

Use of Statistical Methods in Bachelor and Master Theses: Data Requirements, Sample Sizes, and Practical Guidance with Methods and Examples

Dr. Binu Daniel, *Professor of Finance, CBS University of Applied Sciences, Berlin, Germany*
Email: b.daniel@cbs.de

Abstract

This article provides a structured and practical overview of the use of statistical methods in Bachelor's and Master's theses, with a particular focus on data sources, sample size requirements, and methodological decision making. It addresses common challenges faced by students, especially within limited research time frames, by linking methodological rigor with feasibility. The paper distinguishes between three primary data sources: interview based data, questionnaire based data, and secondary data, outlining their characteristics, advantages, and limitations. Clear guidelines for sample sizes are presented for both qualitative and quantitative research, emphasizing concepts such as data saturation and statistical power. In addition, the article explains key statistical methods, including descriptive statistics, correlation analysis, regression analysis, t tests, ANOVA, and qualitative content analysis, supported by practical examples. The importance of data quality is highlighted through the criteria of validity, reliability, and objectivity. Furthermore, the paper offers guidance on selecting appropriate methods based on research objectives and identifies common mistakes to avoid. Overall, the article aims to support students in conducting scientifically sound, methodologically appropriate, and practically feasible empirical research.

1. 1. Introduction

Empirical research is an important component of Bachelor's and master's theses, especially in the areas of business and social sciences. Students will be asked to design a study, collect relevant data, use acceptable statistical tools, and interpret their findings in a scientifically rigorous manner. Many students, however, have difficulty with methodological choices, particularly with respect to data requirements and sample sizes. These questions are often asked: How many interviews are enough? How many survey replies are needed? Is secondary data enough?

These issues are compounded by tight time periods, usually between eight weeks and three months. Therefore, the choice of suitable procedures is of both theoretical and practical interest. A well-designed study needs to balance rigour with viability. This article gives an organised overview of often used data sources and statistical approaches. I will combine paragraphs with bullet points and support the discussion with proven academic references.

2. Types of Data Sources for Thesis Research

The choice of data source profoundly impacts the research strategy and defines the type of analysis that can be performed. The three main types of data used in academic theses are interview-based data, questionnaire-based data and secondary data (Creswell, 2014; Saunders et al., 2019).

2.1 Data from Interviews

Interview research is qualitative and is aimed at in-depth exploration of subjective experiences, perceptions and meanings. It does not rely on massive datasets but rather seeks to gain rich, detailed insights (Bryman, 2016) unlike quantitative research (Bryman, 2016). This method is particularly well-suited for exploratory research where theoretical knowledge is inadequate. For example, a study of leadership behaviour in startups can involve interviewing Company founders on decision-making processes. Common features include small,

targeted sample, Open-ended and flexible questions, Focus on depth rather than generalisation, Data collection and analysis for each iteration.

In qualitative research, data saturation is a significant notion defined as the point at which further interviews yield no new information or insights (Guest et al., 2006). Instead than targeting a specific number of interviews, researchers attempt to get to this stage of saturation. In practice, the basic norms are that you do about 5 to 10 interviews in a Bachelor thesis and roughly 8 to 15 interviews in a Master thesis. However, these statistics are not definitive, as the actual sample size relies on factors such as the intricacy of the research issue and the diversity of the persons engaged.

2.2 Survey data

The questionnaire-based research is quantitative, frequently utilised in student theses, and allows for statistical analysis and hypothesis testing (Field, 2018). Data is collected through structured instruments, often using Likert scales or closed-ended questions. For example, a student could collect survey data from university students to explore the relationship between social media use and academic achievement. The main qualities are: Large sample sizes, Consistent data collecting, Data appropriate for statistical testing, numerical.

Sample size is an important issue in quantitative research since it directly affects statistical power, precision, and the validity of the results (Hair et al., 2019). As a thumb rule, the suggested sample size depends on the level of study. For a Bachelor thesis, you will normally need around 50 to 100 responses, and for a Master thesis, the recommended sample size is around 100 to 250. If you want to do more complex statistical techniques (like regression analysis), you should look for at least 150 observations to have solid results. Moreover, it is generally agreed that researchers should have a minimum of 5 to 10 observations per variable in the study, which is essential for the robustness and stability of the statistical model (Hair et al., 2019).

2.3 Secondary Data

Secondary data is data that has already been acquired by other academics, institutions or organisations. Examples include sources such as the World Bank, OECD, government databases and company reports (Saunders et al., 2019). This form of data is very valuable for students in time-pressured situations because it does not need the collecting of original data. For example, a student might look at the relationship between GDP growth and unemployment, utilising publicly available macroeconomic statistics. The key advantages of secondary data are that it saves time and gives you access to enormous data sets that are generally of good quality. However, there are also considerable restrictions, such as a lack of control over the initial data collection process and the potential that the data may not completely match the specific research objectives. So ultimately, the acceptability of secondary data is determined by its dependability, completeness and relevance to the study issue (Gujarati & Porter, 2009).

3. Data needs and sample size

Having the right amount of data is important for the validity and reliability of study findings.

Research Based on Interviews

Qualitative research is primarily concerned with depth and meaning rather than statistical representation, which implies that the sufficiency of the sample size is judged by saturation, not by fixed numerical cut-offs (Creswell, 2014). This means that they keep collecting data until they are no longer gaining fresh insights or themes from further interviews. To ensure quality and consistency in the research process, a number of critical issues should be considered, including careful selection of relevant participants and the use of a structured interview guide to maintain focus and consistency in how interviews are conducted.

In practice, interviews usually last 30-60 minutes, and the overall number of interviews is decided by the point at which fresh information is no longer emerging, rather than a set number.

Research Using Questionnaire

The size of the sample is very important in quantitative research to determine the accuracy and generalisability of the results. Larger sample sizes tend to give more reliable and consistent estimates (Field, 2018). The minimal number of replies necessary is frequently determined by the sort of statistical analysis that is being conducted. For example, descriptive statistics normally require at least 50 replies to produce relevant summaries, while correlation analysis usually requires 80 or more responses to find correlations between variables. More complex procedures such as regression analysis generally require 100 to 150 or more data for robust and reliable results.

To obtain these sample sizes, it is vital to employ successful tactics for improving response rates, including keeping surveys short and easy to comprehend, employing online distribution channels to reach bigger audiences, and targeting respondents relevant to the study topic.

Secondary Data

The amount of data needed in empirical research is largely determined by the sort of analysis being performed. In time series analysis, for example, it is often advised that there should be at least 20 to 30 observations in order to be able to discover meaningful trends and patterns across time (Gujarati & Porter, 2009). Cross sectional analysis is done across different units (businesses or countries) at a single point of time. Generally something like 20 units is considered minimum acceptable. For panel data analysis, which integrates time series and cross-sectional dimensions, researchers are expected to have many units observed throughout several time periods. This combination provides for more robust and thorough research but requires a sufficiently large and structured data collection to produce credible results.

4. Statistical Methods Important with Examples

Statistical methods allow academics to methodically analyse data and evaluate ideas.

4.1 Summary Statistics

Descriptive statistics are used to describe the basic properties of the data in a study. They provide a simple summary about the sample and the measures. Descriptive statistics are a powerful way to provide basic insights into the data to be analysed before performing more advanced analyses (Field, 2018). They assist researchers to find patterns, central tendencies, and variability in the data. The most frequent measurements are the mean, which reflects the average value, the median, which indicates the middle value in an ordered data collection, and the standard deviation, which reveals the degree of variance or dispersion in the data. For example, a low standard deviation in student satisfaction scores with a mean of 3.8 shows that the responses are tightly grouped around the mean, showing excellent consistency among participants.

4.2 Correlation Analysis

The amount of data needed for empirical research relies heavily on the analysis technique. In time series analysis, for example, it is often advised that there should be at least 20 to 30 observations in order to be able to discover meaningful trends and patterns across time (Gujarati & Porter, 2009). In cross sectional analysis where data are collected at the same point in time for multiple units such as enterprises or countries, a minimum of about 20 units will usually be sufficient. Panel data analysis is a combination of the time series analysis and the cross-sectional analysis; it is a set of several units observed throughout several time periods. While this

combination enables a more robust and thorough analysis, it also necessitates a sufficiently large and structured dataset to generate credible results.

4.3 Regression analysis

Regression analysis is used to study causal relationships and to predict outcomes (Gujarati & Porter, 2009), and it includes both simple regression analysis and multiple regression analysis. Consider a study on academic performance. Here the score in the exam is the dependent variable, and the independent variables may be the number of hours spent studying, sleeping, and using social media. This is a classic example of regression analysis. With this form of model, the researchers may see the effect of each factor on academic outcome, holding the other factors constant. The analytical results can imply, for instance, that studying more hours has a beneficial impact on exam grades, whereas social media usage is negatively linked with them. Regression analysis requires several important conditions to be met for reliable and valid results. These include adequate sample size, absence of strong multicollinearity between independent variables, and key statistical assumptions such as linearity and normality (Hair et al., 2019).

4.4. T-Test

A t test is a statistical procedure that is used to compare the means of two groups to evaluate whether the observed difference is statistically significant (Field, 2018). This allows researchers to analyse if any variation between the averages of groups is likely just random chance or a true effect. For example, a researcher could compare exam scores between students who took online learning and those who learned in the traditional classroom. If the p value calculated is less than the usually recognised threshold of 0.05, then the difference between the two groups is regarded to be statistically significant. This suggests that the type of learning method may have a meaningful impact on academic performance.

4.5 ANOVA (Analysis of Variance)

Analysis of variance (ANOVA) is a statistical procedure used to compare the means of more than two groups to discover if there are statistically significant differences between them (Field, 2018). ANOVA is efficient and reliable since it does a single comparison of all groups rather than performing several comparisons of each group to every other group. For example, a researcher would want to examine satisfaction levels among Bachelor students, Master students and PhD students to see if their experiences differ. The results of ANOVA tell us if there is at least one group mean that is substantially different from the others, however typically we need to do further post hoc tests to determine exactly which groups vary.

4.6 Qualitative Content Analysis

Qualitative content analysis is a systematic technique that enables researchers to evaluate interview material in a structured and relevant way (Mayring, 2015). The first stage is usually to transcribe the interviews, in order to make a written record of the data. Now we code the responses, which means assigning labels to important assertions or concepts. These codes are then combined into larger categories, which help to organise the data into relevant topics. Lastly, the researcher examines these patterns to derive conclusions. For instance, in research investigating job happiness, the data collected from interviews can be categorised into topics such as work environment, salary, and career advancement, which, in combination, present a holistic view of the elements affecting employee satisfaction.

5. Data Quality Requirements

The integrity of any research study depends critically on the quality of the data obtained, ensuring data quality is therefore paramount in guaranteeing the scientific rigour of the research (Saunders et al., 2019). This requires fulfilling three main criteria. Validity relates to whether the data genuinely measure what they are supposed to measure. Second, reliability indicates how much the results are consistent, that is, if the same approach was applied again under comparable circumstances, it would generate similar results. Thirdly, objectivity means that the research method is free from personal bias and that the outcomes of the research are not influenced by the subjective beliefs of the researcher. Failure to appropriately meet these criteria, e.g. by poorly designed questionnaires or inconsistent data collection techniques, could greatly impair the overall quality and reliability of the research.

7. Selecting the Right Approach

The selection of research method should always be consistent with the study purpose and available time and resources since the choosing of a suitable method is vital for the production of meaningful and achievable results (Creswell, 2014). Interviews are especially helpful when you want to investigate new or complicated issues and when you want to obtain a deep grasp of participants' experiences or perspectives. Questionnaires are better appropriate for testing associations between variables and performing statistical analysis with a bigger sample. Secondary data is worth considering when there is limited time and when reliable and relevant datasets are already available. This frees the researcher to concentrate more on analysis than data acquisition.

8. Avoid These Common Mistakes

Many of the problems students face in the process of thesis research are similar and can have a major impact on the quality and authenticity of their work. One of the most common problems is the use of a small sample size in quantitative studies, which results in weak and incorrect statistical results. The analysis is complex and difficult to interpret. Another common difficulty is having too many variables when you have inadequate data. Poor questionnaire design is also a key concern, as ambiguous or biased questions can distort the data collected. Furthermore, some students lack a defined topic for their research and hence have an unfocused study and weak findings. Finally, there is a tendency to apply complex statistical procedures without a complete grasp of them which can lead to wrong analysis and misinterpretation of results.

9. Conclusion

Knowledge of data sources, sample size requirements and analytical methodologies is fundamental for the correct use of statistical methods in Bachelor and Master theses. Interviews bring depth to study, questionnaires allow for statistical testing, and secondary data is cost-effective. Students can generate high quality and scientifically valid work by aligning methodological choices with research objectives and time restrictions. More often than not, simplicity, clarity and methodological correctness are more important than complexity.

References:

- Bryman, A. (2016). *Research methodologies in social relations (5th ed.)* Oxford University Press.
- Creswell, J.W. (2014). *Research design: Qualitative, quantitative, and mixed techniques approaches (4th ed.)*. SAGE Publications.
- Field, A. 1984. (2018). *Discovering statistics using IBM SPSS statistics (5th ed.)*. SAGE Publications
- Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics (5th ed.)* McGraw Hill.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed) Cengage Learning

Mayring, P. (2015) Qualitative content analysis: Theoretical underpinning, fundamental processes and software solution. Klagenfurt.

Saunders, M., Lewis, P., & Thornhill, A. (2019). Research methodology for business students, 8th edition. Pearson.