DegreeLive

B.Com Honours | Semester VI | Calicut University

Data Analysis Using Statistical Software

Course Code: COM6FS113 (3) • Module 3 Notes

Data Analysis Using Statistical Software
Course Code: COM6FS113 (3) • Module 3 Notes

Module III: Preliminary Analysis

Before an empirical researcher or financial analyst can proceed to complex hypothesis testing, econometric modeling, or multivariate factor extractions, the empirical dataset must undergo rigorous preliminary exploration. Preliminary data analysis accomplishes four paramount scientific objectives: (1) calculating foundational descriptive statistics to summarize sample demographics and behavioral distributions; (2) testing critical parametric distributional assumptions, primarily the assumption of univariate normality and evaluating skewness and kurtosis; (3) identifying, diagnosing, and treating univariate and multivariate outliers that threaten statistical validity; and (4) establishing the internal consistency reliability of multi-item psychometric measurement scales using Cronbach's Alpha. In IBM SPSS Statistics, this preliminary phase integrates the Frequencies, Descriptives, Explore, and Reliability procedures alongside an interactive graphical suite. This module examines the methodology, interpretation, and academic reporting of preliminary statistical procedures.

1. Descriptive Statistics: Categorical & Continuous Exploration

Descriptive statistics synthesize large volumes of raw empirical data into concise, interpretable numerical summaries. The appropriate choice of descriptive metrics depends on whether the variable under investigation is categorical (Nominal/Ordinal) or continuous (Interval/Ratio).

Descriptive Analytical Pathways in SPSS Procedural Routing
Path A: Categorical Variables (Nominal / Ordinal):
Executed via `Analyze -> Descriptive Statistics -> Frequencies`. Produces frequency counts, percentages, valid percentages, cumulative percentages, and bar/pie charts.

Path B: Continuous Metric Variables (Interval / Ratio):
Executed via `Analyze -> Descriptive Statistics -> Descriptives` (for streamlined reporting) or `Analyze -> Descriptive Statistics -> Explore` (for comprehensive diagnostic screening of normality, trimmed means, M-estimators, and outliers).

Path C: Bivariate Categorical Association:
Executed via `Analyze -> Descriptive Statistics -> Crosstabs`. Generates two-way contingency tables, cell percentages (row, column, total), and Pearson Chi-Square tests of independence.

1.1 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. Interpreting the resulting output table requires distinguishing between four distinct percentage columns:

- **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). In formal academic research, researchers almost

universally report **Valid Percent** rather than raw Percent because it reflects the true proportion among respondents who provided a usable response.

- **Cumulative Percent:** The progressive running total of valid percentages summed from top to bottom. Highly useful for ordinal variables (e.g., determining what cumulative percentage of investors have completed at least an undergraduate degree or earn below a specified income tier).

Within the `Frequencies` dialog box, SPSS offers three vital sub-dialog configurations:

- **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.
- **Charts Sub-dialog:** Generates Bar Charts, Pie Charts, or Histograms (with optional normal curve overlay) directly linked to frequency counts or percentages.
- **Format Sub-dialog:** Dictates table ordering—Ascending values, Descending values, Ascending counts (from least frequent to most frequent), or Descending counts (ideal for Pareto analysis identifying dominant customer segments). Researchers can also suppress tables with more than N categories to avoid massive output clutter.

```
/* SPSS Syntax: Comprehensive Frequencies Analysis */
FREQUENCIES VARIABLES=Gender Education_Level Investment_Platform
/ORDER=ANALYSIS
/BARCHART FREQ
/MISSING=EXCLUDE.
```

1.2 Analyzing Continuous Variables: Central Tendency and Dispersion

For continuous scale variables (e.g., Age in Years, Annual Income, Monthly Trading Volume, Composite Financial Literacy Index), summary statistics provide insight into distribution shape, location, and variability:

Descriptive Metric	Conceptual Definition	Research Application & Interpretation
Arithmetic Mean	The sum of all observed values divided by the total number of observations.	The premier measure of central tendency for normally distributed symmetric data. Highly vulnerable to extreme distortion from outliers.
Median	The exact 50th percentile; the middle value separating the upper half from the lower half of the sorted distribution.	The preferred measure of central location for heavily skewed data (such as household income, executive bonuses, or trading profits) because it is immune to extreme outliers.
Mode	The single most frequently occurring score in the dataset.	The only meaningful measure of central tendency for purely nominal categorical data. Useful in retail analytics to identify peak demand price points.
Standard Deviation (SD)	The square root of the variance; measures the average distance of observed values from the sample mean.	Quantifies dispersion. In a normal distribution, approximately 68.2% of cases fall within Mean ± 1 SD, and 95.4% fall within Mean ± 2 SD.

Descriptive Metric	Conceptual Definition	Research Application & Interpretation
Standard Error of the Mean (SE Mean)	Standard deviation divided by the square root of sample size (SD / sqrt(N)).	Measures sampling precision; estimates how much the sample mean fluctuates from the true population parameter. Used to construct 95% Confidence Intervals.
Variance	The average of the squared deviations from the mean.	Fundamental building block for inferential models (ANOVA, Regression). Expressed in squared units of measurement.
Interquartile Range (IQR)	The difference between the 75th percentile (Q3) and the 25th percentile (Q1).	Measures the spread of the central 50% of the distribution; resilient against extreme skewness and outliers. Forms the basis of Tukey's boxplot fences.

1.3 The Descriptives Procedure and Automatic Standardized Z-Scores

The `Descriptives` procedure (`Analyze -> Descriptive Statistics -> Descriptives`) is optimized for quickly generating summary tables across multiple continuous variables simultaneously. A distinctive feature of this procedure is the checkbox labeled **"Save standardized values as variables"**.

When checked, SPSS automatically computes standardized Z-scores for every selected variable and appends them as brand new variables in the Data Editor with the prefix `Z` (e.g., `ZIncome`, `ZAge`). A standardized Z-score expresses an observation's relative standing in units of standard deviation above or below the mean:

Standardized Z-Score Formulation:
Z = (X - Mean) / Standard Deviation

Properties of Standardized Z-Scores:
• Mean of Z = 0.00; Standard Deviation of Z = 1.00.
• In a normal distribution, 99.7% of all observations fall between Z = -3.00 and Z = +3.00.
• Enables researchers to compare scores measured on completely different measurement scales (e.g., comparing a respondent's SAT score with their annual income).

1.4 Bivariate Exploration: Crosstabulation & Contingency Tables

While Frequencies and Descriptives examine single variables univariately, `Analyze -> Descriptive Statistics -> Crosstabs` examines bivariate relationships between two categorical variables. In commerce and banking studies, Crosstabs reveals critical interaction patterns, such as investment preference across age cohorts or loan default rates across credit score tiers.

- **Row, Column, and Total Percentages:** In the *Cells* sub-dialog, selecting Row Percentages displays how the categories of the row variable distribute across columns; selecting Column Percentages shows how columns distribute down rows. Misinterpreting row versus column percentages is a notorious research error; column percentages must be examined when the column variable represents the independent grouping factor.
- **Chi-Square Independence Test:** In the *Statistics* sub-dialog, checking *Chi-square* calculates Pearson's Chi-Square statistic to test whether the two categorical variables are statistically independent or significantly associated.
- **Directional Measures:** Computes Lambda, Cramer's V, and Phi coefficients to quantify the magnitude and strength of the categorical association.

2. 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)**. Violations of normality distort p-values, compromise statistical power, and produce invalid conclusions.

2.1 Skewness and Kurtosis: Mathematical Benchmarks

Normality is evaluated quantitatively by examining the numerical values of Skewness and Kurtosis provided in the SPSS Descriptives and Explore outputs:

- **Skewness:** Measures the degree and direction of distribution asymmetry around the central mean:
  - *Symmetric / Normal:* Skewness value equals **0.00**.
  - *Positive Skewness (Right-Skewed):* Values > 0. Distribution has a prolonged tail extending toward positive high values, with the bulk of observations clustered on the left (e.g., individual wealth or option trading earnings). Mean > Median.
  - *Negative Skewness (Left-Skewed):* Values < 0. Distribution has an extended tail trailing toward low negative values, with the majority clustered on the right (e.g., easy exam scores, customer satisfaction with high ratings). Mean < Median.
- **Kurtosis:** Measures the "peakedness" and tail heaviness of the distribution relative to a standard normal curve:
  - *Mesokurtic (Normal):* Kurtosis value equals **0.00** (SPSS calculates excess kurtosis where normal = 0).
  - *Leptokurtic (Positive Kurtosis):* Values > 0. Distribution exhibits an excessively sharp central peak with heavy, fat tails. Indicates higher vulnerability to extreme outlier occurrences.
  - *Platykurtic (Negative Kurtosis):* Values < 0. Distribution is relatively flat, dispersed, with thin tails.

Standardized Normality Z-Score Testing (Field, 2018):
Z_Skewness = Skewness Statistic / Standard Error of Skewness
Z_Kurtosis = Kurtosis Statistic / Standard Error of Kurtosis

Decision Rules:
• For Small Samples (N < 50): Values of Z between ± 1.96 confirm normality at p = 0.05 level.
• For Medium Samples (50 ≤ N ≤ 200): Values between ± 2.58 confirm normality at p = 0.01 level.
• For Large Samples (N > 200): Z-scores become hypersensitive to trivial deviations; researchers evaluate absolute Skewness values between ± 1.00 (or ± 2.00) and absolute Kurtosis values between ± 2.00 (or ± 7.00) as acceptable for parametric modeling (Kline, 2015; George & Mallery, 2020).

2.2 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**.

Hypothesis Framework for Normality Tests Decision Logic
Null Hypothesis (H0): The distribution of sample scores does not differ significantly from a normal distribution (Distribution is Normal).
Alternative Hypothesis (H1): The distribution of sample scores differs significantly from a normal distribution (Distribution is Non-Normal).

CRITICAL INTERPRETATION RULE:
Unlike standard hypothesis tests where researchers seek statistical significance ($p < 0.05$) to reject the null, **in normality testing, researchers desire a NON-SIGNIFICANT result ($p > 0.05$)**.
• *$p > 0.05$:* Fail to reject H_0 → Normality assumption is satisfied → Proceed with Parametric Tests (t-test, ANOVA, Pearson r).
• *$p \leq 0.05$:* Reject H_0 → Normality is violated → Deploy Non-Parametric Tests (Mann-Whitney, Kruskal-Wallis, Spearman rho) or apply data transformations.

- **Kolmogorov-Smirnov (K-S) Test with Lilliefors Significance Correction:** Suitable for large samples ($N > 50$). It compares the empirical cumulative distribution function of the sample against a theoretical normal cumulative distribution. The Lilliefors correction adjusts for the fact that sample mean and variance are estimated from the sample rather than known a priori.
- **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.

2.3 Graphical Assessment of Normality: Visual Diagnostics

In large datasets, formal statistical tests (K-S and Shapiro-Wilk) frequently flag trivial, harmless deviations as statistically significant due to excessive statistical power. Methodologists strongly advise combining quantitative tests with visual diagnostic plots:

- **Histogram with Normal Curve Overlay:** A visual Gaussian bell curve is superimposed over sample bars, allowing instantaneous visual detection of skewness, bimodality, and extreme tail outliers.
- **Normal Probability Q-Q Plot (Quantile-Quantile Plot):** Plots observed sample quantiles against expected normal quantiles. If the variable is perfectly normally distributed, the data points fall directly along the straight diagonal reference line. If data curves upward or downward away from the line, non-normality is present. S-shaped curves indicate heavy or light tails, while banana-shaped curves indicate skewness.
- **Detrended Normal Q-Q Plot:** Plots the actual deviation (distance) of each point from the normal line. In a normal distribution, points cluster randomly in a horizontal band above and below the central zero line with no systematic curved pattern.
- **P-P Plot (Probability-Probability Plot):** Plots empirical cumulative probabilities against theoretical normal cumulative probabilities. Particularly sensitive to non-normality in the middle of the distribution, whereas Q-Q plots are more sensitive to anomalies in the tails.

3. 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. Outliers are classified as univariate (extreme on a single variable) or multivariate (unusual combination of scores across multiple variables).

3.1 Univariate Outlier Detection Techniques

SPSS provides four primary diagnostic tools for unearthing univariate outliers:

- **1. The 5% Trimmed Mean (Explore Output):** SPSS calculates the mean after removing the top 5% and bottom 5% of extreme cases. Researchers compare the Arithmetic Mean with the 5% Trimmed Mean. If the difference between the two means is substantial (e.g., greater than 0.20 on a 5-point scale or visibly shifts the value), extreme outliers are exerting undue leverage over the sample mean.

Degreelive

- **2. Boxplot (Tukey's Fences):**
  - *Mild Outliers (represented by a circle 'o'):* Cases that lie between 1.5 and 3.0 times the Interquartile Range (IQR) beyond the outer edge of the box (Q1 - 1.5*IQR or Q3 + 1.5*IQR). SPSS explicitly displays the case number next to the circle.
  - *Extreme Outliers (represented by an asterisk '*'):* Cases that lie more than 3.0 times the IQR beyond the outer box edge (Q1 - 3.0*IQR or Q3 + 3.0*IQR). These represent severe anomalies that require immediate investigative action.
- **3. Standardized Z-Scores (Tabachnick & Fidell Benchmark):** Cases with absolute Z-scores exceeding ± 3.29 ($p < 0.001$) are flagged as potential univariate outliers in large samples, while ± 2.58 is applied in medium samples.
- **4. M-Estimators (Robust Location Statistics):** Available in `Analyze -> Descriptive Statistics -> Explore -> Statistics -> M-estimators`. Calculates Huber's M-estimator, Tukey's biweight, Hampel's M-estimator, and Andrews' wave to provide robust alternatives to the arithmetic mean that downweight extreme cases.

3.2 Multivariate Outlier Detection: Mahalanobis Distance

A respondent may not have an extreme score on any single variable, but their combination of scores across multiple variables may be exceedingly bizarre (e.g., an individual reporting an Age of 18 with an Annual Income of INR 50,00,000 and 25 years of investment experience). Such cases are **multivariate outliers**.

Mahalanobis Distance (D-squared) Protocol:
Mahalanobis distance measures the distance of a case from the multivariate centroid (multidimensional mean) of all independent variables, accounting for the covariance matrix.

SPSS Execution Trick:
1. Navigate to `Analyze -> Regression -> Linear`.
2. Enter a dummy or unique Case ID variable as the *Dependent* variable, and enter all continuous research variables as *Independent* variables.
3. Click `Save` and check **Mahalanobis** under Distances.
4. SPSS calculates Mahalanobis distance for each case and saves it as a new variable `MAH_1`.

Evaluation Rule:
Compare `MAH_1` values against the Chi-Square critical value table at degrees of freedom equal to the number of independent variables at **$p < 0.001$**.
• Any case where MAH_1 exceeds the critical Chi-Square threshold is confirmed as a multivariate outlier and should be evaluated for exclusion.

3.3 Outlier Remediation Strategies

Once identified, researchers must choose an academically defensible strategy for handling outliers:

- **Data Entry Verification:** Determine whether the outlier is an erroneous data entry (e.g., entering an age of 250 instead of 25). Correct the entry using original questionnaires.
- **Deletion (Trimming):** Exclude the outlier case from further analysis. Must be fully disclosed and justified in research methodology reports. Recommended when the outlier clearly represents an unrepresentative sampling error.
- **Winsorization (Capping):** Rather than deleting the case, replace extreme values with the highest (or lowest) acceptable benchmark score (e.g., setting extreme values to the 99th percentile score, or Mean + 3 SD). Retains sample size without skewing distribution parameters.

- **Mathematical Data Transformation:** Transform the entire variable using mathematical operations to pull in extreme positive tails:
  - *Logarithmic Transformation (LG10):* Ideal for severe positive skewness and long right tails (e.g., income, asset valuations).
  - *Square Root Transformation (SQRT):* Ideal for moderate positive skewness.
  - *Inverse / Reciprocal Transformation (1/X):* Effective for extreme positive skewness.
  - *Reflection:* For negative skewness, the variable is first reflected (subtracting all values from [Max + 1]) before applying log or square root transformations.

4. Graphical Data Representation & The SPSS Chart Editor

Data visualization is an essential analytical skill that communicates complex empirical trends and relationships transparently. SPSS provides two graphical environments: the modern **Chart Builder** (drag-and-drop interface) and **Legacy Dialogs**.

4.1 Core Graphical Architectures in Business Research

1. Bar Charts (Categorical Distributions)

Visualizes discrete group counts or summary means:
• *Simple Bar:* One categorical variable across categories.
• *Clustered Bar:* Two categorical variables plotted side-by-side (e.g., investment platform preference clustered by gender).
• *Stacked Bar:* Subcategories stacked vertically to display proportional composition (100% stacked bar highlights relative shares).

2. Boxplots (Tukey's Five-Number Summary)

Visualizes median, quartiles, and outliers simultaneously:
• *Central Solid Line:* The Median (50th percentile).
• *Box Body:* Spans the Interquartile Range (Q1 to Q3).
• *Whiskers:* Extend to min/max non-outlying values (1.5 * IQR).
• *Outliers (o):* Values 1.5 to 3.0 IQR beyond the box edge.
• *Extreme Values (*):* Values > 3.0 IQR beyond the box.

3. Line Graphs (Temporal & Trend Curves)

Connects data points sequentially to display movements over time or across ordinal categories (e.g., tracking annual corporate return on equity, quarterly EPS trends, or customer satisfaction trajectories across successive fiscal quarters).

4. Scatter Plots (Bivariate Relationships)

Plots paired scores of two continuous variables on Cartesian coordinates (X and Y axes). Essential preliminary diagnostic prior to executing Pearson correlation and linear regression, visually confirming linearity, homoscedasticity, and bivariate outliers.

4.2 Editing Charts via SPSS Chart Editor

Charts produced by default in the SPSS Output Viewer are rarely ready for direct inclusion in thesis dissertations or commercial business reports. Double-clicking any graph in the Output Viewer launches the dedicated **SPSS Chart Editor**:

DegreeLive

- **Modifying Visual Properties:** Double-clicking data bars, lines, or scatter points launches the *Properties Palette*, enabling the analyst to alter fill colors, border thickness, line styles, opacity, and marker shapes.
- **Adjusting Scale Axes:** Change minimum, maximum, and major increment tick intervals on numerical axes. Avoid truncating vertical axes in bar charts to prevent misleading visual exaggerations.
- **Adding Reference Lines:** Add horizontal or vertical reference lines at specific benchmark thresholds (e.g., adding a horizontal reference line at the overall sample mean score or at an industry regulatory benchmark).
- **Displaying Data Value Labels:** Click the *"Show Data Labels"* icon to print exact percentages, frequencies, or means directly on top of graph bars or data markers.
- **Adding Linear Fit Lines:** In scatter plots, clicking the *"Add Fit Line at Total"* button instantly calculates and draws the Ordinary Least Squares (OLS) linear regression line, displaying the equation and R-squared goodness-of-fit metric. Analysts can also add Lowess smoothing curves to evaluate non-linear patterns.
- **Transposing Coordinate Axes:** Click the *"Transpose Coordinate System"* icon to swap X and Y axes instantly, converting vertical bar charts into horizontal bar charts (highly recommended when category labels are lengthy).

4.3 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*. EMF preserves scalable vector resolution, ensuring sharp printing without pixelation when resized in Word.
- **Copy as Editable Word Table:** Right-click and select `Copy`. Pasting directly into Word (`Ctrl + V`) embeds the output as an active Microsoft Word table, allowing complete reformatting of borders, cell padding, and typography according to formal university guidelines.
- **Automated Batch Export:** Navigate to `File -> Export` in the Output Viewer. Choose **Word/RTF (*.doc)** as the export format. SPSS exports the entire output log, frequency summaries, ANOVA tables, and high-resolution chart images into a single structured Microsoft Word document in seconds.
- **Syntax-Driven Exporting (OUTPUT EXPORT):** Researchers can automate document generation directly via syntax, specifying document type, margins, and destination path.

```
/* SPSS Syntax: Automated Output Export to Word Document */
OUTPUT EXPORT
/CONTENTS EXPORT=VISIBLE
/DOC DOCUMENTFILE='D:\Research_Project\Preliminary_Analysis_Report.doc'
NOTESCAPTIONS=NO WIDETABLES=WRAP.
```

5. Scale Reliability Analysis: Cronbach's Alpha

In empirical commerce and management investigations, researchers measure latent psychological constructs (such as Customer Trust, Perceived Risk, Brand Equity, Financial Literacy, or Algorithmic Trading Acceptance) using multi-item Likert scales. Prior to computing composite scores or conducting inferential modeling, the researcher must prove that the individual questionnaire items measure the same underlying construct consistently.

5.1 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):

Classical Test Theory Equation:
Observed Score (X) = True Score (T) + Measurement Error (E)

Definition of Reliability (R_xx):
Reliability is the ratio of true score variance to observed score variance:
R_xx = Variance(True Score) / Variance(Observed Score)

• A perfectly reliable instrument has zero error variance (Reliability = 1.00).
• An unreliable instrument is dominated by measurement error (Reliability approaches 0.00).
Researchers differentiate between four distinct forms of reliability:

- **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:** Splitting multi-item scales into two halves (e.g., 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. The universal standard metric for internal consistency is **Cronbach's Alpha**, formulated by Lee Cronbach in 1951.

5.2 Mathematical Formulation of Cronbach's Alpha

Cronbach's Alpha mathematically represents the average of all possible split-half reliability coefficients that could be calculated from an instrument:

Cronbach's Alpha Formula:
Alpha = [k / (k - 1)] * [1 - (Sum of Item Variances / Total Scale Variance)]

Where:
• k = Total number of measurement items in the scale.
• Sum of Item Variances = Sum of individual variances for each of the k survey items.
• Total Scale Variance = Variance of the total composite summated score across respondents.

5.3 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. (Never place items from different theoretical constructs into the same reliability run, as Cronbach's Alpha assumes unidimensionality).
- **Step 3: Model Selection:** Ensure **Alpha** is selected in the Model dropdown (the default). Other available models include Split-half, Guttman, Parallel, and Strict Parallel.
- **Step 4: Configure Statistical Output:** Click the `Statistics` button. In the dialogue box, check:
  - **Descriptives for:** Check *Item*, *Scale*, and **Scale if item deleted**.
  - **Inter-Item:** Check *Correlations*.

Degreelive

    - **Summaries:** Check *Means* and *Variances*.
    - **ANOVA Table:** Optional; F-test for item equality.
- **Step 5: Execute:** Click `Continue` and then `OK` (or `Paste` to generate syntax).

```
/* SPSS Syntax for Scale Reliability Analysis */
RELIABILITY
/VARIABLES=Risk_Tol_1 Risk_Tol_2R Risk_Tol_3 Risk_Tol_4
/SCALE('Risk Tolerance Construct') ALL
/MODEL=ALPHA
/STATISTICS=DESCRIPTIVE SCALE CORR
/SUMMARY=TOTAL.
```

5.4 Comprehensive Interpretation of Reliability Output Tables

SPSS produces three critical output tables that require systematic academic scrutiny:

- **1. Case Processing Summary:** Confirms the number of valid cases included and reports cases excluded due to missing data. Ensure that listwise case exclusion has not depleted the sample size.
- **2. Reliability Statistics:** Displays the computed **Cronbach's Alpha** coefficient and the number of items (N of Items):
    - *Alpha ≥ 0.90:* Excellent reliability (though values above 0.95 may suggest excessive item redundancy).
    - *0.80 ≤ Alpha < 0.90:* Good / Highly Reliable; the gold standard for published empirical commerce research.
    - *0.70 ≤ Alpha < 0.80:* Acceptable threshold for exploratory behavioral studies (Nunnally & Bernstein, 1994).
    - *0.60 ≤ Alpha < 0.70:* Questionable / Poor; marginally acceptable only in initial preliminary exploratory research.
    - *Alpha < 0.60:* Unacceptable; the scale is plagued by measurement error and must be refined or abandoned.
- **3. Inter-Item Correlation Matrix:** Displays bivariate correlations between every pair of items. Methodologists recommend that inter-item correlations should fall between **0.20 and 0.50** (Clark & Watson, 1995). Correlations below 0.20 indicate items that fail to measure the core construct, while correlations above 0.80 suggest identical, redundant items.
- **4. Item-Total Statistics (Diagnostic Table):** The premier table for instrument diagnostic refinement:
    - *Scale Mean if Item Deleted:* Shows the scale mean if that specific item were removed.
    - *Scale Variance if Item Deleted:* Shows scale variance without that item.
    - *Corrected Item-Total Correlation:* Indicates the correlation between each individual item and the composite score of all remaining items. **Any item exhibiting a value below 0.30 is problematic**, indicating it does not correlate well with the overall scale and should be considered for deletion.
    - *Squared Multiple Correlation:* Indicates the proportion of variance in that item explained by all other items.
    - *Cronbach's Alpha if Item Deleted:* Shows what the overall Cronbach's Alpha would become if that specific item were removed. If deleting an item causes Alpha to increase significantly (e.g., from 0.64 to 0.82), that item is suppressing scale reliability and should be purged.

DegreeLive

5.5 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:

- **Failure to Reverse-Code Negatively Phrased Items:** Questionnaires commonly include reverse-worded items to minimize acquiescence bias (e.g., "I feel anxious when investing in stocks" alongside positive risk-tolerance items). If the researcher forgets to reverse-code this item (converting 1 to 5, 2 to 4, etc.), it correlates negatively with other items, causing Cronbach's Alpha to collapse. Always inspect the Inter-Item Correlation Matrix; negative correlations signal unreversed items.
- **Multidimensionality:** Cronbach's Alpha assumes that all items measure a single unidimensional construct. If items from two distinct theoretical constructs (e.g., "Financial Literacy" and "Risk Tolerance") are lumped into a single reliability run, Alpha is artificially suppressed. Conducting Exploratory Factor Analysis (EFA) confirms dimensionality before running reliability.

5.6 Presenting Scale Reliability in Publication Format

In academic dissertations and journal articles, researchers synthesize reliability statistics into a standardized APA-style summary table:

Scale Construct	No. of Items	Sample Mean	Std. Deviation	Cronbach's Alpha	Reliability Status
Financial Literacy	5	3.82	0.68	0.864	Good / Highly Reliable
Risk Tolerance	4	3.14	0.82	0.812	Good / Highly Reliable
Trading Platform Trust	6	4.05	0.54	0.891	Good / Highly Reliable
Algorithmic Adoption Intention	4	2.78	0.95	0.785	Acceptable / Reliable

6. 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:

APA Style Reporting Templates Scholarly Standards

Reporting Descriptive Statistics:
"The mean investment portfolio size among retail investors was INR 3,45,200 (SD = 1,12,450, Mdn = INR 3,10,000, IQR = 95,000). The distribution showed moderate positive skewness (Skewness = 0.84, SE = 0.14) and mesokurtic peakedness (Kurtosis = 0.42, SE = 0.28)."

Reporting Normality Tests:
"Univariate normality was evaluated using the Shapiro-Wilk test. Scores on the Financial Literacy Index were normally distributed, $W(250) = 0.991$, $p = 0.142$. In contrast, Monthly Trading Volume violated normality, $W(250) = 0.864$, $p < 0.001$; inspection of the normal Q-Q plot confirmed severe right skewness, prompting logarithmic transformation prior to parametric testing."

Reporting Scale Reliability:
"The 6-item Trading Platform Trust scale demonstrated high internal consistency reliability, Cronbach's alpha = 0.891. All corrected item-total correlations exceeded the recommended 0.30 threshold (ranging from 0.62 to 0.78), and no substantial increase in alpha was observed upon individual item deletion."

Summary Synthesis

Analytical Procedure	Key SPSS Tools & Metrics	Core Methodological Function
Categorical Exploration	Frequencies (Valid Percent, Cumulative Percent).	Summarizes demographic sample composition and response distributions accurately.
Continuous Exploration	Descriptives, Explore (Mean, Median, SD, IQR, Z-scores).	Quantifies central tendency and dispersion while evaluating skewness and kurtosis.
Normality Testing	Kolmogorov-Smirnov, Shapiro-Wilk, Q-Q and P-P Plots.	Tests fundamental parametric assumptions; non-significant $p > 0.05$ validates parametric modeling.
Outlier Detection & Diagnostics	5% Trimmed Mean, Boxplots (o and *), Mahalanobis Distance.	Identifies extreme influential scores that threaten statistical validity before hypothesis testing.
Data Visualization	Chart Editor, Histograms, Boxplots, Scatter Plots, EMF Export.	Transforms raw data into publication-grade graphical representations exported to Word/PDF.
Scale Reliability	Reliability Analysis (Cronbach's Alpha, Item-Total).	Verifies internal consistency (alpha ≥ 0.70/0.80) and purges problematic survey items.

Deg

Acing your exams is just a click away!
Visit www.degreelive.in to download the next module for free.

DegreeLive