INSPECT-LB

Institut National de Santé Publique, d'Épidémiologie Clinique et de Toxicologie

STATISTICS AT A GLANCE:

A Practical Guide for Junior Researchers

JUNE 2024

Introduction

"Publish or perish," they say.

In many fields, including health sciences, the ability to analyze data and draw reliable conclusions through statistical methods is essential for publishing reliable and credible research. This guide aims to equip you with the knowledge and skills to confidently plan, execute, and report statistical analyses by following the basic processes, assumptions, and considerations. Whether you're dealing with descriptive statistics, inferential techniques, or advanced multivariate approaches, this resource will provide valuable insights to help you navigate the complexities of statistical analysis successfully.

Getting Started

1. **Define the Objective:** Define a clear research objective or question you aim to address through statistical analysis.
2. **Define the variables:** Identify the variables you want to investigate link or the groups you want to compare based on your objective.
3. **Formulate Hypotheses:** Formulate precise null and alternative (research) hypotheses that align with your objective.
4. **Develop a Statistical Analysis Plan:** Prepare a statistical analysis plan (SAP) in advance, including dummy tables for presenting your results.
5. **Determine variable types:** Determine the types of variables (categorical, continuous, etc.) you will use to test your hypotheses.
6. **Calculate sample size:** Calculate the minimum sample size required to demonstrate your primary hypothesis using appropriate formulas or software, e.g., G*Power, Düsseldorf, Germany, or EpiInfo™, CDC, Atlanta, USA.
7. **Select Applicable Scenarios:** Based on the variable types, select the most applicable scenario from the provided tables (Tables 1-4).
8. **Choose Statistical Test(s):** Choose the appropriate statistical test(s) and decide on the format for presenting the results. Finalize your dummy tables accordingly.
9. **Execute the Planned Analyses:** After collecting data according to a standardized form or questionnaire, execute the planned analyses as per the SAP using statistical software, and copy the results into your dummy tables.

Tips for Effective Statistical Analysis

- **Know Your Objectives:** Avoid random statistical procedures. A well-prepared statistical analysis plan (SAP) ensures organized and time-efficient data analysis.
- **Check Normality:** Check the normality of all continuous variables of your study before starting. Avoid checking normality for multinomial or dichotomous data, as it would be irrelevant.

- **Clean Your Data:** Check plausibility and assess a random sample of questionnaires/forms. Correct or remove false or I values. Do not work on raw databases.
- **Document All Your Analytical Steps and Procedures:** Keep a syntax of all your operations to save time in case you need to replicate or repeat the analysis.
- **Assess Interactions Only When Necessary:** Focus on evaluating main effects unless your research question involves interactions.
- **Validate Scales:** Use a previously validated scale for a similar population. Otherwise, re-validate the scale for your specific context.

Describing Statistical Analyses in the Methods Section

The following examples are written in the past tense to show how the methods section should be presented.

Descriptive analysis

- **Software:** SPSS (Statistical Package for Social Sciences) version 28.0 was used to examine the collected data. For the descriptive analysis, means and standard deviations were calculated for quantitative variables, while frequencies and percentages were reported for categorical variables.
- **Non-Normal Distribution:** Medians and interquartile ranges were presented for variables with non-normal distribution. Normality was visually assessed using histograms, ensuring skewness and kurtosis values are less than 2 in absolute value. These parameters are consistent with normality if the sample size is > 300.

Bivariate analysis

- **Significance Threshold:** A p-value lower than 0.05 was considered significant. The Bonferroni correction was applied for multiple comparisons, particularly in *post hoc* tests.
- **Continuous variables:** The Student's t-test was used to compare means between two groups and ANOVA for three groups or more after checking the homogeneity of variances with Levene's test. If variances were not homogeneous, the corrected t-test/Mann-Whitney U test and the Kruskal-Wallis test were employed, respectively. Significant ANOVA and Kruskal-Wallis results were followed by *post hoc* analyses with Bonferroni adjustment.
- **Correlation coefficient:** Spearman correlation coefficient was calculated for associations between continuous variables, and gamma coefficient was used for ordinal variables.
- **Paired Tests:** Repeated measures ANOVA was used for paired or repeated measures.
- **Categorical Variables:** Associations between categorical variables were assessed using the chi-square test or Fisher's exact test when expected cell counts were lower than 5.
- **Agreement Measure:** The Kappa coefficient was measured to assess agreement between measurement types.

Multivariable analysis

- **Continuous Dependent Variables:** After verifying the normality of the residues, the linearity of the relationship, the absence of multicollinearity, and the homoscedasticity assumptions, multiple linear regressions were carried out for the multivariable analysis to evaluate the correlates of dependent variables throughout the entire sample. A stepwise approach was employed to arrive at the most parsimonious model. The supplied data included the beta coefficient, its 95% Confidence Interval, and the p-value.
- **Qualitative Dependent Variables:** Logistic regression models were utilized for multinomial or dichotomous dependent variables, and the Hosmer-Lemeshow test was employed to assess model fit.
- **Repeated Measures:** Generalized Estimating Equations were utilized for repeated measures with missing values and/or non-normal distribution, accounting for correlated observations.
- **Variable Selection:** Independent variables included in the models were those with conceptual relevance and those with $p < 0.2$ in bivariate analyses, while considering the maximum number of variables allowed based on the sample size. Sociodemographic factors and other relevant independent variables were included as needed.

Scale validation

- **Construct Validity:** The construct validity of newly developed scales was assessed using exploratory factor analysis, after verifying the assumptions of sampling adequacy (inter-item correlations, anti-image matrices, Keiser-Meyer-Olkin values, Bartlett's test of sphericity, and commonality). Items were assessed for factor loadings using No/Varimax/Promax rotation.
- **Structural Validity and Reliability:** Structural validity was assessed using Spearman correlation coefficients, while reliability (internal consistency) was evaluated using Cronbach's alpha.
- **Test-Retest Reliability:** The test-retest reliability was evaluated using the Intraclass Correlation Coefficient (ICC), while convergent/concurrent/discriminant validity with other scales and face validity with anchor questions were assessed using Spearman coefficients.
- **Sensitivity to Change:** Sensitivity to change and minimal important difference were assessed as appropriate.

Additional Tips for Writing the Methods Section

Use a guiding checklist, according to your study type: https://inspect-lb.org/methodological-checklists/

Tips for Discussing the Results

- **Summarize Key Results:** Start the discussion by summarizing the most important results that respond to your objective(s), focusing on descriptive and multivariable results rather than bivariate analyses.
- **Compare with the Literature:** Compare each result with existing literature, highlighting similarities and attempting to explain differences. Avoid repeating background information already covered in the introduction or using ideas from the literature that are not relevant to your results.
- **Differentiate Causation from Association:** When reporting associations from observational studies, use cautious language and avoid implying causation, which can only be established through well-designed experimental studies.
  If investigating potential causal relationships, consider applying Bradford Hill causality criteria, such as the strength of associations, dose-response relationships, cause-effect specificity, reversibility, temporality, biological plausibility, and consistency with other evidence. The more elements you can demonstrate, the clearer the causality between exposure and outcomes. Otherwise, it is only an association.
- **Discuss Limitations:** In this section, address study design, sample size considerations (particularly for non-significant results), and potential biases, e.g., selection, information (differential and non-differential), and residual confounding.
- **Adopt a Modest Tone:** Acknowledge the strengths of your study while suggesting further research to confirm or build upon your findings, except in the case of conclusive meta-analyses.
- **Practical Implications:** Suggest implications for clinical, public health, social, or political contexts.
- **Conclude Effectively:** End with a conclusion that directly answers your research question(s) stated in the objective(s).

Tables

Table 1: Bivariate Analysis for Comparing Two Groups. This table outlines the appropriate statistical tests for bivariate analyses when comparing two groups, whether the grouping variable is the dependent or independent variable.

Table 2: Bivariate Analysis for Comparing Multiple Groups. This table provides guidance on selecting suitable statistical tests for bivariate analyses involving more than two groups, regardless of whether the grouping variable is the dependent or independent variable.

Table 3: Bivariate Analysis for Continuous Variables. This table focuses on the appropriate statistical tests for bivariate analyses when both the dependent and independent variables are continuous (quantitative).

Table 4: Multivariable Analysis Techniques. This table covers advanced multivariable analysis methods, offering recommendations on selecting the appropriate statistical techniques based on the characteristics of the dependent variable and the assumptions of the analysis.

Table 1. Bivariate Analysis for Comparing Two Groups

Variables	Group a***	Group b***	Statistical test result (p-value)
Quantitative normal* variable 1; Homogeneous variances**	Ma1(SDa1)	Mb1(SDb1)	Student T-test
Quantitative normal* variable 2; Heterogeneous variances**	Ma2(SDa2)	Mb2(SDb2)	Corrected T-Test; or Mann Whitney non-parametric test
Quantitative ordinal/non-normal* variable 3	Median 3a (IQR)	Median 3b (IQR)	Mann Whitney non-parametric test Gamma coefficient can be used
Dichotomous variable 4 **Modality 1** **Modality 2**	 Na41(%) Na42(%)	 Nb41(%) Nb42(%)	 Chi2 or Fisher exact test (if expected count less than 5); OR or RR [95% CI]
Qualitative variable 5 **Modality 1** **Modality 2** **Modality 3**	 Na51(%) Na52(%) Na53(%)	 Nb51(%) Nb52(%) Nb53(%)	 Chi2 or Fisher exact test (if expected count less than 5); use logistic regression for two by two comparison {OR or RR [95% CI]}
Quantitative variable 6 **Comparison between two dependent groups (v6 normal*)** **Comparison between two dependent groups (v6 abnormal*)**	Ma6(SDa6)	Mb6(SDb6)	 Paired sample's Student test Wilcoxon signed-rank test
Qualitative variable 7 **Comparison between two dependent groups** **Agreement between variable 7 and 2 measures (a/b)** **Agreement between variable 7 and more than 2 measures**	Na7(%)	Nb7(%)	 Mc Nemar test Cohen's Kappa coefficient Fleiss' Kappa coefficient
Ordinal variable 8 **Comparison between two dependent groups** **Agreement between variable 8 and 2 measures (a/b)** **Agreement between variable 8 and more than 2 measures**	Na7(%)	Nb7(%)	 Wilcoxon signed-rank test Weighted Kappa coefficient Fleiss' kappa coefficient

Normality is tested using a visual inspection of the histogram and checking skewness and kurtosis* **(<|2|)*; testing normality is always recommended, and mandatory for sample sizes n<100; **Variances are compared using Levene's test; ***a=Group a; b=Group b; M=Mean; SD= Standard Deviation; N=Frequency (number of individuals in the group); IQR=Interquartile Range; OR=Odds Ratio (in all kinds of studies; mainly in a case-control study); RR=Relative Risk (in cross-sectional as a Prevalence Ratio, in cohort studies as an Incidence Ratio, or in an experimental study as a Hazard Ratio); 95% CI=95% Confidence Interval*

Table 2. Bivariate Analysis for Comparing Multiple Groups

Groups	Measure*	Statistical test result - p-value (overall) & 95% CI	p-value *(post hoc)*
Qualitative variable **Group 1** **Group 2** **Group 3**	Quantitative M1(SD1) M2(SD2) M3(SD3)	ANOVA (normal** distribution and homogeneous variances***); Kruskall-Wallis (if non-normal** distribution or heterogeneous variances***; you might compare Medians[IQR] here)	Bonferroni adjustment for two-by-two groups comparison
Qualitative variable **Group 1** **Group 2** **Group 3**	Qualitative N1(%) N2(%) N3(%)	Contingency table Chi2 or Fisher exact test (if expected count <5) Phi coefficient (Phi) and Goodman & Kruskal's lambda (λ) can also be used	Use logistic regression for two-by-two groups comparison
Qualitative variable **Group 1** **Group 2** **Group 3**	Ordinal; dichotomous	Contingency table Chi2 or Fisher (if expected count <5) Rank bi-serial correlation coefficient can also be used Cochrane's Q (proportions consistency; assesses whether % of successes is the same between groups)	Use logistic regression for two-by-two groups comparison
Repeated measures **Time 1** **Time 2** **Time 3**	Quantitative M1(SD1) M2(SD2) M3(SD3)	Repeated Measures ANOVA Friedman test if non-normal** distribution	Bonferroni adjustment for two-by-two groups comparison
Repeated measures **Time 1** **Time 2** **Time 3**	Qualitative N1(%) N2(%) N3(%)	Cochrane's Q test Exact version, if small sample size (n<4)	Use GEE for two-by-two measures' comparison

**1=Group 1; 2=Group 2; M=Mean; SD=Standard Deviation; N=Frequency (number of individuals in the group); IQR=Interquartile Range; GEE=Generalized Estimating Equation; **Normality is tested using a visual inspection of the histogram and checking skewness and kurtosis* (<|2|)*; testing normality is always recommended, and mandatory for sample sizes n< 100; ***Variances are compared using Levene's test*

Table 3. Bivariate Analysis for Continuous Variables

Variables	Measure	p-value
Quantitative variable (both have normal* distribution)	Correlation coefficient R	Pearson; 95% CI**
Quantitative variable (at least one variable has non-normal* distribution)	Correlation coefficient R	Spearman; 95% CI**
Ordinal by non-normal quantitative (medians' comparison across ordinal groups)	Jonckheere-Terpstra test	Medians; IQR
Ordinal by normal quantitative (means comparison across ordinal groups)	Trend test (or ANOVA with post hoc tests)	Means; 95% CI**
Ordinal by ordinal	Kendall Tau-b; Gamma coefficient	95% CI**
Repeated measures (test-retest) – Single rater reliability!	Intraclass correlation coefficient-single	95% CI**
Repeated measures (test-retest) – Different raters' reliability!	Intraclass correlation coefficient-average	95% CI**
Special situation: Agreement between a continuous and a dichotomous variable (Linear association and normal distribution)	Point bi-serial correlation coefficient (Rank bi-serial if assumptions not fulfilled)	95% CI**

**Normality is tested using a visual inspection of the histogram and checking skewness and kurtosis (<\|2\|); testing normality is always recommended, and mandatory for sample sizes n< 100; **95% CI=95% Confidence Interval; ! This is a case of agreement; the Bland-Altman plot can also be used.*

Table 4. Multivariable Analysis Techniques

Dependent variable	Independent variables	Multivariable Operation	Association measures to report
Dichotomous	Any type	Dichotomous logistic regression*	ORa [95% CI]; p-value
Ordinal	Any type	Ordinal logistic regression	aBeta [95% CI]; p-value
Multinomial	Any type	Multinomial logistic regression	aBeta [95% CI]; p-value
Time to event	Any type	Cox regression**	HRa [95% CI]; p-value
Count	Any type	Poisson regression	RRa [95% CI]; p-value
Continuous	Continuous; dichotomous	Multiple regression***	aBeta [95% CI]; p-value
Continuous	Any type	GLM-MANOVA	aBeta [95% CI]; p-value; EMM/group
Multiple (continuous, dichotomous)	Any type	GLM-MANCOVA	aBeta [95% CI]; p-value; EMM/ group
Repeated measures (no missing values; normal distribution)	Any type	GLM-RANCOVA	aBeta [95% CI]; p-value; EMM/group
Repeated measures (missing values or non-normal distribution)	Any type	GEE on re-structured data	aBeta [95% CI]; p-value; EMM/group
Latent variable! (variables grouping to form subscales and a total scale)	Continuous; Likert; dichotomous	Exploratory factor analysis Confirmatory factor analysis	Factor loading (check assumptions); reliability measures; fit measures

**Hosmer-Lemeshow test should be non-significant; **Proportional hazard hypothesis should be fulfilled (log-minus-log parallel); ***Assumptions: Linearity, residuals' normality, homoscedasticity, non-colinearity; GLM=General Linear Model; MANOVA=Multicariate Analysis of Variance; MANCOVA= Multivariate Analysis of Covariance; RANCOVA=Multivariate Analysis of Covariance; GEE=Generalized Estimating Equation; aBeta=Adjusted Beta; ORa=Adjusted Odds Ratio; HRa=Adjusted Hazard Ratio; RRa=Adjusted Relative Risk; EMM=Estimated Marginal Mean; [95% CI]=95% Confidence Interval; !To group people into profiles (homogeneous groups of people), use a cluster analysis.*