Test for Equality of Variances:** 05: Fail to reject H0 rarr; Variances are equal rarr; Read the top row: "Equal variances assumed".
- **p ≤ 0.05:** Reject H0 rarr; Variances are unequal rarr; Read the bottom row: "Equal variances not assumed" (Welch's t-test), which adjusts the degrees of freedom downward to protect against Type I error.
- **Interpreting Significance:** 05, the difference between the two group means is statistically significant.
- **Effect Size (Cohen's d):** Calculated as ($Mean_1$ - $Mean_2$) / Pooled SD.

Paired Samples T-Test (Dependent / Repeated Measures): , Evaluating Financial Knowledge scores Before Training vs After Training). Calculates the difference score for each participant (D = Post - Pre) and tests whether the average difference differs significantly from zero.

Post-Hoc Multiple Comparisons in ANOVA: A statistically significant ANOVA F-test confirms that at least one group mean differs from another, but it does not specify which exact pairs of groups differ.

- **Equal Variances Assumed (Levene's p > 0.05):** Tukey's HSD (Honestly Significant Difference): The premier, balanced post-hoc test when group sample sizes are approximately equal.
- **Scheffe Test:** Highly conservative; recommended when group sample sizes are unequal or complex comparisons are tested.
- **Bonferroni Correction:** Adjusts alpha level per test; ideal for a limited number of planned pairwise contrasts.
- **Duncan's Multiple Range Test:** A stepwise range test that groups means into homogeneous subsets.
- 05): Games-Howell Test: Widely considered the most powerful and robust post-hoc test when group variances and sample sizes are unequal.

Two-Way ANOVA (Factorial Design)

Analysis of Covariance (ANCOVA)

Academic Reporting: APA Style Formatting Guidelines: To adhere to high scholarly publication standards, statistical findings from inferential analyses must be reported in accordance with the American Psychological Association (APA 7th Edition) guidelines:

DegreeLive