

VU LEARNING HUB

vulearninghub.xyz

HANDOUT

STA5209

Data Analysis by Statistical Software-I

STA5209 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

STA5209

Data Analysis by Statistical Software-I

HANDOUT

This course teaches practical data analysis with statistical software (SPSS/R): data management, descriptive statistics and visualization, hypothesis testing, ANOVA, correlation and regression, and reporting results reproducibly.

Table of Contents

Module 1	Statistical Software Environment
Module 2	Descriptive Statistics
Module 3	Data Visualization
Module 4	Probability and Sampling Distributions
Module 5	Hypothesis Testing Framework
Module 6	t-Tests and ANOVA
Module 7	Correlation and Regression
Module 8	Categorical Data Analysis
Module 9	Nonparametric Tests and Questionnaire Analysis
Module 10	Reporting Results and Research Workflow

MODULE 1: STATISTICAL SOFTWARE ENVIRONMENT

Learning Objectives

- Navigate the software interface (SPSS/R)
- Distinguish variable types and measurement scales
- Import, save, and manage datasets
- Write reproducible analysis scripts

Why Software, Why Reproducibility

Modern statistics is **computational**: datasets are too large for hand calculation, and software (SPSS, R, Stata, Python) implements methods reliably. But point-and-click analysis is **not reproducible** — no record of what was done. Professional practice: **script everything** (R scripts, SPSS syntax files) so any result can be regenerated exactly. A good analysis folder holds raw data (never edited), scripts, and outputs — the **reproducible workflow**.

Data and Variable Types

Variables by **measurement scale**: **nominal** (categories, e.g., gender), **ordinal** (ordered, e.g., Likert scale), **interval** (equal gaps, no true zero — temperature °C), **ratio** (true zero — income, height). Also **discrete vs. continuous**. The scale **dictates valid methods**: means need interval/ratio; medians suit ordinal; chi-square suits nominal. Declaring variable types and value labels in software (SPSS Variable View; R factors) prevents silent errors.

Getting Data In and Clean

Import from CSV/Excel; check **data integrity** — dimensions, ranges, duplicates. **Missing values**: distinguish MCAR/MAR/MNAR; options — listwise deletion, mean imputation (biases variance), **multiple imputation** (best practice). **Outliers**: detect via boxplots and z-scores; investigate before deleting — an outlier may be the discovery. **Recoding**: reverse-scoring Likert items, creating age groups, dummy variables for regression.

Key Terms & Definitions

Measurement scales	Nominal, ordinal, interval, ratio — Stevens' taxonomy.
Reproducible workflow	Scripted analysis regenerating every result from raw data.
MCAR / MAR / MNAR	Missing-data mechanisms: completely/at random, not at random.
Dummy variable	0/1 coding of a categorical level for regression.

Example: The Mean of Gender

Coding gender 1/2 and taking the mean (1.4) is meaningless — nominal data has no numeric scale. Software will happily compute it; only the analyst's declared measurement level stands between a number and nonsense. Declare types first, analyze second.

Module Summary

- Script analyses (R/SPSS syntax) — reproducibility is professionalism.
- Measurement scale (nominal → ratio) dictates valid methods.
- Import carefully: check dimensions, ranges, duplicates.
- Handle missing values and outliers deliberately, not by default.

Exam Pointers (Very Important)

- ★ Why is scripted analysis superior to point-and-click?
- ★ Classify five variables by measurement scale.
- ★ Compare strategies for handling missing data.
- ★ What should a reproducible analysis folder contain?

MODULE 2: DESCRIPTIVE STATISTICS

Learning Objectives

- Compute and interpret measures of center and spread
- Detect skewness and outliers numerically
- Choose appropriate summaries per variable type
- Produce publication-quality summary tables

Center and Spread

Center: mean (sensitive to outliers), **median** (robust; best for skewed data like income), mode (nominal data). **Spread:** range, **IQR** (robust), **variance/standard deviation** (in the original units via SD), coefficient of variation (unit-free comparison). **Empirical rule:** for bell-shaped data, ~68/95/99.7% within 1/2/3 SD. **z-scores** standardize: $z = (x - \bar{x})/s$ — the basis of outlier flags ($|z| > 3$) and many tests.

Shape: Skewness and Kurtosis

Skewness: right-skewed (mean > median; income, waiting times) vs. left-skewed. **Kurtosis:** tail heaviness vs. normal. Why it matters: many methods assume approximate normality — strong skew suggests **transformation** (log for right-skewed) or **nonparametric** methods. Report skewness/SE as a quick normality screen alongside plots.

Summaries by Group

Real questions are comparative: mean blood pressure **by treatment group**, income **by province**. Software: SPSS Split File / Compare Means; R `aggregate()/dplyr group_by()`. Present as **mean ± SD (n)** tables. **Crosstabs** for categorical pairs with row/column percentages. Rule: every table needs a title, units, sample sizes, and a one-line interpretation — a table without a message is decoration.

Key Terms & Definitions

IQR	Interquartile range: $Q3 - Q1$; robust spread measure.
z-score	$(x - \bar{x})/s$; positions a value in SD units.
Skewness	Asymmetry of a distribution.
Five-number summary	Min, $Q1$, median, $Q3$, max.

Example: Mean vs. Median Income

A town's mean income is Rs. 85,000 but median Rs. 45,000 — a few high earners drag the mean up. Reporting only the mean would mislead policy; the median describes the typical resident. Shape-aware summarizing is honest summarizing.

Module Summary

- Center: mean/median/mode — match to variable type and skew.
- Spread: range, IQR, SD; z-scores standardize and flag outliers.
- Skewness guides transformation vs. nonparametric choices.
- Summarize by group; tables need titles, units, n, interpretation.

Exam Pointers (Very Important)

- ★ When is the median preferable to the mean?
- ★ Interpret skewness = +1.8 for a dataset.
- ★ Construct a grouped summary table from raw data.
- ★ Explain the empirical rule and z-scores.

MODULE 3: DATA VISUALIZATION

Learning Objectives

- Choose the right chart for each variable type
- Build histograms, boxplots, bar charts, scatterplots
- Avoid misleading graphics
- Apply visualization best practices

Chart Grammar

Match chart to data: **bar chart** for nominal/ordinal counts; **histogram** for continuous distributions (bins reveal shape); **boxplot** for group comparisons and outlier display; **scatterplot** for bivariate relationships; **line chart** for time series. Software: SPSS Chart Builder; R ggplot2 (grammar of graphics: data + aesthetics + geoms). **Plot before you test** — graphs reveal what statistics hide.

Reading Distributions and Relationships

Histograms show **shape, center, spread, outliers, and modes** — bimodality suggests hidden subgroups. Boxplots compare **medians, IQRs, and outliers across groups** at a glance. Scatterplots reveal **form, direction, strength** of association plus nonlinearity and influential points — always inspect before computing correlation (Anscombe's quartet: identical correlations, wildly different plots). Add **jitter** for overplotted discrete data; consider **log scales** for wide-range data.

Honest Graphics

Misleading tricks: **truncated y-axes** exaggerating differences, 3-D effects distorting areas, dual axes implying false relationships, cherry-picked ranges. Best practices: start bars at zero, label axes with units, use color for meaning (not decoration), keep **data-ink ratio** high (Tufte), write informative titles. A visualization's job: **reveal truth efficiently** — beauty serves clarity, never replaces it.

Key Terms & Definitions

Anscombe's quartet	Four datasets, same statistics, different plots — plot first.
Jitter	Small random displacement reducing overplotting.
Data-ink ratio	Tufte: maximize ink devoted to data.
Bimodal	Two-peaked distribution suggesting subgroups.

Example: The Truncated Axis Trick

Two bars (51 vs. 49) with the axis starting at 48 look dramatically different; starting at 0 they look identical — which they nearly are. Always check the axis before believing the drama: honest graphics start bars at zero.

Module Summary

- Bars for categories, histograms for distributions, boxplots for groups, scatterplots for relationships.
- Plot before testing — Anscombe's quartet proves why.
- Honest axes: bars start at zero; label units; color with meaning.
- High data-ink ratio; titles state the message.

Exam Pointers (Very Important)

- ★ Which chart for: Likert responses? Income distribution? BP by group?
- ★ Explain Anscombe's quartet and its lesson.
- ★ Identify three misleading-graphic tricks and their fixes.
- ★ How do boxplots reveal outliers and group differences?

MODULE 4: PROBABILITY AND SAMPLING DISTRIBUTIONS

Learning Objectives

- Review probability rules for inference
- Explain the sampling distribution concept
- Apply the Central Limit Theorem
- Construct confidence intervals in software

Probability Essentials

Rules: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$; $P(A \text{ and } B) = P(A)P(B|A)$; **conditional probability** $P(B|A) = P(A \text{ and } B)/P(A)$. **Bayes' theorem** inverts conditionals — the logic of diagnostic testing (sensitivity, specificity, **positive predictive value** depends on prevalence!). Distributions: **binomial** (counts of successes), **normal** (z-scores, empirical rule), **t** (small samples), **chi-square** and **F** (test statistics' distributions).

Sampling Distributions and CLT

A **statistic** varies sample to sample; its **sampling distribution** describes that variation. **Standard error** = SD of the sampling distribution (SE of mean = $s/\sqrt{n}$ — precision grows with $\sqrt{n}$, so quadrupling n halves SE). **Central Limit Theorem**: for large n , the sampling distribution of the mean is **approximately normal regardless of population shape** — the license for t-tests and confidence intervals on skewed data given enough n .

Confidence Intervals

A 95% CI: $\bar{x} \pm t^* \cdot SE$ — the range of **plausible population values**, not a probability about this interval. Interpretation done right: values inside are compatible with the data; width shows **precision**. Software: SPSS Explore/One-Sample T Test; R `t.test()`. **CIs beat p-values** for communication — they show magnitude and uncertainty together. Report as 'mean difference 4.2 (95% CI: 1.1 to 7.3)'.

Key Terms & Definitions

Standard error	SD of a statistic's sampling distribution; $s/\sqrt{n}$ for the mean.
Central Limit Theorem	Sample means become normal as n grows.
Confidence interval	Range of plausible population parameter values.
Bayes' theorem	$P(A B) = P(B A)P(A)/P(B)$; inverts conditionals.

Example: The Screening Paradox

A test 99% sensitive and 99% specific for a disease with 0.1% prevalence: a positive result means only ~9% chance of disease (Bayes). Most positives are false — because the disease is rare. Predictive value depends on prevalence, not just test accuracy: base rates rule.

Module Summary

- Probability rules + Bayes; key distributions: binomial, normal, t, χ^2 , F.
- Sampling distribution → standard error; CLT justifies normal-based inference.
- 95% CI = plausible values + precision; report magnitude with uncertainty.
- Interpretation: compatible values, not probability the truth is inside.

Exam Pointers (Very Important)

- ★ State the Central Limit Theorem and its practical importance.
- ★ Compute and interpret a 95% CI for a mean.
- ★ Use Bayes' theorem on a diagnostic-testing problem.
- ★ Distinguish standard deviation from standard error.

MODULE 5: HYPOTHESIS TESTING FRAMEWORK

Learning Objectives

- State null and alternative hypotheses correctly
- Interpret p-values without common fallacies
- Distinguish Type I/II errors and power
- Choose between parametric and nonparametric tests

The Logic of Testing

H_0 (no effect/difference) vs. H_1 (research claim); **one- vs. two-sided** alternatives (two-sided is the default — one-sided needs strong justification). **p-value** = probability of data this extreme **if H_0 were true** — NOT $P(H_0 \text{ is true})$, NOT the probability of a fluke finding, NOT effect size. Decision: reject H_0 if $p < \alpha$ (usually 0.05). **Statistical significance** $\neq$ **practical importance**: with huge n , trivial differences are 'significant' — always report **effect sizes** (Cohen's d , r).

Errors and Power

Type I error (α): rejecting a true H_0 — false positive. **Type II error** (β): missing a real effect — false negative. **Power** = $1 - \beta$: probability of detecting a true effect; target ≥ 0.80 . Power grows with **effect size, sample size, and α** , and shrinks with variability. **Power analysis** (G*Power, R pwr) sizes studies **before** data collection — underpowered studies waste resources and mislead.

Assumptions and Nonparametric Alternatives

Parametric tests assume: **independence**, approximate **normality** (of residuals), **equal variances** (homogeneity). Check: Q-Q plots, Shapiro-Wilk, Levene's test. When violated: **transform** (log), or go **nonparametric** — Mann-Whitney U (vs. t-test), Wilcoxon signed-rank (vs. paired t), Kruskal-Wallis (vs. ANOVA), Spearman (vs. Pearson). Nonparametrics use ranks — robust but slightly less powerful when assumptions hold.

Key Terms & Definitions

p-value	$P(\text{data this extreme} \mid H_0 \text{ true})$.
Power	$1 - \beta$; probability of detecting a true effect.
Effect size	Standardized magnitude (Cohen's d); separate from p .
Homogeneity of variance	Equal variances across groups; checked by Levene's test.

Example: Significant but Trivial

$n = 50,000$ finds drug lowers BP by 0.3 mmHg, $p < 0.001$ — 'highly significant' yet clinically meaningless. The p-value answered 'is there an effect?'; the effect size ($d = 0.02$) answered 'does it matter?'. Report both, decide on the latter.

Module Summary

- $p = P(\text{data} | H_0)$; significance $\neq$ importance — report effect sizes.
- Errors: Type I (α , false +), Type II (β , false –); power = $1 - \beta \geq 0.80$.
- Power analysis before collecting data.
- Check assumptions (Q-Q, Levene); use rank-based alternatives when violated.

Exam Pointers (Very Important)

- ★ Define the p-value and list three common misinterpretations.
- ★ Explain Type I/II errors and power with an example.
- ★ When should you choose Mann-Whitney over a t-test?
- ★ Design a power analysis for comparing two means.

MODULE 6: T-TESTS AND ANOVA

Learning Objectives

- Run one-sample, independent, and paired t-tests
- Explain why ANOVA replaces multiple t-tests
- Interpret one-way ANOVA and post-hoc tests
- Check assumptions and report results in APA style

The t-Test Family

One-sample: is the mean μ ? ($t = (x - \mu)/SE$). **Independent samples:** do two groups differ? — check **Levene's test**; if variances unequal use **Welch's t** (the safer default). **Paired:** before-after or matched pairs — analyze the **differences** (one-sample t on d). Software: SPSS Compare Means; R `t.test()`. Always accompany with **means, SDs, CI of the difference, and Cohen's d** — the test alone is half the story.

One-Way ANOVA

Comparing 3+ groups with t-tests **inflates Type I error** (with 5 groups, 10 tests → ~40% family-wise error). **ANOVA:** $F = \text{between-group variance} / \text{within-group variance}$ — one omnibus test. Logic: if group means differ, between-group variation dwarfs within-group noise. **Assumptions:** normality within groups, homogeneity (Levene), independence. **Effect size:** η^2 (eta-squared) = $SS_{\text{between}}/SS_{\text{total}}$.

Post-Hoc and Reporting

A significant F only says 'somewhere differs' — **post-hoc tests** locate it: **Tukey HSD** (all pairs, controls error), **Bonferroni** (conservative, flexible), **Dunnnett** (vs. control). **Planned contrasts** (decided before seeing data) are more powerful than post-hocs. **APA report:** 'F(2, 87) = 6.42, p = .003, $\eta^2 = .13$; Tukey HSD showed...'. Include a means plot with error bars — the picture carries the result.

Key Terms & Definitions

Welch's t-test	t-test not assuming equal variances; safer default.
Family-wise error	P(any false positive) across multiple tests.
Tukey HSD	Post-hoc pairwise test controlling family-wise error.
Eta-squared (η^2)	Proportion of variance explained by group membership.

Example: The Multiple-Testing Trap

Testing 4 diets pairwise = 6 t-tests at $\alpha = .05$ → 26% chance of at least one false 'discovery'. ANOVA's single F-test holds the line at 5%; Tukey then finds which diets truly differ. One omnibus gate, then careful pairwise looks — that's the discipline.

Module Summary

- t-tests: one-sample, independent (Welch default), paired (test differences).
- ANOVA: one F-test for 3+ groups; F = between/within variance.
- Post-hocs (Tukey/Bonferroni/Dunnett) locate differences; report η^2 .
- APA style: F(df), p, effect size, means plot.

Exam Pointers (Very Important)

- ★ When do you use a paired vs. independent t-test?
- ★ Why not run multiple t-tests instead of ANOVA?
- ★ Interpret: $F(3, 76) = 4.21, p = .008, \eta^2 = .14$.
- ★ Compare Tukey HSD, Bonferroni, and Dunnett post-hocs.

MODULE 7: CORRELATION AND REGRESSION

Learning Objectives

- Compute and interpret Pearson and Spearman correlations
- Fit and interpret simple linear regression
- Extend to multiple regression with diagnostics
- Avoid causation and overfitting traps

Correlation

Pearson r: linear association, -1 to $+1$; assumes interval data and linearity — **plot first** (Anscombe!).
Spearman ρ : rank-based; for ordinal data or monotonic nonlinear relations. Interpretation: direction, strength ($|r| \approx .1$ small, $.3$ medium, $.5$ large — Cohen), and r^2 = **shared variance**. Mantra: **correlation $\neq$ causation** — third variables and reverse causation abound. **Correlation matrix** (heatmap) screens many pairs at once.

Simple Linear Regression

Model: $y = \beta_0 + \beta_1 x + \varepsilon$; **least squares** minimizes $\sum \text{residuals}^2$. $\beta_1 = r(s_y/s_x)$: **slope in context** ('each extra study hour predicts +2.1 marks'). R^2 = variance explained. **Inference**: t-test on β_1 (is the slope nonzero?), CI for the slope. **Prediction**: $\hat{y}$ for new x — with **prediction intervals** (wider than CIs; include individual variation). Never **extrapolate** beyond the data range.

Multiple Regression and Diagnostics

$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$: each β_j = effect **holding others constant**. Categorical predictors via **dummy variables**. **Diagnostics** (non-negotiable): residual-vs-fitted (linearity, **homoscedasticity**), Q-Q of residuals (normality), **VIF** (< 5 – 10 ; multicollinearity check), Cook's distance (influential points). **Adjusted R^2** penalizes added predictors; **overfitting** (too many predictors for n) memorizes noise — validate or keep models parsimonious.

Key Terms & Definitions

Least squares	Fitting by minimizing summed squared residuals.
R^2	Proportion of y-variance explained by the model.
Multicollinearity (VIF)	Redundant predictors inflating SEs.
Homoscedasticity	Constant residual variance across fitted values.

Example: The Dummy-Variable Trick

Predicting salary from experience + gender (M/F): code female = 1, male = 0; the gender coefficient is the adjusted pay gap holding experience constant. One 0/1 column turns a categorical comparison into a regression coefficient with a CI and p-value.

Module Summary

- Pearson (linear) vs. Spearman (rank); r^2 = shared variance; $\neq$ causation.
- Regression: least squares; interpret slope in context; R^2 = explained variance.
- Multiple regression: dummies for categories; adjusted R^2 ; beware overfitting.
- Diagnostics mandatory: residuals, Q-Q, VIF, Cook's distance.

Exam Pointers (Very Important)

- ★ Interpret $r = -.62$ between stress and sleep hours.
- ★ Fit $y = \beta_0 + \beta_1 x$: interpret the slope and $R^2 = .49$.
- ★ List and explain four regression diagnostics.
- ★ What is multicollinearity and how is VIF used?

MODULE 8: CATEGORICAL DATA ANALYSIS

Learning Objectives

- Build and interpret contingency tables
- Run chi-square tests of independence and goodness-of-fit
- Compute odds ratios and relative risk
- Introduce logistic regression

Contingency Tables

Crosstab: rows × columns of counts with **row/column percentages** (percent within the causal variable). **Chi-square test of independence:** $\chi^2 = \sum (O-E)^2/E$; compares observed to expected-under-independence counts. **Assumptions:** independent observations, **expected counts** ≥ 5 (else Fisher's exact test). **Goodness-of-fit** χ^2 : does one variable match hypothesized proportions (e.g., fair die)?

Effect Sizes for 2×2 Tables

χ^2 says 'associated'; **effect sizes** say 'how much': **relative risk (RR)** = risk in exposed / risk in unexposed (cohort studies); **odds ratio (OR)** = odds of outcome in exposed vs. unexposed (case-control; ≈ RR when outcome rare). **Phi/Cramér's V:** 0–1 association strength for larger tables. Report with **95% CIs** — an OR of 2.1 (CI: 0.9–4.8) is inconclusive despite exceeding 1.

Logistic Regression (Introduction)

For binary outcomes: model **log-odds** = $\beta_0 + \beta_1 x_1 + \dots$ — the **logit link** keeps predictions in (0,1). Coefficients exponentiate to **odds ratios**: $e^{\beta_j} = \text{OR per unit } x_j$. **Wald tests** assess predictors; **pseudo- R^2** and **classification tables / ROC-AUC** assess fit. Example: predicting diabetes (yes/no) from BMI and age — each coefficient is an adjusted odds ratio, the lingua franca of medical research.

Key Terms & Definitions

Chi-square test	$\chi^2 = \sum (O-E)^2/E$; tests independence or fit.
Odds ratio	Ratio of outcome odds between groups.
Logit	$\log(p/(1-p))$; link function for binary outcomes.
Fisher's exact test	Exact test for small expected counts.

Example: Smoking and Disease: OR in Action

Case-control: 90/100 cases smoked vs. 45/100 controls → $\text{OR} = (90 \times 55) / (10 \times 45) = 11$. Smokers' odds of disease are 11× higher. One 2×2 table, one ratio — the entire evidentiary basis of early smoking epidemiology was arithmetic like this.

Module Summary

- Crosstabs with correct percentages; χ^2 tests independence (expected ≥ 5).
- RR for cohorts, OR for case-control; report with CIs.
- Phi/Cramér's V for larger tables.
- Logistic regression: logit link; coefficients = log odds ratios.

Exam Pointers (Very Important)

- ★ Run a χ^2 test of independence on a given 2×2 table.
- ★ Compute and interpret an odds ratio with its CI.
- ★ When is Fisher's exact test needed?
- ★ Interpret a logistic coefficient $\beta = 0.69$.

MODULE 9: NONPARAMETRIC TESTS AND QUESTIONNAIRE ANALYSIS

Learning Objectives

- Select rank-based tests for non-normal data
- Analyze Likert-scale questionnaire data
- Assess reliability with Cronbach's alpha
- Perform basic factor-analysis screening

Rank-Based Tests

When normality fails and n is small: **Mann-Whitney U** (2 independent groups), **Wilcoxon signed-rank** (paired), **Kruskal-Wallis** (3+ groups + Dunn post-hoc), **Spearman ρ** (correlation), **Friedman** (repeated measures). They test **ranks/medians**, not means — report **medians and IQRs** with them. Power is slightly lower than parametric when assumptions hold, far more trustworthy when they don't.

Likert Scales and Reliability

Single Likert items are **ordinal** — summarize with medians, test with nonparametrics. **Summed scales** (10 items → total score) are often treated as interval (means, t-tests) — defensible with ≥ 5 points and symmetric wording, but report the choice. **Reverse-score** negatively worded items first! **Cronbach's α** ≥ 0.70 = acceptable internal consistency; low α means the items don't measure one thing — fix the scale, not the analysis. **Item-total correlations** flag bad items.

Validity Screening

Content validity: expert review of items. **Construct validity**: **exploratory factor analysis (EFA)** — do items cluster into expected dimensions? Check **KMO (> 0.6)** and Bartlett's test before EFA; retain factors by **scree plot / eigenvalues > 1** ; rotate (varimax) for interpretation. **Criterion validity**: correlation with a gold standard. A questionnaire is an **instrument** — validate the instrument before trusting its measurements.

Key Terms & Definitions

Cronbach's alpha	Internal-consistency reliability; $\geq .70$ acceptable.
KMO measure	Sampling adequacy for factor analysis; want $> .60$.
Reverse scoring	Flipping negatively worded Likert items before summing.
Construct validity	Evidence the scale measures its intended concept.

Example: The Unreversed Item

A 10-item satisfaction scale: 9 items positive, 1 negative ('I dislike this service') — forgotten reverse-scoring gives $\alpha = 0.31$ and nonsense totals. Reverse that one item: $\alpha = 0.87$. Data quality lives in these details; the fanciest test can't rescue an unscored scale.

Module Summary

- Nonparametrics (Mann-Whitney, Wilcoxon, Kruskal-Wallis, Spearman) for non-normal data.
- Likert items ordinal; summed scales quasi-interval; always reverse-score negatives.
- Cronbach's $\alpha \geq .70$ for reliability; item-total correlations flag bad items.
- Validate with EFA (KMO, scree, rotation) before trusting scores.

Exam Pointers (Very Important)

- ★ Choose the right nonparametric test for three scenarios.
- ★ Compute Cronbach's α interpretation for $\alpha = .65$ vs. $.85$.
- ★ Why must negatively worded items be reverse-scored?
- ★ Outline the steps of an exploratory factor analysis.

MODULE 10: REPORTING RESULTS AND RESEARCH WORKFLOW

Learning Objectives

- Write APA-style results sections
- Build a complete analysis pipeline
- Avoid p-hacking and HARKing
- Present findings with tables and figures

APA-Style Reporting

Template: **descriptives first** (n, means $\pm$ SD or medians/IQR), then **inferential tests** with test statistic, df, p, and **effect size**: 't(48) = 2.31, p = .025, d = 0.65'. Report **exact p-values** (p = .032, not p < .05; p < .001 when tiny). **Tables**: numbered, titled, with notes defining abbreviations. **Figures**: error bars defined in captions (SD? SE? CI?). Never report a p-value without the **descriptives and effect size** that give it meaning.

The Full Pipeline

End-to-end workflow: **1.** import & audit data → **2.** clean (missing, outliers, recodes) → **3.** descriptives + plots → **4.** assumption checks → **5.** inferential tests → **6.** effect sizes + CIs → **7.** tables/figures → **8.** APA write-up — all in **one script**, rerunnable end-to-end. **Version control** (even dated folders) and a **README** describing the data dictionary make the analysis auditable months later.

Research Integrity

p-hacking: trying analyses until p < .05 — inflates false positives; **HARKing**: presenting post-hoc findings as a priori hypotheses. Defenses: **pre-registration** of analysis plans, **correcting for multiple comparisons**, distinguishing **confirmatory vs. exploratory** analysis (explore freely, label honestly). **Open data and code** let others verify. Integrity practices aren't bureaucracy — they're what make statistical evidence **evidence**.

Key Terms & Definitions

APA style	Standard reporting format: statistic(df) = value, p, effect size.
p-hacking	Fishing for significance by trying many analyses.
HARKing	Hypothesizing after results are known.
Pre-registration	Publishing the analysis plan before seeing data.

Example: The Garden of Forking Paths

An analyst tries 20 reasonable specifications; one gives p = .04 and gets published as 'the' analysis. The reported p-value assumed one test, but 20 were run — the true false-positive rate is ~64%. Pre-registration closes the garden gate: one plan, one test, honest p.

Module Summary

- Report: descriptives → test(df) = value, exact p, effect size (APA).
- Pipeline: import → clean → explore → check → test → report, all scripted.
- Avoid p-hacking/HARKing: pre-register, correct, label exploratory honestly.
- Share data + code; integrity makes statistics evidence.

Exam Pointers (Very Important)

- ★ Write an APA results paragraph for a given t-test output.
- ★ What are p-hacking and HARKing? Give defenses.
- ★ Outline a complete reproducible analysis pipeline.
- ★ Why must effect sizes accompany p-values?