

VU LEARNING HUB

vulearninghub.xyz

HIGHLIGHTED HANDOUT

STA641

Statistical Packages and its Applications

STA641 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

STA641

Statistical Packages and its Applications

STUDY HANDOUT

A complete topic-wise study handout for STA641 Statistical Packages and its Applications, prepared from statistical-computing lecture series. Covers getting data into SPSS and Excel, descriptive statistics, charts, hypothesis testing, t-tests, ANOVA, correlation, regression and chi-square - with menu paths, output interpretation, worked examples and exam pointers. 

VU Learning HUB | vulearninghub.xyz

Table of Contents

1. Why Statistical Packages
2. Getting Started with SPSS
3. Data Entry, Import and Cleaning
4. Descriptive Statistics in SPSS
5. Charts and Graphs
6. Probability Refresher for Applications
7. Hypothesis Testing Framework
8. One-Sample and Paired t-Tests
9. Independent-Samples t-Test
10. One-Way ANOVA
11. Correlation Analysis
12. Simple Linear Regression in SPSS
13. Multiple Regression
14. Chi-Square Tests
15. Non-Parametric Alternatives
16. Working in Excel: Analysis ToolPak
17. Reporting Results: Tables and Write-ups
18. Common Mistakes and How to Avoid Them
19. Revision: Midterm Practice Set
20. Revision: Final-Term Practice Set

Lecture 1: Why Statistical Packages

Key terms: SPSS, Excel, statistical software, reproducibility

Hand calculation built understanding; packages do the arithmetic. SPSS gives menu-driven statistics with publication-ready tables, while Excel handles data entry, cleaning and quick analysis everywhere. The analyst's job shifts from computing to choosing the right procedure and interpreting output correctly - the package computes, you think.

- SPSS: menus + syntax; Excel: Data Analysis ToolPak + formulas.
- Always keep raw data untouched; work on copies.
- Save syntax/scripts - point-and-click alone is not reproducible.

EXAM POINTER: Why do statisticians use packages, and what remains the analyst's responsibility? 

Lecture 2: Getting Started with SPSS

Key terms: SPSS interface, Data View, Variable View

SPSS has two main windows: Data View (the spreadsheet of cases and variables) and Variable View (where each variable's name, type, labels, values and missing codes are defined). Define variables before typing data: short names, descriptive labels, value labels for codes (1 = Male, 2 = Female), and proper missing-value codes. Output appears in the Viewer; charts can be edited there.

- Variable View first, Data View second - always.
- Value labels turn cryptic codes into readable output.
- Set decimal places sensibly; nominal variables need none.

Lecture 3: Data Entry, Import and Cleaning

Key terms: data entry, import, cleaning, missing values

Data enters via typing, Excel/CSV import, or database connections. Cleaning follows: check ranges (age 999?), duplicates, inconsistent codes, and missing values - distinguish true missing from zeros. SPSS's Frequencies is the fastest audit: one table per variable exposes every anomaly. Document every cleaning decision.

- Import check: variable types and value labels survive the trip?
- Frequencies audit catches out-of-range and miscoded values.
- Never delete suspicious cases silently - investigate, then decide.

Lecture 4: Descriptive Statistics in SPSS

Key terms: Frequencies, Descriptives, Explore, descriptives

Analyze > Descriptive Statistics > Frequencies gives counts, percentages and charts for categorical data. Descriptives gives mean and SD for scale variables. Explore is the workhorse: descriptives plus normality tests, stem-and-leaf and boxplots, split by groups. Report centre, spread and shape - never a mean without its SD.

- Categorical: frequencies and percentages; scale: mean, SD, median.
- Explore: tick Normality plots to test assumptions.
- Copy key tables to your report; don't dump raw output.

Lecture 5: Charts and Graphs

Key terms: bar chart, histogram, boxplot, scatter plot

Graphs reveal what tables hide. Bar charts compare categories; pie charts show composition (use sparingly); histograms show distribution shape; boxplots compare groups and flag outliers; scatter plots show bivariate relationships. SPSS Chart Builder or legacy dialogs create them; always title, label axes, and choose the chart that answers the question.

- Distribution shape: histogram with normal curve overlay.
- Group comparison: clustered bar or side-by-side boxplots.
- Every chart needs: title, labelled axes, and a source note.

EXAM POINTER: Which chart for: income distribution? sales by region? age vs spending? 

Lecture 6: Probability Refresher for Applications

Key terms: probability, normal distribution, sampling distribution

Applied work leans on a few ideas: the normal distribution (bell curve) models many measurements; the sampling distribution of the mean lets us infer populations from samples; the Central Limit Theorem says means become normal as n grows. P-values and confidence intervals - the language of every test below - rest on these foundations.

- Check normality before t-tests/ANOVA: Shapiro-Wilk, Q-Q plots.
- $n > 30$ per group makes most tests robust to non-normality.
- A p-value is not the probability H_0 is true - interpret carefully.

Lecture 7: Hypothesis Testing Framework

Key terms: H_0 , H_1 , p-value, alpha, Type I/II errors

Every test follows the ritual: state H_0 (no effect/difference) and H_1 , choose α (usually 0.05), compute the test statistic and p-value, decide. $p < \alpha$ rejects H_0 . Type I error: false alarm (rejecting true H_0); Type II: miss (keeping false H_0). Power ($1 - \beta$) is the chance of detecting a real effect - bigger samples mean more power.

- Write hypotheses before seeing the data - no HARKing.
- Report exact p-values ($p = .032$), not just 'significant'.
- Non-significant is not 'no effect' - it may be low power.

Lecture 8: One-Sample and Paired t-Tests

Key terms: one-sample t-test, paired t-test, before-after

One-sample t-test: does the sample mean differ from a known value? (Is average delivery time 30 minutes?) Analyze > Compare Means > One-Sample T Test. Paired t-test: before-after on the same subjects (diet programme weights) - Analyze > Compare Means > Paired-Samples T Test. Both assume roughly normal differences; report mean difference with its confidence interval.

- Paired = same subjects twice; independent = two separate groups.
- Check difference scores for normality, not the raw variables.
- Effect size (Cohen's d) tells practical importance.

Lecture 9: Independent-Samples t-Test

Key terms: independent t-test, Levene's test, equal variances

Compares means of two independent groups (treatment vs control exam scores). Menu:  Analyze > Compare Means > Independent-Samples T Test; grouping variable defines the groups. Levene's test checks equal variances - read the correct row ('equal variances assumed' or not). Report: $t(df) = \text{value}$, $p = \text{value}$, means and SDs per group.

- Levene's $p > .05$: use the equal-variances row.
- Boxplots first: visual check of group differences and outliers.
- Non-normal small samples: consider Mann-Whitney U instead.

EXAM POINTER: Walk through an independent t-test: menus, Levene's, and full reporting.

Lecture 10: One-Way ANOVA

Key terms: ANOVA, F-test, post hoc, Tukey

ANOVA compares three or more group means (fertiliser A/B/C yields) - running many t-tests inflates false positives, so one F-test first. Analyze > Compare Means > One-Way ANOVA; a

significant F means some groups differ, then post hoc tests (Tukey) identify which pairs. Assumptions: normality within groups, equal variances (Levene's), independent observations.

- Significant F only says 'not all equal' - post hocs locate differences.
- Tukey HSD for pairwise comparisons with equal variances.
- Report: $F(df1, df2) = \text{value}, p, \text{group means}, \text{post hoc results}.$

Lecture 11: Correlation Analysis

Key terms: Pearson r, Spearman, correlation matrix

Pearson's r measures linear association (-1 to +1): Analyze > Correlate > Bivariate. Check the scatter plot first - r misses curves and is dragged by outliers. Spearman's rho handles ordinal data and monotonic (not just linear) relationships. Correlation never proves causation; it quantifies co-movement for the regression to come.

- Plot before correlating - Anscombe's quartet warning.
- Ordinal/Likert data: Spearman, not Pearson.
- Report r, p-value, n - and the scatter plot.

Lecture 12: Simple Linear Regression in SPSS

Key terms: regression, R-squared, coefficients, residuals

Analyze > Regression > Linear: dependent Y, independent X. Output: Model Summary (R and R-squared - fit), ANOVA table (is the model significant?), Coefficients (intercept a, slope b with p-values). Write the equation: predicted Y = a + bX. Check residuals: scatterplot of residuals vs predicted should show no pattern; histogram roughly normal.

- b: expected Y change per unit of X; test its p-value.
- R-squared: proportion of Y's variation explained.
- Patterned residuals mean the model is misspecified - investigate.

Lecture 13: Multiple Regression

Key terms: multiple regression, adjusted R-squared, multicollinearity

Multiple predictors enter together: sales = f(price, ads, season). Each coefficient is the effect holding others constant. Adjusted R-squared penalises useless predictors. Check multicollinearity (VIF - above 10 is trouble) and use dummy variables for categories. Method choices: Enter (all at once, preferred) vs stepwise (data-dredging risk).

- VIF check: Analyze > Regression > Linear > Statistics > Collinearity diagnostics.
- Stepwise capitalises on chance - prefer theory-driven Enter.

- Validate on fresh data when prediction is the goal.

EXAM POINTER: Interpret a multiple regression output: coefficients, R-squared, VIF. 

Lecture 14: Chi-Square Tests

Key terms: chi-square, contingency table, independence

Chi-square tests categorical associations: Analyze > Descriptive Statistics > Crosstabs > Statistics > Chi-square. Example: is product preference related to gender? Requirements: independent observations, expected counts ≥ 5 in most cells (else Fisher's exact). Report chi-square(df) = value, p, and the contingency table with percentages.

- Crosstabs: put percentages in the direction you compare.
- Expected count < 5 : combine categories or use Fisher's exact.
- Effect size: Cramer's V for strength of association.

Lecture 15: Non-Parametric Alternatives

Key terms: Mann-Whitney, Wilcoxon, Kruskal-Wallis, non-parametric

When normality fails and samples are small, rank-based tests rescue the analysis: Mann-Whitney U (two independent groups), Wilcoxon signed-rank (paired), Kruskal-Wallis (3+ groups). They test medians/ranks rather than means. In SPSS: Analyze > Nonparametric Tests > Legacy Dialogs. Less power than parametric tests when assumptions hold - so check assumptions first.

- Ordinal data or skewed small samples: go non-parametric.
- Report medians and IQR instead of means and SDs.
- Kruskal-Wallis significant: follow with pairwise comparisons.

Lecture 16: Working in Excel: Analysis ToolPak

Key terms: Excel, ToolPak, formulas, pivot tables

Excel's Data Analysis ToolPak (enable in Add-ins) runs descriptives, t-tests, ANOVA, correlation and regression without SPSS. Formulas (AVERAGE, STDEV, T.TEST, CHISQ.TEST) give quick answers; pivot tables summarise massive tables by drag-and-drop. Excel is ideal for cleaning, quick checks and sharing - SPSS for heavy inference and publication tables.

- Enable: File > Options > Add-ins > Analysis ToolPak.
- Pivot tables: the fastest exploratory summary in Excel.
- Mind Excel's statistical quirks - verify critical results in SPSS.

Lecture 17: Reporting Results: Tables and Write-ups

Key terms: APA style, reporting, tables

A results section reports, not dumps: describe the sample, present descriptives, then each test with its statistic, df, p-value and effect size, in APA-ish style - 'Men ($M = 72.3$, $SD = 8.1$) scored higher than women ($M = 68.4$, $SD = 7.9$), $t(98) = 2.45$, $p = .016$, $d = 0.49$.' Tables carry numbers; text carries interpretation. Every claim needs its statistic.

- One table per analysis, numbered and titled.
- Round sensibly: means to 1-2 decimals, p to 3.
- Interpretation answers 'so what?' - not just ' $p < .05$ '.

Lecture 18: Common Mistakes and How to Avoid Them

Key terms: mistakes, p-hacking, assumptions, interpretation

Classic errors: testing without checking assumptions, confusing statistical with practical significance, p-hacking (testing until something is significant), treating correlation as causation, and ignoring missing-data patterns. Defences: plan analyses in advance, check assumptions routinely, report effect sizes and confidence intervals, and pre-register when stakes are high.

- Assumption checks are part of the analysis, not optional extras.
- A tiny p with a trivial effect is a large-sample artefact.
- Transparency - syntax, data, decisions - is the ultimate defence.

Lecture 19: Revision: Midterm Practice Set

1. Explain Variable View vs Data View and the steps of defining a survey variable. (10 marks)
2. Describe a data-cleaning audit using Frequencies. (5 marks)
3. Compare bar chart, histogram and boxplot: when is each used? (10 marks)
4. State the hypothesis-testing steps and explain Type I/II errors. (10 marks)
5. When do you use paired vs independent t-test? Illustrate. (10 marks)

Lecture 20: Revision: Final-Term Practice Set

1. Interpret an independent t-test output including Levene's test. (10 marks)
2. Explain one-way ANOVA and post hoc testing with an example. (10 marks)
3. Interpret a regression output: R-squared, coefficients, residuals. (10 marks)
4. When is chi-square appropriate? State its requirements. (5 marks)
5. Write an APA-style results paragraph for a given t-test output. (10 marks)

End of Handout - Good luck with your exams!