TECHNISCHE
UNIVERSITÄT
WIEN

DIPLOMARBEIT

Normality Tests for small Sample Sizes: A Comparison with Power Simulation

ausgeführt am

Institut für
Finanz- und Versicherungsmathematik
TU Wien

unter der Anleitung von

Senior Scientist Priv.Doz.in. Tijana Levajkovic, PhD

durch

Lisa Retzer, BSc
Matrikelnummer 11802445

Wien, 12. Mai 2025

Kurzfassung

Im Rahmen eines Datenerhebungsprozesses bei der UNIQA Insurance Group stellte sich die Frage, ob eine Teilmenge der beobachteten Variablen als normalverteilt angenommen werden kann. Diese Fragestellung bildet den Ausgangspunkt dieser Diplomarbeit, die sich mit statistischen Tests und grafischen Methoden zur Bewertung der Hypothese befasst, dass eine Stichprobe von einer Normalverteilung stammt.

Aufgrund historischer Änderungen in der Datenerhebung sowie in Berechnungsverfahren oder -annahmen sind die verfügbaren Stichproben sehr klein. Dies stellt eine besondere Herausforderung dar, da viele Tests auf asymptotischen Eigenschaften beruhen, die bei kleinen Stichproben nicht erfüllt sind. Daher liegt der Fokus der Tests und der durchgeführten Powersimulation auf sehr kleinen Stichproben, die dem beobachteten Datensatz in Streuung und Größenordnung ähneln. Obwohl an die Tests unter solchen Bedingungen keine hohen Erwartungen gestellt werden können, besteht das Ziel darin, eine Auswahl robuster Verfahren zu auszuwählen, die dennoch zuverlässige Aussagen zur Gültigkeit der Normalverteilungsannahme ermöglichen.

In dieser Diplomarbeit werden sowohl grafische Methoden vorgestellt, die einen Einblick in die Struktur und Streuung der Daten erlauben, als auch statistische Tests sowie die theoretischen Grundlagen, auf denen sie basieren. Alle Verfahren wurden in R implementiert. Der verwendete Code ist im Appendix beigelegt, sodass die Simulation reproduzierbar ist und leicht auf größere Stichproben oder alternative Verteilungen erweitert werden kann.

Für die Powersimulation werden mehrere nicht-normalverteilte Alternativverteilungen mittels Maximum-Likelihood Schätzung an die realen Stichproben angepasst, um zufällige Datensätze zu erzeugen. Anschließend werden die Tests auf diesen Datensätzen ausgewertet, um ihr Verhalten untersuchen zu können.

Die Ergebnisse zeigen, dass einige Tests in kleinen Stichproben eine höhere Power haben als andere. Darunter sind der Shapiro-Wilk Test, eine robuste Version des Jarque-Bera Tests sowie modifizierte Versionen des Anderson-Darling und des Cramér-von-Mises Tests, die Korrekturen für die Stichprobengröße beinhalten. Zusätzlich bietet diese Diplomarbeit eine Anleitung zur Anwendung der vorgestellten Methoden auf reale Datensätze und liefert damit einen Rahmen für statistische Entscheidungen bei der Überprüfung der Normalverteilungsannahme.

Abstract

As part of a data gathering process at the UNIQA Insurance Group, the question arose whether a subset of observed variables could be reasonably assumed to follow a normal distribution. This motivated the central topic of this thesis: the evaluation of statistical tests and graphical methods for assessing the hypothesis that a sample originates from a normal distribution.

Due to historical changes in the data collection and calculation procedures, the observed samples are very small. This presents a particular challenge, as many statistical tests for normality rely on asymptotic properties that may not hold for small samples. Therefore, the comparison of tests in this thesis, along with a power simulation, focuses specifically on very small sample sizes that resemble the observed data in both spread and magnitude. While normality tests are not generally expected to perform well under such conditions, the goal is to identify a selection of robust procedures that can provide reliable insight into the validity of the normality assumption.

This thesis presents both graphical methods, which offer intuitive insights into the shape and spread of the data, and statistical tests, including background on the theoretical ideas they are based on. All methods are implemented in R, with accompanying code provided in the appendix, enabling full reproducibility and allowing the simulation framework to be easily extended to larger samples or alternative distributions.

For the power simulation, several non-normal distributions are fitted to the observed samples using maximum likelihood estimation, providing a realistic foundation for generating random data with similar characteristics. The performance of the tests is then evaluated on these simulated datasets to gain insight into their behavior.

The results show that some tests demonstrate a more satisfying power than others in the small-sample setting. In particular, the Shapiro-Wilk test, a robust version of the Jarque-Bera test, and modified versions of the Anderson-Darling and Cramér-von Mises tests, which contain adjustments for the sample size, stand out as the more effective options for evaluating normality under such constraints. Finally, this thesis provides practical guidance on how these methods can be applied to real-world data, offering a concrete procedure for statistical decision-making when normality is in question.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 12. Mai 2025

Lisa Retzer

Contents

List of Figures

List of Tables

1. Introduction

The normal distribution is one of the most well-known distribution families in statistical analysis, having a fundamental role. On the theoretical side, it serves as the focal point of several important results, most notably the Central Limit Theorem, which explains the convergence of statistics toward normality. On the practical side, the normal distribution frequently appears in the modeling and analysis of real-world data, where it often provides a reasonable approximation of underlying patterns.

Many techniques, including t-tests, analysis of variance or regression procedures, rely on a normality assumption of the data to produce valid results. While these methods might remain robust under mild deviations, more substantial deviations from normality, such as heavier tails or skewness, can significantly impact their performance, especially in small sample sizes. Therefore, assessing whether a sample originates from a normal distribution is a crucial step for many methods of statistical analysis.

Furthermore, the structure of the probability density function of the normal distribution, characterized by a symmetric bell-shaped curve and an exponential term, gives it a range of properties, making it particularly useful for modeling.

To evaluate normality, a wide range of tools is available. For this thesis, they are divided into two categories: graphical methods and statistical hypothesis tests. Graphical methods such as histograms, boxplots or Q-Q plots provide a visual representation of the data and can reveal structural patterns, outliers, and deviations from normality that may influence the assessment of normality. However, these methods do not follow a standardized decision-making process and interpretation is inherently subjective, particularly when applied to small samples, where random variation can introduce bias into the visual representation.

Statistical hypothesis tests for normality offer a more objective and reproducible way to evaluate the null hypothesis that a sample is coming from a normal distribution. Presented in this thesis are regression-based tests, like the Shapiro-Wilk test, moment-based tests, like the Jarque-Bera test, and goodness-of-fit tests, such as the Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests. These tests differ in their approach for detecting departures from normality.

To compare the performance of the normality tests, this thesis conducts a power simulation, estimating how often each test correctly rejects the null hypothesis when applied to random samples drawn from a range of non-normal distributions. This approach allows for an analysis of the sensitivity of each test to specific types of deviation from normality, such

as lighter or heavier tails, additionally it allows for testing in small sample scenarios. Tests with a low power are less likely to reject non-normal data, which could lead to incorrect conclusions and an inappropriate use of statistical procedures that assume normality. Thus, a power simulation is a valuable tool for understanding the practical reliability of these tests.

While many statistical procedures rely on the assumption of normality, there is no single test that can universally and reliably verify this assumption across all scenarios, particularly when working with small samples. Therefore, the main objective of this thesis is not to identify a perfect test for normality, but rather to offer a deeper understanding of the ideas and procedures behind a selected set of normality tests and graphical methods, acknowledging that these represent only a portion of the many techniques available in statistical literature.

Each method examined in this thesis is based on a different concept for detecting departures from normality, whether through comparing moments such as skewness and kurtosis, measuring deviations between empirical and theoretical distributions, or assessing the linear relation of expected and observed values of order statistics. By exploring what each method measures, this thesis provides information on their fundamental approach and assumptions.

Another aspect of the thesis is the correct application of the normality tests, including the assumptions they rely on, their behavior under small sample sizes, and the impact of unknown parameters. The tests are accompanied by graphical methods such as Q-Q plots and boxplots, which can serve as complementary tools to provide visual context and additional diagnostics. This combined approach helps to form a more robust and nuanced process when evaluating whether a data set is derived from a normal distribution.

The theoretical background of each method is presented in a high-level overview, with the aim of making the underlying ideas accessible while still providing the formulas and statistical concepts that form the basis for their implementation. In particular, this includes how these tests are internally implemented in R and highlighting the specific formulas or transformations used behind the R functions used in practice.

The initial motivation for this study originated directly from a need for practical application. In the context of a data gathering process at the UNIQA Insurance Group for some observed variables, the question arose whether it is reasonable to assume that they follow a normal distribution. Therefore, applying the presented procedures to actual observed data not only illustrates the methods in a realistic setting but can also highlight potential limitations or challenges that might not be visible in simulated environments.

The structure and distributional properties of real-world samples are often more complex than those of random samples drawn in controlled environments, making statistical analysis more challenging, particularly aspects such as small sample sizes, outliers, deviations from theoretical assumptions, or measurement inconsistencies from the data itself. These factors can strongly influence the outcome of statistical tests and graphical methods. By applying the tests to real samples, this thesis aims to give additional context to whether the methods provide reliable guidance under practical conditions and to demonstrate how

theoretical expectations might differ from practical outcomes.

Evaluating methods on real data is particularly important because, in practice, decisions might have to be made based on incomplete or even inconsistent information and limited sample sizes. While simulation studies help to provide an understanding of the behavior of tests under idealized and controlled conditions, real-world applications will introduce the methods to data with additional complexities that might not be fully captured in a simulation. Therefore, the practical evaluation provides valuable insights into the sensitivity and interpretability of the different methods.

The real-world dataset used in Chapters 3 and 8 was provided by the UNIQA Insurance Group. Due to various changes over time, including changes in the method of data collection, changes in calculations or underlying assumptions, or shifts in the origin of the data, the observed samples were reduced in size to ensure consistency in the quality of the data and the validity of any conclusions drawn with the statistical methods. The data set consists of 12 data samples, which are referred to as $S1$-$S12$ in the plots in Chapters 3, 7 and 8. Furthermore, the number of observations in each sample lies between 5 and 10.

Beyond its role in illustrating the application of normality tests and graphical methods, the real-world dataset also serves as a starting point for the random sample generation used in the power simulation study in Chapter 7. To ensure that the alternative distributions used in the simulation closely resemble the structure and characteristics of the observed data, a series of alternative distributions were fitted to the real-world samples using the maximum likelihood estimation method, as presented in Section 2.2. For the simulation study, this allows to produce samples that behave similarly to the real-world data in terms of spread, shape, and magnitude, rather than relying on arbitrary or idealized alternative distributions.

Following the introduction, Chapter 2 introduces the concept of statistical inference, providing the theoretical foundation for statistical hypothesis testing and explaining the idea behind power simulations. This chapter establishes the necessary background on how a statistical test is conducted and how their performances can be compared.

Chapter 3 discusses graphical methods for assessing normality. It explains the procedures behind simple raw data plots, such as histograms and boxplots, and explores more analytical techniques, the so-called probability plots, including Q-Q plots and P-P plots. Additionally, this chapter compares typical graphical behavior observed in simulated data with the graphical displays of the real-world dataset, offering practical insight into the interpretation of these diagnostic plots.

Chapters 4 through 6 present the three main categories of normality tests covered in this thesis. Chapter 4 focuses on the Shapiro-Wilk test, a regression-based normality that evaluates the linear relationship between observed values of the order statistics and their expected values under a normality assumption.

Chapter 5 addresses tests based on distributional moments, particularly the Jarque-Bera test and its robust modification. These tests assess normality by measuring skewness and kurtosis. The robust version is designed to be less sensitive to outliers.

Chapter 6 introduces goodness-of-fit tests, which are not restricted to normality but can evaluate more general distributional assumptions. This chapter presents the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Cramér-von Mises test. In addition, modified versions of the Anderson-Darling and Cramér-von Mises tests are presented that account for the sample size.

Chapter 7 details the power simulation study. It describes the methodology for fitting various non-normal distributions to the real-world samples using maximum likelihood estimation, generating new samples, and evaluating the rejection rates of different tests. The simulation results suggest that none of the original versions of the normality tests perform well on small samples. Only tests that incorporate adjustments or modifications based on simulation studies themselves, such as the Shapiro-Wilk test, the robust version of the Jarque-Bera test, and the adjusted versions of the Anderson-Darling and Cramér-von Mises tests, show more satisfying rejection rates.

Finally, Chapter 8 evaluates the normality tests on a real-world dataset. This application assesses how the methods perform when faced with actual data, where ideal conditions cannot be guaranteed. It provides insight into the limitations observed in the simulation study, offering guidance for real-world applications.

The content and methods presented in this thesis are based on a variety of literature. General information on normality tests and graphical methods was primarily drawn from sources [7] and [20], which offer a broad overview of the theoretical foundation and practical applications of these techniques. For the theoretical background of statistical inference and the procedure of hypothesis testing, main references include [18], [19], [2], [3], and [13]. The Shapiro-Wilk test, as a regression-based approach to normality testing, is discussed in detail in [17]. Literature concerning moment-based tests, in particular the Jarque-Bera test and its robust modification, is found additionally [5] and [10]. The class of goodness-of-fit tests is again covered in [7] and [20], additionally the Kolmogorov-Smirnov test is addressed in [15], the Anderson-Darling test in [6], [8] and [7], and the Cramér-von Mises test in [14], [7], and [1]. These sources collectively form the theoretical background of this thesis.

The contribution of this thesis lies in providing a compact and structured overview of common graphical methods and statistical tests for assessing normality. These methods are not only theoretically described, but also compared through a simulation study that focuses specifically on small sample sizes, a testing environment in which normality tests often behave differently, when they rely on asymptotic conditions. To support reproducibility and further studies, all the R-code used in the simulation and for the tests is included in

the appendix, allowing the study to be easily extended to larger sample sizes or alternative distributions. In addition, the thesis offers practical guidance on how these methods can be applied to real-world data, demonstrating how to interpret results from both statistical tests and graphical diagnostics in application.

Notation

Throughout the thesis, it is assumed a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω denotes the set of all possible outcomes, $\mathcal{F}$ is a σ-algebra of measurable events, and $\mathbb{P}$ is a probability measure defined on $(\Omega, \mathcal{F})$. A real random variable X is a measurable function $X : \Omega \to \mathbb{R}$, assigning a real number to each outcome $\omega \in \Omega$. This thesis focuses exclusively on continuous random variables.

Unless stated otherwise, all random variables considered in this thesis are assumed to be at least square-integrable. In Section 2.3 it is assumed that the fourth moments are finite, which is a common requirement for the validity of certain statistical tests, particularly those based on skewness and kurtosis, such as moment-based tests.

The only exception to this assumption is the Cauchy distribution, which is included in the simulation study due to its special role in power studies. The Cauchy distribution has undefined mean and variance and does not satisfy the square-integrability or finite-moment condition. However, it is often used in simulation studies to evaluate the sensitivity of statistical tests to extreme heavy-tailed behavior.

The cumulative distribution function of a continuous random variable X is defined as

$$F_X(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}\,.$$

It uniquely defined the random variable and it has the following properties:

(i) $F_X(x)$ is non-decreasing and right-continous for $x \in \mathbb{R}$,

(ii) $\lim_{x\to\infty} F_X(x) = 1$ and

(iii) $\lim_{x\to-\infty} F_X(x) = 0$.

This thesis considers only continuous random variable, and therefore the cumulative distribution function of any random variable X is absolutely continuous. That is, there exists a density function $f_X : \mathbb{R} \to \mathbb{R}$ such that:

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\,ds \quad \text{for all } x \in \mathbb{R} \quad \text{and}$$
$$f_X(x) = \frac{d}{dx} F_X(x)\,.$$

Thus, in this thesis, any cumulative distribution function is assumed to be continuous and strictly increasing on its support supp $:= \{x \in \mathbb{R} : F_X(x) \geq 0\}$. Therefore, the inverse of the distribution function F_X^{-1} exists and is defined as:

$$F_X^{-1}(p) = \inf\{x \in \mathbb{R} : F_X(x) \geq p\}, \quad p \in (0, 1).$$

Important examples for continuous distributions are the following.

- The normal distribution is the central object in this thesis. A random variable X is said to follow a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean μ (location parameter) and variance σ^2 (scale parameter), $\mu \in \mathbb{R}, \sigma^2 > 0$, if it has a density function $f_X(x)$ of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , x \in \mathbb{R}$$

and the cumulative distribution function

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-\mu)^2}{2\sigma^2}} \, ds \quad , x \in \mathbb{R}\,. \tag{1.1}$$

A special case is the standard normal distribution $Z \sim \mathcal{N}(0, 1)$ with mean $\mu = 0$ and variance $\sigma^2 = 1$. Its cumulative distribution function $F_Z(x)$ is commonly denoted by $\Phi(x)$.

- The standard uniform distribution on the interval $(0, 1)$, denoted by $U(0, 1)$, is also relevant in this thesis, particularly in the context of goodness-of-fit tests where transformed samples under the null hypothesis are expected to follow a standard uniform distribution. The density function is given by

$$f^*(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} ,$$

and its cumulative distribution function is

$$F^*(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases} .$$

If the connection between a random variable and its distribution function is clear from the context, the cumulative distribution function will be denoted simply by F instead of F_X. The cumulative distribution function of the standard normal distribution is denoted by Φ, and the cumulative distribution function of the standard uniform distribution on $(0, 1)$ is denoted by F^*. To indicate that a random variable X follows a distribution with cumulative distribution function F, the notation

$$X \sim F$$

is used throughout this thesis. Additional distributions used in this thesis, particularly in the power simulation study, are presented and discussed in Chapter 7.

Let $(X_1, X_2, \ldots, X_n), n \in \mathbb{N}$ be a sequence of real-valued random variables, where each $X_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$. These variables are assumed to be independent and identically distributed (i.i.d.) unless stated otherwise. These random variables represent a random sample on the population level, that is, the theoretical object before any data is observed. A realization of this sample refers to the observed values $(x_1, x_2, \ldots, x_n)$ which are outcomes

of a particular observation. This distinction between the random variables $(X_1, X_2, \ldots, X_n)$ and their observed values $(x_1, x_2, \ldots, x_n)$ is maintained throughout the thesis.

Given a sample $(x_1, x_2, \ldots, x_n)$ the empirical distribution function is a non-parametric estimator of the cumulative distribution function F of the underlying population X. It is defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{x_i \leq t\}} ,$$

where $\mathbb{I}_{\{x_i \leq t\}}$ is the indicator function, which equals to 1 if $x_i \leq t$ and 0 otherwise. The empirical distribution function is a step function that increases by $1/n$ at each data point and satisfies $F_n(t) \to F(t)$ almost surely as $n \to \infty$. Given a sample $(x_1, x_2, \ldots, x_n)$, the empirical distribution function F_n is a non-parametric estimator for the cumulative distribution function.

The ordered (sorted) sample, also referred to as the order statistics, is the sorted version of the sample values, denoted by

$$X_{(1)} \leq X_{(2)} \leq \ldots X_{(n)} ,$$

where $X_{(1)}$ coincides with the minimum value of the random sample and $X_{(n)}$ with the maximum value of the sample of size n. The realization of the ordered sample is denoted by $x_{(1)} \leq x_{(2)} \leq \ldots x_{(n)}$. Throughout this thesis, statistical procedures, both graphical and hypothesis test, are applied to the observed sample, while theoretical properties are often stated in terms of the underlying random sample on population level.

Let X be a real-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The k-th moment of X is defined as

$$\mu_k = \mathbb{E}(X^k), \quad k \in \mathbb{N} .$$

The k-th central moment of X, assuming that $\mathbb{E}(X) = \mu$, is defined as

$$\mu'_k = \mathbb{E}\big((X - \mu)^k\big) \quad k \in \mathbb{N} .$$

In particular, the first moment is the expectation and the second central moment is the variance. Given a sample $(x_1, x_2, \ldots, x_n)$ drawn from the distribution of X, the corresponding sample estimates of these moments are defined as follows:

- The sample mean, the estimate of the first moment, is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

- The k-th sample moment is

$$m_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k \quad , k \in \mathbb{N} .$$

- The k-th central sample moment is

$$m'_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \quad , k \in \mathbb{N}\,.$$

- The sample variance, the estimate of the second central moment, is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\,,$$

  while the modified sample variance

$$s_1^2 = \frac{n-1}{n}s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,.$$

2. Statistical Inference

Consider having a fully specified distribution with known distributional parameter θ from a parameter space Θ. Then one can perform calculations, such as calculating probabilities or moments of the distribution, or draw a data sample. Statistical inference does the opposite. Here, a data sample is the starting point of the analysis, through which the underlying distribution is observed, see e.g. [3, p. 127]. This sample is assumed to follow a probability distribution which is at least partly unknown. This missing knowledge results in uncertainty about how the real distribution behaves and how to complete it for modeling purposes. The goal of statistical inference is to derive information about the underlying distribution, the parameter θ or another statistic of the distribution based on the observations [13, p. 3].

This chapter contains a general introduction to statistical hypothesis testing and presents methods for estimating distributional parameters. Both of these areas are part of statistical inference. At the end of this chapter, estimators for the sample skewness and kurtosis are presented, including estimators adjusted for the sample size. These will be used to create a better overview of the simulation results in Chapter 7.

2.1. Hypothesis Testing

At the beginning of statistical testing, there is always a statement about the assumed distribution or population parameter, the so-called null hypothesis. The observed sample is supposed to provide guidance for making a decision about accepting or rejecting the hypothesis, to complete the underlying model with a plausible working hypothesis. This procedure is called the test of hypothesis and aims at measuring how strong the evidence against the null hypothesis is, to give a reliable decision [13] [19] [2].

A hypothesis test always consists of two hypotheses: the null hypothesis H_0 for which a decision is made, and the alternative hypothesis H_1, which is the complementary statement. If the null hypothesis is rejected, the alternative hypothesis is assumed to hold true for the sample [2, p. 373]. Even if a hypothesis is said to hold true for a sample, it cannot be assumed that the hypothesis is literally the truth behind the probability distribution of the sample. It is at most an approximation of reality. Instead of saying that a hypothesis is true or false or that it is accepted or rejected, one should rather consider if it is an appropriate working hypothesis or if the test results show that the sample deviates from the hypothesis to a significant degree to where its assumption becomes questionable [19, p. 87].

Usually, the hypothesis makes a statement about statistics of the distribution, i.e., a function of the random sample, such as expectation or variance, the distributional parameter θ, or even about the class of the distribution [3, p. 153].

When testing if a distributional parameter θ lies in a subset $\Theta_0 \subseteq \Theta$ of all possible values and the assumed distribution F is part of a family of distributions Ω_0. For the statistical tests presented in this thesis, the set Ω_0 consists of the family of all normal distributions $\Omega_0 = \{F_X : X \sim \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ and $\Theta_0 = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$, where F_X is of the form (1.1). The null and the alternative hypothesis have the following forms:

$$\begin{aligned} H_0 : &\quad \theta \in \Theta_0 \text{ and } F \in \Omega_0 \,, \\ H_1 : &\quad \theta \notin \Theta_0 \text{ or } F \notin \Omega_0 \,. \end{aligned} \tag{2.1}$$

For more details on general hypothesis testing see e.g. [2] [13].

Another component that is typically defined for a hypothesis test is the test statistic $W = W(X_1, \ldots, X_n)$, a function of the random sample $(X_1, \ldots, X_n)$, in which the unknown parameter does not appear [2, p. 374].

The test statistic measures the discrepancy between the data sample and the null hypothesis H_0. Under the null hypothesis, the test statistic W follows a certain distribution and is supposed to give a ranking system as to how likely it is that the observed sample is derived from the distribution specified by the null hypothesis [19, p. 87, 88]. With knowledge of the distribution under the null hypothesis and given that it holds true for the sample, one can consider how probable it is to observe a discrepancy, measured by the test statistic W, that is at least as extreme as the observed discrepancy $d := W(x_1, \ldots, x_n)$ in the sample. This probability is the p-value with

$$p\text{-value} = \mathbb{P}(|W| \geq |d|) \,. \tag{2.2}$$

The smaller the p-value, the larger is the discrepancy measured in the sample and the more evidence the test shows that speaks against the null hypothesis. An exceptionally small p-value means that a discrepancy at least as extreme as the one observed in the sample occurs only rarely if the null hypothesis is true. The other case, where the p-value gets exceptionally large, does not have such an intuitive interpretation, i.e., there is not enough evidence to reject the null. Thus, the p-value and therefore hypothesis testing in general is used to find evidence against the null hypothesis instead of evidence that supports it [19, p. 88].

Depending on the definition of the test statistic, either sufficiently large values or extreme values of the test statistic for the particular sample speak against the null hypothesis. For the tests presented in this thesis, the exact definition of the threshold for extreme observations is controlled by the level of significance denoted by α, as mentioned below.

Assuming that the population parameter θ is known, one can clearly assess whether the null hypothesis or the alternative hypothesis is true. However, because of the uncertainty

about the real distribution and because the underlying sample is subject to random variations, there exists the possibility of making a mistake by incorrectly accepting or rejecting the null hypothesis. For making these mistakes, probabilities can be defined which allow for comparison between several tests [2, p. 382] [3, p. 153] [13, p. 56]. For a hypothesis test of the form (2.1) there are two types of errors that can be made, the Type I error and the Type II error. The following table, as seen in [3], shows the possible outcomes of a test and in which case one of the errors is made.

Decision made	H_0 is true (H_1 is false)	H_0 is false (H_1 is true)
Reject H_0 (accept H_1)	Type I error	correct decision
Accept H_0 (reject H_1	correct decision	Type II error

Table 2.1.: Possible outcomes of a test and cases when a Type I and Type II error is made. Table as seen in [3, p. 156].

The Type I error occurs when the null hypothesis H_0 is rejected even though it is actually true. This is the case in the upper left corner of Table 2.1, while the Type II error occurs when the null hypothesis H_0 is accepted even though it is actually not true. This is the case in the lower right corner of Table 2.1 [2, p. 382].

The probabilities of certain outcomes have a specific name. The probability of making a Type I error is called the level of significance α of the test:

$$\alpha = \mathbb{P}(\text{ reject } H_0 \,|\, H_0 \text{ is true }).$$

Given that H_0 is not true, the probability or rejecting the null H_0 is called the power of test and is denoted by $1 - \beta$:

$$1 - \beta = \mathbb{P}(\text{ reject } H_0 \,|\, H_0 \text{ is not true }).$$

Both notations refer to the decision to reject the null hypothesis. These and their counter probabilities are shown in Table 2.2 below, see also [3, p. 157].

Decision made	H_0 is true (H_1 is false)	H_0 is false (H_1 is true)
Reject H_0 (accept H_1)	level of significance $= \alpha$	power of test $= 1 - \beta$
Accept H_0 (reject H_1	$\mathbb{P}(\text{ accept } H_0 \,\|\, H_0) = 1 - \alpha$	$\mathbb{P}(\text{ accept } H_0 \,\|\, H_1) = \beta$

Table 2.2.: Probabilities of possible outcome of the test. Table as seen in [3, p. 157].

The power of a test is a measurement for how sensitive the test is in detecting departures from the null hypothesis. If a hypothesis test has a high power, the test is referred to as a powerful test [3, p. 168].

Since only correct decisions are of interest, an ideal hypothesis test has a low level of significance α and a high power of test $1 - \beta$ [3, p. 157].

To some extent, these error probabilities can be controlled. Usually, the power of a test depends on the size n of the sample $(X_1, \ldots, X_n)$, because the more information there is about a population, the more likely it is to make a correct inference about the underlying distribution. The power can also be influenced by several other elements, such as the variation in the variable of interest or the significance level. By considering to increase or decrease the sample size, the Type II error can be constrained. [3, p. 168] [2, p. 382-385]

When the size of the observed sample is given, one cannot control both error probabilities simultaneously. In that case, usually only the probability of committing a Type I error is constrained, such that it is fixed at a specified level α. The most common choices are $\alpha \in \{0.01, 0.05, 0.10\}$. When comparing multiple tests in this scenario, the best performing test is the one with the lowest probability of making a Type II error, i.e. the test with the highest power [2, p. 385]. This is the idea behind the power simulation presented in Chapter 7, since the sample sizes in the observed data set mentioned in Chapter 1 cannot be increased. In this thesis, all tests are evaluated with a significance level of $\alpha = 0.05 = 5\%$.

Lehmann and Romano mention that in the past choosing the level α as a seemingly arbitrary value, such as 5%, rather served for convenience to reduce the sizes of the required distribution tables to evaluate (2.2) and to standardize the testing procedure, which later became a habit. However, the level of significance should rather be chosen with consideration of the power of test, since there is a connection between those two. Additionally, they add, there is no point in using a test whose power, i.e., the ability to detect samples for which the null hypothesis is actually not true, is already quite low [13, p. 57].

This or any other rule for choosing a level of significance in connection with the power was not considered for the simulation study in Chapter 7. For some of the tests, the formulas for the test statistic or the p-value are specifically calibrated to a significance level of 5%, such as the robust version of the Jarque-Bera test in Chapter 5. This calibration is also visible in Section 7.4 in Figure 7.8.

When the Type I error is bounded by a significance level of α, the rule for rejecting the null hypothesis for all normality tests in this thesis is:

$$\alpha \geq p \quad \Leftrightarrow \quad \text{reject } H_0 \,.$$

2.2. Point Estimation

When sampling from a distribution described by a probability function $f(x|\theta)$, the distributional parameter θ can determine characteristics of a random sample $(X_1, \ldots, X_n)$, for example for a normal distribution, the expectation μ will determine the location and the variance σ^2 will give a scale of the spread in the sample. Therefore, finding a good estimator for the point θ provides a good foundation for understanding the underlying distribution. In the following, main steps in point estimation from [2] are summarized.

Without specifying a relationship between the estimator $\hat{\theta}$ and the estimated parameter θ or the range of the estimator, Casella and Berger [2] define a point estimator as a function of the sample $(X_1, \ldots, X_n)$:

$$\hat{\theta} = W(X_1, \ldots, X_n).$$

In the case of the normal distribution, a natural choice for the point estimator of the expectation μ is the sample mean. However, not every parameter θ will have such an intuitive explanation. Thus, the following methods, the maximum likelihood method and the method of moments, will give a universal approach to finding a point estimator [2, p. 312].

The maximum likelihood method was preferred over the method of moments, since it is more accessible for distributions that have undefined moments when combined with a numerical optimization. The estimator is used to fit a series of alternative distributions to the real-world data set to be able to draw random samples for the power simulation in Chapter 7.

In the case when the random sample comes from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with two unknown parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, the maximum likelihood estimators and the method of moments estimators coincide, where the estimator for μ is the sample mean and the estimator for σ^2 is the modified sample variance s_1^2. [2]

2.2.1. Maximum Likelihood Estimator

The maximum likelihood method is the most important method in parameter estimation, when fitting a distribution to a sample [4, p. 498]. The approach of the method can be easily illustrated with a discrete distribution. Assume the sample follows a discrete distribution, then one can calculate the probability that the distribution takes these exact values by taking the product of the probability mass function evaluated at each value. The higher the probability determined by the probability mass function with parameter θ of observing the sample values, the more likely the distributional parameter θ appears. Thus, one might consider that the most plausible estimator of θ is the one that maximizes this probability. Similarly, the approach is done for the continuous case. [13, p. 16]

The purpose of the likelihood function $L(\theta|x)$ is to distinguish it from the calculation of probabilities. A distribution specified through its probability function $f(x|\theta)$, as a function of x for a specified distributional parameter θ, describes the probability for the outcome

of the sample values drawn from the distribution. However, statistical inference aims to make statements about the distribution and its parameter θ based on an observed sample. Here, f would need to be considered as a function of the parameter θ and is determined by the sample $(x_1, x_2, \ldots, x_n)$. Instead of reusing the notation of f, which already denotes the probability density function of the population under θ, the term "likelihood" is used to emphasize the change in the point of view where the sample is given and the parameter is analyzed [19, p. 8].

Assume $x = (x_1, \cdots, x_n)$ is a given sample of n independent observations, which is used to fit a continuous distribution with density function $f(x, \theta)$ and unknown d-dimensional distributional parameter $\theta = (\theta_1, \cdots, \theta_d)$. The likelihood function $L(x|\theta)$ of the sample x is given by:

$$L(\theta|x) = L(\theta_1, \ldots, \theta_d|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_d) \tag{2.3}$$

and it only depends on the distributional parameter θ [2, p. 315] [4, p. 499].

The maximum likelihood estimator $\hat{\theta}$ for the sample x is found by maximizing the likelihood function L, i.e.

$$\hat{\theta} = \text{argmax}_{\theta \in \mathbb{R}^d} L(\theta|x_1, \cdots, x_n)\,.$$

Since the logarithm $\ln : \mathbb{R}_+ \to \mathbb{R}, x \to \ln x$ is a strictly increasing function, the locations of the maximum of the likelihood function $L(\theta|x)$ and of the log-likelihood function $\ln L(\theta|x)$ with respect to θ are overlapping.

$$\ln L(\theta|x) = \ln L(\theta|x_1, \cdots, x_n) = \ln\left(\Pi_{i=1}^{n} f(x_i, \theta)\right) = \sum_{i=1}^{n} \ln\left(f(x_i, \theta)\right)$$

Therefore, the maximum likelihood estimator of θ can also be found by maximizing the log-likelihood function, i.e.

$$\hat{\theta} = \text{argmax}_{\theta \in \mathbb{R}^d} L(\theta|x_1, \cdots, x_n) = \text{argmax}_{\theta \in \mathbb{R}^d} \ln L(\theta|x_1, \cdots, x_n)$$

The estimator $\hat{\theta} = (\hat{\theta_1}, \ldots, \hat{\theta_k})$ defined as the value of θ for which the likelihood function attains its maximum, can be derived by setting the partial derivatives of L with respect to all components of θ to zero, forming the following normal equations, and checking the second order condition for continuous distributions.

$$\frac{\partial}{\partial \theta_i} L(\theta|x) = 0, \qquad i = 1, \ldots, k$$

In some literature as is the case for [19], the likelihood function is defined proportional to the above definition (2.3) instead of equal to it, because the actual value of the likelihood is not important, only the comparison is. [19, p. 8, 9]

It can be shown that the maximum likelihood estimator has several asymptotic properties. Most importantly, the maximum likelihood estimator $\hat{\theta}$ is consistent, i.e., it converges in probability against the true parameter value θ_0, that is, $\hat{\theta} \to \theta_0$ as the sample size $n \to \infty$. Furthermore, it converges asymptotically in distribution against a normal distribution, i.e., $\sqrt{n}(\hat{\theta} - \theta_0) \approx \mathcal{N}(0, I(\theta_0)^{-1})$ as $n \to \infty$, where $I(\theta_0)$ denotes the Fisher information [2].

Although the maximum likelihood estimator possesses asymptotic optimum conditions as the size of the sample increases and leads to a satisfactory estimate in many cases, this might not be the case for finite samples. [13, p. 16]

Furthermore, there are drawbacks to this method that are related to finding the maximum of the likelihood function. First, finding the global maximum and verifying it can become a difficult task, also when this is done by a numerical method. The second drawback is the numerical sensitivity of the maximization procedure to changes in the data sample. [2, p. 315]

For the simulation described in Chapter 7, the maximum likelihood method was used to fit alternative distributions. This is done in R. In the R code, the maximum likelihood estimates are found with the optimization function `optim` by minimizing the negative log-likelihood function. Some existing R functions that calculate the maximum likelihood estimates use this function internally. The results can depend on the initial starting value, since `optim` uses iterative methods to find a local minimum. The default method is the Nelson-Mead method, which was used in the R code.

2.2.2. Method of Moments

The estimator found by the method of moments is also a well-known point estimator. For the estimation of $\theta = (\theta_1, \ldots, \theta_d)$ from a random sample derived from the distribution with density function $f(x|\theta)$, the first k sample moments $(m_1, \ldots, m_d)$ and theoretical moments $(\mu_i, \ldots, \mu_d)$ are used in the following set of equations to derive the solution for the point $(\theta_1, \ldots, \theta_d)$:

$$\begin{aligned} m_1 &= \mu_1(\theta_1, \ldots, \theta_d) \\ m_2 &= \mu_2(\theta_1, \ldots, \theta_d) \\ &\vdots \\ m_d &= \mu_d(\theta_1, \ldots, \theta_d), \end{aligned}$$

where $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ is the k-th sample moment and $\mu_k = \mu_k(\theta_1, \ldots, \theta_d) = \mathbb{E}(X^k)$ is the k-th moment, where $X \sim f(x|\theta)$ follows the same distribution as the underlying random sample [2, p. 312].

The method of moments can often lead to estimators that still require improvement, e.g. the estimator for the distributional parameters of a binomial distribution could take negative values. In addition, it is not applicable to every distribution, since the moments up to

order d have to exist. For example, for the Cauchy distribution, which is used in the power simulation, the moments do not exist, so the method of moments is not applicable. [2]

2.3. Skewness and Kurtosis

In tests based on moments, the skewness and kurtosis can be used to detect whether a sample comes from a normal distribution. One way to differ between a normal distribution and a non-normal distribution is to focus on the skewness and kurtosis of a random variable X with expectation $\mathbb{E}(X) = \mu$. These are given by the following expressions in the terms of the central moments, as mentioned in [7, p. 42]:

$$\text{Skewness:} \quad \beta_1 = \frac{\mu_3'}{\mu_2'^{3/2}} = \frac{\mathbb{E}(X-\mu)^3}{\{\mathbb{E}(X-\mu)^2\}^{3/2}}$$
$$\text{Kurtosis:} \quad \beta_2 = \frac{\mu_4'}{\mu_2'^{2}} = \frac{\mathbb{E}(X-\mu)^4}{\{\mathbb{E}(X-\mu)^2\}^{2}},$$

where $\mathbb{E}(X) = \mu$ is the expectation of X and $\mu_k' = \mathbb{E}(X-\mu)^k$ is the k-th central moment of X. For the realization $(x_1, x_2, \ldots, x_m)$ of a random sample $(X_1, \cdots, X_n)$ the k-th central sample moment is estimated as

$$m_k' = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

with the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The estimators for the central sample moments can be used to estimate skewness and kurtosis of the sample [5, p. 20].

Furthermore, to give a more comprehensive overview of the results of the power simulation results in Chapter 7, the sample estimates for skewness and kurtosis of the generated samples are added to the graphs.

Based on [12], estimators G_1 and G_2 that contain an adjustment for the sample size are displayed, instead of the traditional estimators β_1 and β_2. These adjusted estimators have a smaller mean-squared error for small samples from non-normal distributions and are also implemented in other programs, such as SAS or Microsoft Excel.

$$\text{Skewness:} \quad G_1 = \frac{K_3}{K_2^{3/2}} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{s_1^3}$$
$$\text{Kurtosis:} \quad G_2 = \frac{K_4}{K_2^{2}} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{s_1^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

The estimators G_1 and G_2 are derived based on the traditional estimators g_1 and g_2 of β_1 resp. β_2,

$$\text{Skewness:} \quad g_1 = \frac{m_3'}{{m_2'}^{3/2}}$$
$$\text{Kurtosis:} \quad g_2 = \frac{m_4'}{{m_2'}^2} - 3\,.$$

where $m_k' = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = m_2'$, but replacing the m_r' by K_r defined as:

$$K_2 = \frac{n}{n-1} m_2'$$
$$K_3 = \frac{n^2}{(n-1)(n-2)} m_3'$$
$$K_4 = \frac{n^2}{(n-1)(n-2)(n-3)}\Big((n+1)m_4' - 3(n-1){m_2'}^2\Big)$$

For further details, see [12].

3. Graphical Approach

The importance of the wide range of diagnostic plots, recommended for various statistical methods, lies in the sense of pattern and details that cannot be provided otherwise. Although statistical tests might give a more objective judgment for the suitability of the null hypothesis for the observed sample, they do not give detailed insight into the reason for the results of the tests, nor do they offer a sense for statistical effects, such as clusters or outliers, that may influence the results [20, p. 15].

The empirical distribution function (EDF) of an observed sample $(x_1, x_2, \ldots, x_n)$ is defined as a step function $F_n(t)$:

$$F_n(t) = \begin{cases} 0 & t < x_{(i)} \\ i/n & x_{(i)} \leq t < x_{(i+1)} \text{ for } i = 1, \cdots, n-1 \\ 1 & x_{(n)} \leq t\,, \end{cases} \tag{3.1}$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ is the ordered sample. This definition reflects the proportion of observed values less than or equal to t. For a sample of size n, the steps increase by $1/n$ for each observation [20, p. 100].

$F_n(x)$ is an unbiased estimator for the cumulative distribution function $F(x)$. Considered on its own, the empirical distribution function is an elementary statistics, since the entire sample can be recovered from it. Therefore, it has an important role in statistical inference, such as goodness-of-fit testing. However, using it on its own will not give much insight into the underlying distribution [18, p. 56].

A disadvantage of methods based on the empirical cumulative function is their sensitivity to random occurrences in the samples. Thus, relying only on these could influence the final decision-making. This becomes especially important for smaller sample sizes [7, p. 9].

Figure 3.1 displays the comparison of the EDF with normal distribution density functions, where expectation μ and variance σ^2 are estimated by the sample mean and sample variance resp. for each observed sample in the real-world data set, mentioned in Chapter 1. For some of the plots, the normality assumption seems more suitable. However, the sample sizes range between 5 and 10. This small amount of observations might introduce a bias into the visual representation and therefore into the interpretation of the sample's real distribution.

Since in general, cumulative distribution functions of continuous random variables are curved, it is not easy to visually judge a normality assumption based on a comparison

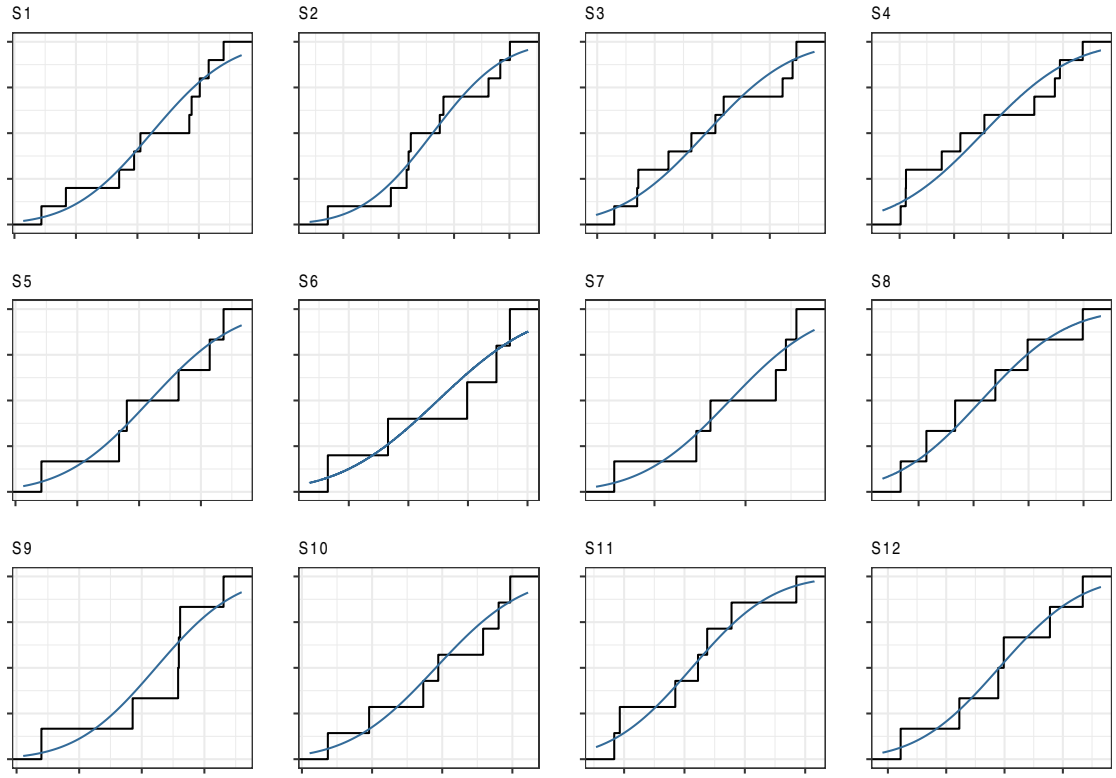

Figure 3.1.: Comparison of the empirical distribution function and the density of a normal distribution with sample mean and sample standard deviation for each data sample $S1$ - $S12$. The values on the x-axis are the observed data points and the values on the y-axis are the estimated probabilities, as defined in (3.1).

with the sample's empirical distribution function in the form of a step function. Therefore, the following plots provide methods based on the raw data or on the deviation from straight lines for an easier decision making.

Furthermore, this section presents basic plotting procedures for observed data samples. Since the plots focus on evaluating whether a sample follows a normal distribution, any distribution function that is used in the construction of the plots refers to the normal distribution. Because of the small sample sizes, they might not always provide a meaningful judgment and might even be misleading in cases where the sample size is not depicted in the plot, as is the case for a plain boxplot or a histogram.

3.1. Raw Data Plots

Raw data plots give a feeling for shape and symmetry of the data sample and other characteristics, such as clusters and outliers. They do not involve any complex computation and could also be plotted by hand [20, p. 17].

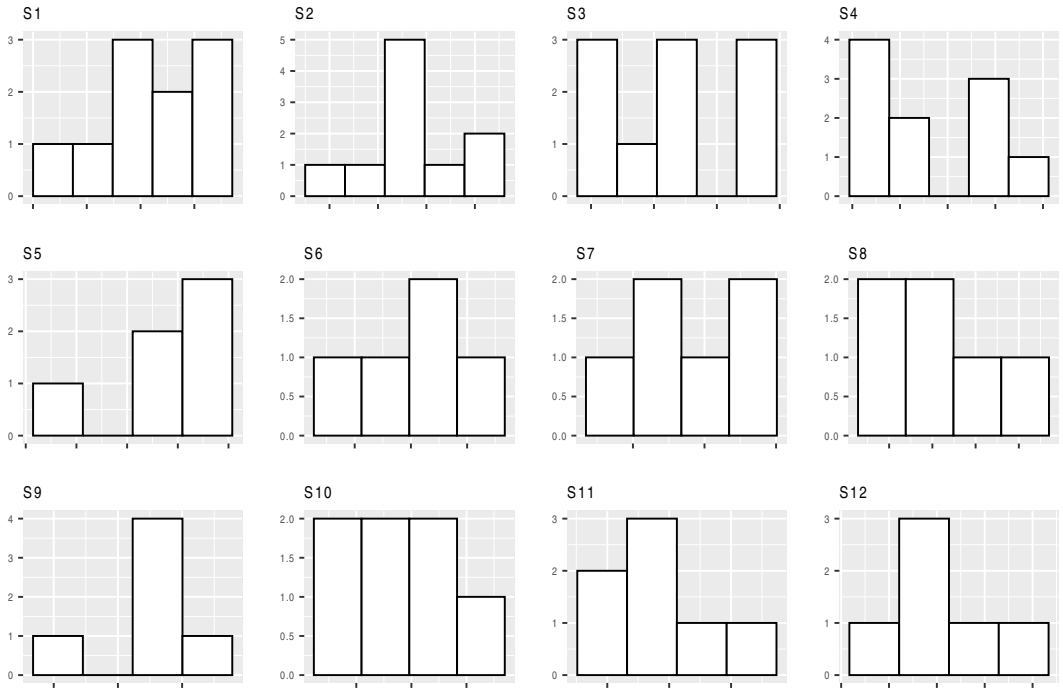

Figure 3.2.: Histograms of samples S_i of sizes $n_i \in \{10, 10, 10, 10, 6, 5, 6, 6, 6, 6, 7, 7, 6\}$ with number of bins set to $\lceil \sqrt{n_i} \rceil$ for $i \in \{1, \ldots, 12\}$.

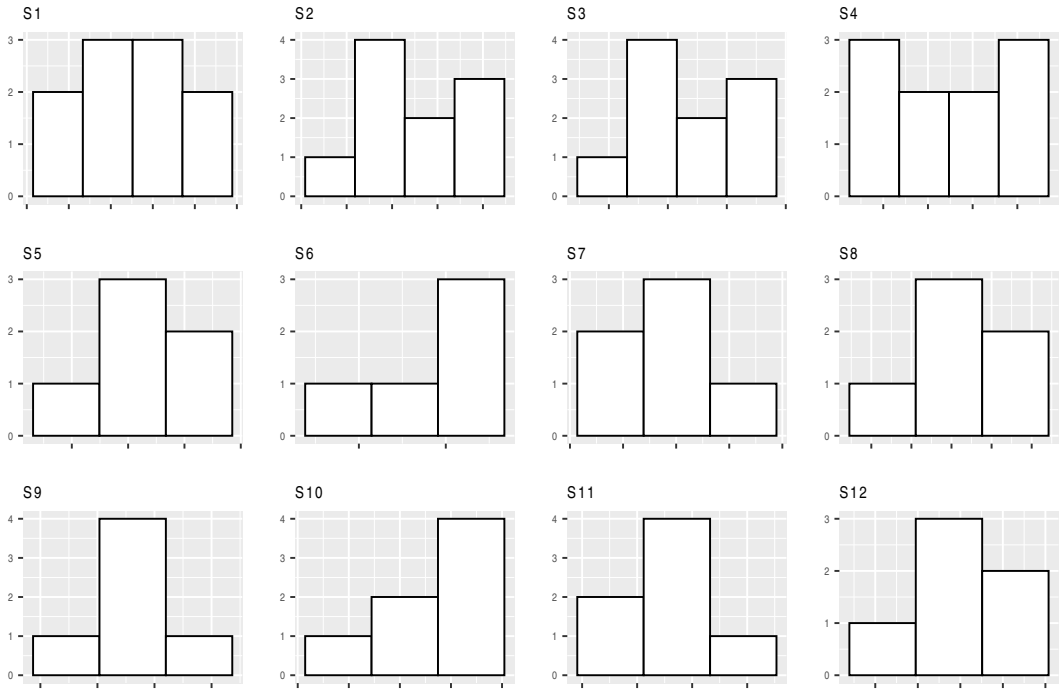

Figure 3.3.: Histograms of samples S_i of sizes n_i where the number of bins is determined by " Sturges " as $\lceil \log_2(n_i) + 1 \rceil$

3.1.1. Histograms

A histogram displays the absolute or relative frequencies of observations in pre-specified bins and, therefore, can be used to approximate the sample distribution. Distributional characteristics, which can be visible in a histogram, include symmetry, spread of data points, or outliers [20, p. 17].

A histogram of a normally distributed sample is expected to show a symmetric bell shape approximating the density function. Figures 3.2 - 3.3 show histograms of the observed samples with two different methods to determine the number of bins used. In Figure 3.2, the number of bins for each histogram is set to $\lceil\sqrt{n_i}\rceil$, where n_i is the number of observations in each sample $i \in \{1, \cdots, 12\}$. For Figure 3.3, the number of bins is determined by the rule of Sturges as $\lceil\log_2(n_i)+1\rceil$, resulting in a different number of bins for the observed samples.

Especially with such small sample sizes, the number of bins displayed in the plots can greatly influence the form of the histograms. Increasing the number of bins decreases the number of observations per bin and can lead to bins with just one observation for such small samples, making the usage of a histogram inappropriate for such small samples.

3.1.2. Boxplots

To understand the construction of boxplots, the concept of quantiles are defined first. Assume $X \sim F$ is a continuous real-valued random variable, following the distribution of the cumulative distribution function F and $(x_1, x_2, \ldots, x_n)$ is an observed sample that follows the same distribution as X. For $p \in (0, 1)$ the p-quantile is a value x_p that satisfies

$$F(x_p) = \mathbb{P}(X \leq x_p) \geq p \quad \text{and} \quad \mathbb{P}(X \geq x_p) \geq 1 - p\,,$$

that is, x_p is a value in the range of the sample that is greater than at least $100\,p\%$ of the sample values and smaller then at least $100(1-p)\%$ of the sample values.

For $p \in \{0.25, 0.5, 0.75\}$ the corresponding quantiles, also commonly named 25%-, 50%- and 75%-quantiles, are also called first quartile ($p = 0.25$), second quartile or median ($p = 0.5$) and third quartile ($p = 0.75$).

Quantiles also play a role in probabilities plots, particularly Q-Q plots, presented in the next sections.

A boxplot shows less detail compared to a histogram, but provides visible indication about the spread and symmetry of the sample. The box is bounded by the first quartile (25%-quantile) and the third quartile (75%-quantile) and surrounds the median (50%-quantile) as a location parameter of the sample.

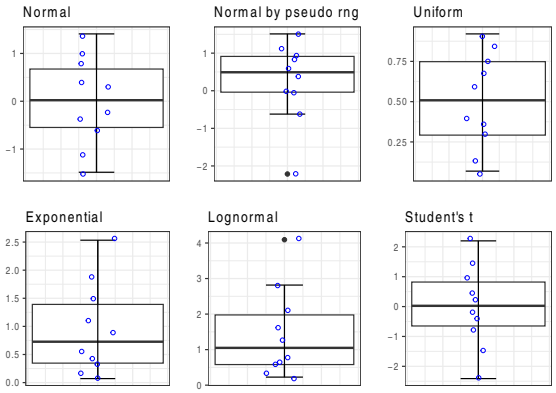

Figure 3.4.: Distinct patterns in boxplots of alternative distributions, sample size $n = 10$.

Outside the box, the whiskers usually extend to a maximum of 1.5 times the interquartile range, the difference between the third and the first quartiles, away from the lower and the upper quartile, or the maximum and minimum value of the sample if these are closer. Data points outside the whiskers are considered outliers. Symmetry in the sample is indicated by an even spread between the depicted quartiles and the whisker ends [20, p. 18].

Figure 3.4 shows idealized and expected patterns of the normal distribution and alternative distributions.

All samples were generated by the quasi-random number generator `randomLHS` in R, except for the middle plot in the upper row. This is a pseudo-random sample generated by the R function `runif`. For the example of a uniformly distributed sample, a quasi-random generator provides a more idealized sample, with random numbers more equally spread across the range of the specified distribution, whereas a pseudo-random generator will produce more naturally looking random numbers, that might even appear to be spread more sparsely across the range of the distribution.

In some literature, as it is the case with [7] [20], example plots for the display of difference between certain distributions tend to use quasi-random numbers, since they are better for identifying distributional characteristics. However, pseudo-random numbers inherently contain some random variation, making it easier to compare with real-world data instead of depicting ideal and highly controlled random samples. Therefore, the plots in the following subsection will always contain alternatively distributed samples generated with quasi-random numbers to depict which distinct patterns, and a normal sample generated with pseudo-random numbers to show that some kind of variation in the data can still occur.

As visible in the plots of Figure 3.4, for a sample coming from a symmetric distribution, the boxplot is expected to be symmetric around the median. For asymmetric distributions, such as the exponential and log-normal distributions, the boxplots are not symmetric but show a shorter and a longer whisker, which can have outliers. Furthermore, compared to the width of the box, lighter tailed distributions show shorter whiskers, such as the uniform distribution, whereas heavier tailed distributions show longer whiskers, such as the Student's t-distribution. Therefore, a heavy-tailed distribution might also show outliers.

For the observed data samples considered in this thesis, most of the boxplots in Figure 3.5 are not symmetric and do not show a proportion of the box to the whisker length expected for a normal distribution. However, the observed samples only have 5 to 10 values.

The boxplots in Figure 3.5 additionally display the observed data. If that were not the case, the plots could easily give a misleading representation of the underlying distribution when the observer does not assume or expect small samples. Thus, boxplots for such small samples might not give a reasonable representation or judgment about the symmetry of the data.

While boxplots can be useful for identifying potential outliers, they are not that well suited for assessing normality in small datasets, since the whole sample is summarized by a set of five statistics, i.e., the minimum, the 25%-quantile, the median, the 75%-quantile and the maximum.

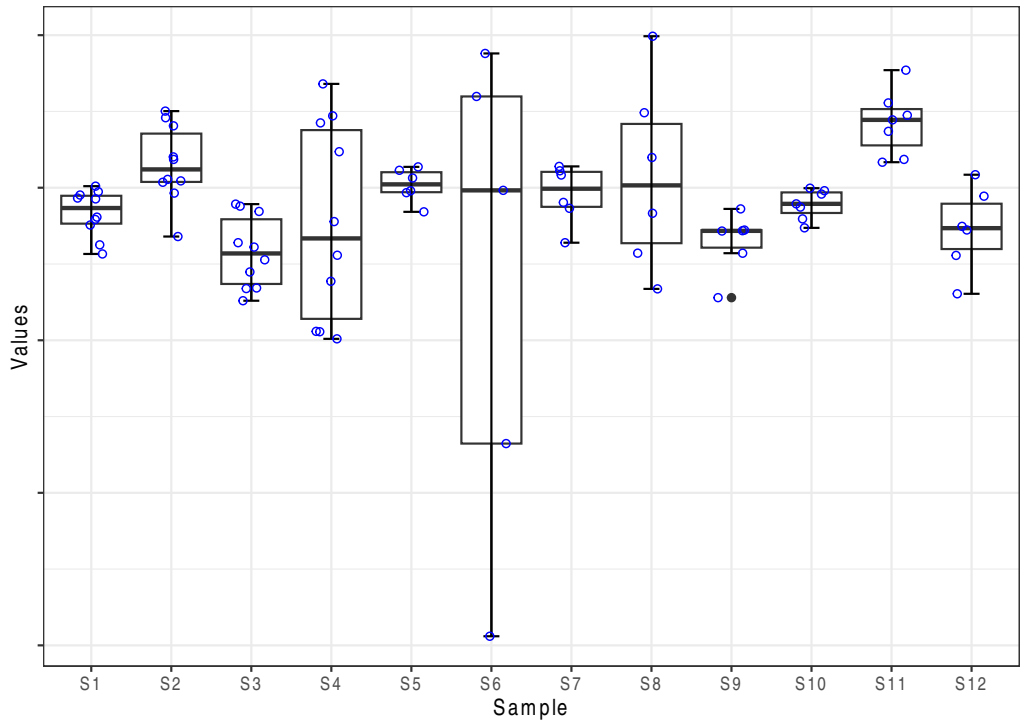

Figure 3.5.: Boxplots of the real-world samples. Outliers are marked with dots, when they lay further than 1.5 times the interquartile range away from the box. The observed values are displayed with circles.

3.2. Methods of Probability Plotting

Probability plots give a graphical method for comparing observed values of order statistics of the sample with their theoretical values of the null hypothesis. The i-th order statistic corresponds to the i-th smallest observation in a random sample: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(i)} \leq \cdots \leq x_{(n)}$. As part of probability plots, a normal Q-Q plot (quantile-quantile) or a P-P plot (percent-percent) displays an approximately linear relation to indicate how well the null hypothesis fits the sample. Small deviation from the linear relation is typically caused by randomness, whereas a distinct pattern in the plots could indicate that the sample follows an alternative distribution [7, p. 16] [20, p. 26].

Often a probability plot is accompanied by a straight line to determine the fit of the null hypothesis. There are different procedures for its construction. The simplest, as used in most Q-Q plots, is to find an upper quantile and a lower quantile of the x- and y-coordinates and connect the points. The quantiles could be 10%-90%, 2.5%-97.5% or 25%-75%, which is the most popular. The deviation is then evaluated with respect to the vertical distance between the points and the line [7, p. 29].

Other procedures for the line result from a transformation. Assume the underlying sample follows a distribution with location and scale parameters, μ resp. σ. These do not have to be the expectation (mean) and standard deviation. With the transformation $z = \frac{x-\mu}{\sigma}$, the distribution function $F(x)$ of the sample can be rewritten as $F(x) = G\left(\frac{x-\mu}{\sigma}\right) = G(z)$. This transformed sample z is said to come from the standardized random variable Z, i.e. the mean of Z is zero and its variance is one, that follows the distribution function $G(z)$ [7, p. 25].

Using this method, the straight line in probability plots is created by plotting z on x with the relation:

$$z = G^{-1}(G(z)) = G^{-1}(F(x)) = \frac{x-\mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}x$$

or

$$x = \mu + \sigma z \tag{3.2}$$

For application, the distribution function F of the sample is estimated by the empirical distribution function F_n, as defined in (3.1) [7, p. 25].

For the estimation of the unknown coefficients μ and σ in (3.2), the least-square linear regression method could be used to obtain the estimates $\hat{\mu}$ and $\hat{\sigma}$. Their regression estimates are then given by:

$$\hat{\sigma} = \frac{\sum_i (z_i - \bar{z})x_i}{\sum_i (z_i - \bar{z})^2} \quad \text{and} \quad \hat{\mu} = \bar{x} - \hat{\sigma}\bar{z}\,. \tag{3.3}$$

Other estimation techniques lead to regression tests, as presented in Chapter 4 [7, p. 29].

For small sample sizes, probability plots for z against x can show curvature in the tails, even if the null hypothesis is actually true. Therefore, it is recommended to plot the expectations of the i-th order statistics $\mathbb{E}(Z_{(i)}), i = 1, \cdots, n$ of the standardized values of the null hypothesis instead of the transformed random variable Z for plotting positions [7, p. 31, 58] as it is described in the following section.

3.2.1. Plotting Positions

Generally, in probability plots the observed values of the order statistics $x_{(1)} \leq \cdots \leq x_{(n)}$ on the y-axis are plotted against their theoretical or expected values $z_i = G^{-1}(F_n(x_{(i)}))$ on the x-axis. These theoretical abscissa are based on their cumulative distribution function corresponding to the distributional assumption of the null hypothesis and are called plotting positions. Here, they are denoted by $p_i, i = 1, \cdots, n$. To some extent, the plots can be affected by the plotting points [20, p. 18, 19] [7, p. 25].

The mathematician Gunnar Blom proposed a general estimate for the plotting positions independent of the underlying distribution:

$$F_n(x_{(i)}) = p_i := \frac{i - \alpha_i}{n - 2\alpha_i + 1}$$

with different constants $0 \leq \alpha_i \leq 1$ for every sample size n and the null distribution function G. For a normality assumption under the null hypothesis ($G(x) = \Phi(x) := \int_{-\infty}^{x} \frac{e^{-y^2/2}}{2\pi}\, dy$) the estimates for the mean values of the order statistics are derived with the inverse cumulative distribution function of the standard normal distribution:

$$z_i = \mathbb{E}(Z_{(i)}) = \Phi^{-1}(p_i)$$

As Blom only observed small variation for various null distributions, he suggested letting $\alpha_i, i = 1, \cdots, n$ be constant with $\alpha = 0.5$ or $\alpha = 0.375$, which then leads to the plotting positions:

$$p_i = \frac{i - 0.5}{n}, \qquad \text{for } \alpha = 0.5 \tag{3.4}$$

$$p_i = \frac{i - 0.375}{n + 1/4}, \qquad \text{for } \alpha = 0.375 \tag{3.5}$$

Another common set of plotting positions is constructed with $\alpha = 0$ and results in

$$p_i = \frac{i}{n + 1} \tag{3.6}$$

as an estimate for $p_i = \mathbb{E}(\Phi^{-1}(Z_{(i)})) = \mathbb{E}(U_{(i)})$, where $U_{(i)}, i = 1, \cdots, n$ denotes the order statistics of the standard uniform distribution $U(0, 1)$ and $Z_{(i)}, i = 1, \cdots, n$ the order statistics for the standard normal distribution $\mathcal{N}(0, 1)$ [20, p. 19, 20].

3.2.2. Quantile-Quantile Plots (Q-Q)

A quantile-quantile plot (Q-Q) usually compares the quantiles of two different distributions. However, if it compares the empirical quantiles of the sample with the hypothesized quantiles of the null distribution, the Q-Q plot follows the theory of a probability plot as described in Section 3.2 [7, p. 58].

Q-Q plots visualize the observed values of the order statistics from the sample $x_{(i)}, i = 1, \cdots, n$, compared to their expected values $\mathbb{E}(Z_{(i)})$ under the null hypothesis with distribution function $G(x) = \Phi(x) := \int_{-\infty}^{x} \frac{e^{-y^2/2}}{2\pi}\, dy$ estimated by:

$$z_i = G^{-1}(F_n(x_{(i)})) \overset{H_0}{=} \Phi^{-1}(p_i)$$

In case of testing a normality assumption, the plots visualize the pairs $\left(\Phi^{-1}(p_i), x_{(i)}\right)$ [20, p. 21].

Unlike regression tests, presented in Chapter 4, which depend on the comparison of the empirical distribution function and the assumed null distribution, Q-Q plots are only evaluated visually, and therefore the choice of plotting positions is not as crucial. The most popular plotting positions are given by (3.4) and (3.5) [20, p. 21].

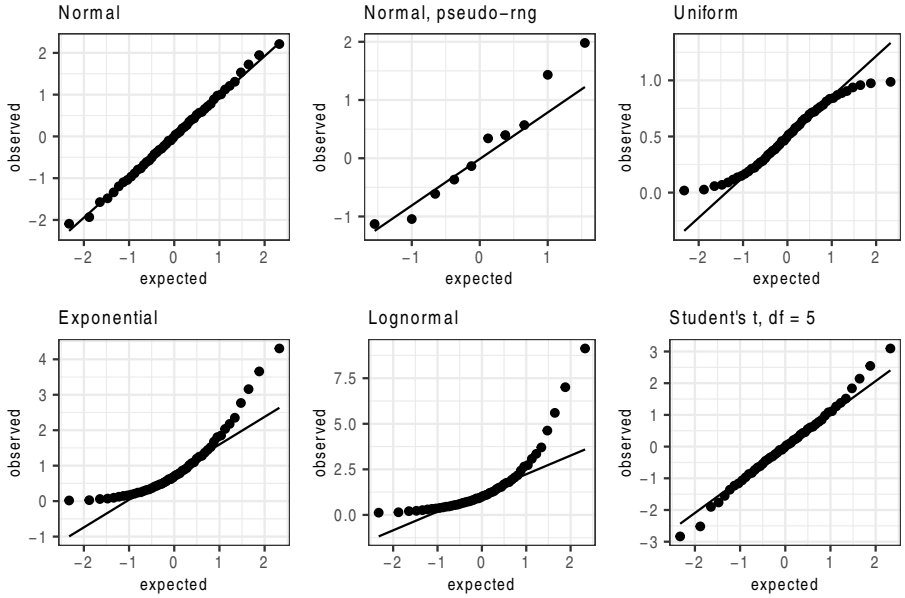

Figure 3.6.: Q-Q plots of different distributions for the comparison of quantiles with a normal distribution. All of them, except for the middle in the upper row, are generated with a quasi-random number generator (rng). The one exception was generated with the standard random generator `rnorm` from R for pseudo-random numbers.

Figure 3.6 visualizes the behavior of different distributions in Q-Q plots that test for a normality assumption. Furthermore, one can see the difference between a normal sample generated by a quasi-random number generator and a pseudo-random number generator. While the normally distributed sample in the upper left corner follows the line almost perfectly, the plot next to it with a pseudo-random sample looks rougher with more random variation. The patterns in the other four plots result from the alternative distributions the samples are derived from. The standard uniform distribution has light tails since the generated values cannot be outside the interval (0,1). This produces the S-shape of the plotted points. The Student's t-distribution with 5 degrees of freedom has heavier tails than the normal distribution. This is visible in the deviation in the lower and upper ranges of the interval. The exponential and log-normal distributions are right-skewed, making the right tail very light and therefore creating the curves in the plots.

The straight line in this figure is created by connecting the 25%- and 75%-quantile of the sample data and the standard normal distribution.

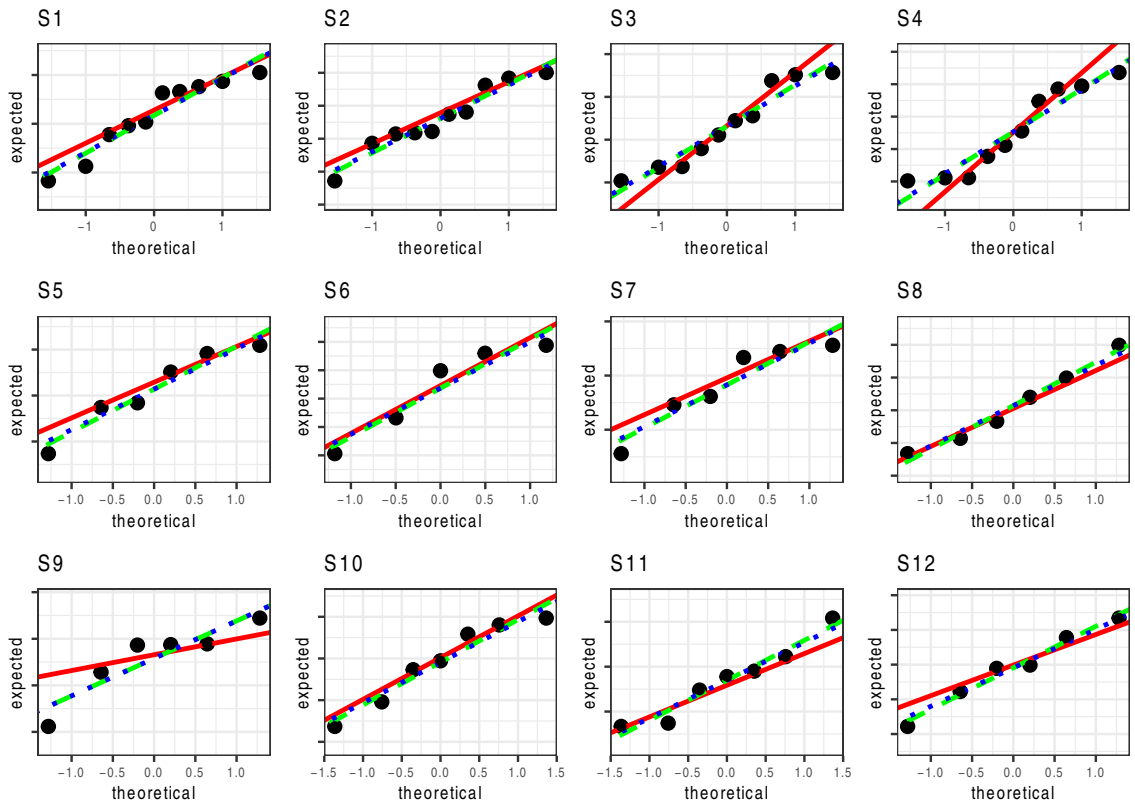

Figure 3.7.: Q-Q plots for the observed data variable with three different lines.

As described in [7, p. 29], there are multiple ways to construct the line in a Q-Q plot: (a) with the quantiles, shown in red in Figure 3.7, (b) with linear regression of the model

$x_{(i)} = \hat{\mu} + \hat{\sigma} z_{(i)}$, where the $z_{(i)}, i = 1, \cdots, n$ are estimated by $\mathbb{E}(Z_{(i)}) = \Phi^{-1}(p_i)$ in green, and (c) by sample estimates $\hat{\mu}$ and $\hat{\sigma}$ in blue.

In Figure 3.7 one can see the difference in the lines for simplifying the judgment of following a linear relation. The green and blue dashed lines, using estimates for

$$x = \hat{\mu} + \hat{\sigma} z$$

as described in (3.3), are almost identical and deviate from the red line, which is constructed by quantiles. Especially for small samples, this can influence how much and in what pattern the points deviate from it.

As mentioned in [7], for small sample sizes, probability plots can show curvature in the tails, even though the null hypothesis is actually true, possibly leading to an unclear representation of the behavior of the data at the tails. Considering the small samples sizes, the Q-Q plots of the observed variables do not seem to display extreme patterns. In addition, since these variables contain data from practice, some randomness is expected, that is, the data may not show a perfect linear relation.

3.2.3. Percent-Percent Plots (P-P)

A percent-percent plot (P-P) compares the percentages of two different distributions. Although its usefulness is limited, a P-P plot is good at detecting deviations around the median of a distribution [7, p. 58].

In general, P-P plots are recommended for a comparison of two samples instead of a sample and a null distribution. Despite this, there is a standardized version of a P-P plot for this purpose, similar to a Q-Q plot [20, p. 23].

P-P plots are based on the relation $F(x) = G(z)$ with the transformation

$$z = (x - \mu)/\sigma$$

that was described in (3.2). Under the null hypothesis, the transformed sample z follows the standard normal distribution ($G = \Phi$), when testing for normality with appropriate estimates $\hat{\mu}$ and $\hat{\sigma}$ for the unknown parameters. Furthermore, when a random variable Z is distributed with cumulative distribution function G, the transformation $G(Z) \sim U(0, 1)$ follows a standard uniform random variable. This is known as the Probability Integral Transform (PIT).

In addition, if the expectations of the order statistics of a sample are estimated by the sorted sample itself, i.e by $x_{(1)} \leq \cdots \leq x_{(n)}$, then the values

$$w_i := \Phi\Big(\frac{x_{(i)} - \mu}{\sigma}\Big)$$

estimate the mean value of the order statistics of a uniform distribution resp. their percentage points, since for the cumulative distribution function of a uniform distribution it

holds: $F_{U(a,b)}(u) = u$ for $a \leq u < b$. Thus, for visualizing the percentage points (p_i, w_i), the accompanying line always starts at $(0, 0)$ and ends at (1,1) [20, p. 24].

In the plots below, the plotting positions $p_i = i/(n+1)$ as in (3.6) are used, since these estimate the expected values of the order statistics of a standard uniform distribution $U(0, 1)$ [20, p. 20].

Smaller deviation from the 45° line is explained by randomness under the null hypothesis, whereas specific patterns indicate an alternative distribution [20, p. 24].

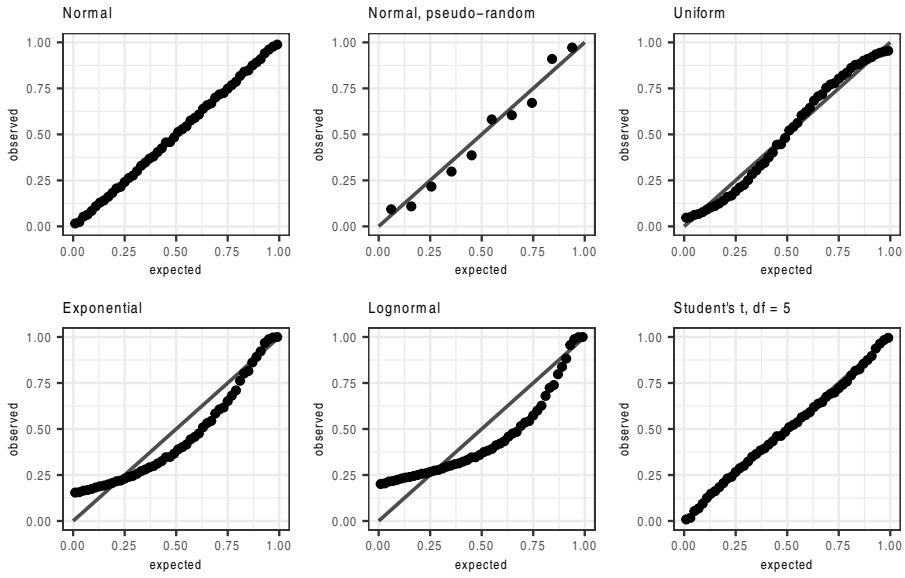

Figure 3.8.: P-P plots of different distributions for the comparison of percentage points with a normal distribution. All of them, except for the middle in the upper row are generated with a sobol random number generator (rng). The one exception was generated with the standard random generator `rnorm` from R.

In Figure 3.8 some example patterns of alternative distributions are displayed, including an example with two normally distributed samples, one generated as quasi-random numbers, the other by the R function `rnorm` as pseudo-random numbers. Similarly to Figure 3.6, the normal sample, generated as quasi-random numbers in the upper left corner, shows an almost perfect linear relation, whereas the normal sample with pseudo-random numbers looks a bit more scattered, but still follows a linear relation.

With a P-P plot, asymmetric distributions seem to be better identifiable than symmetric alternative distributions with lighter or heavier tails.

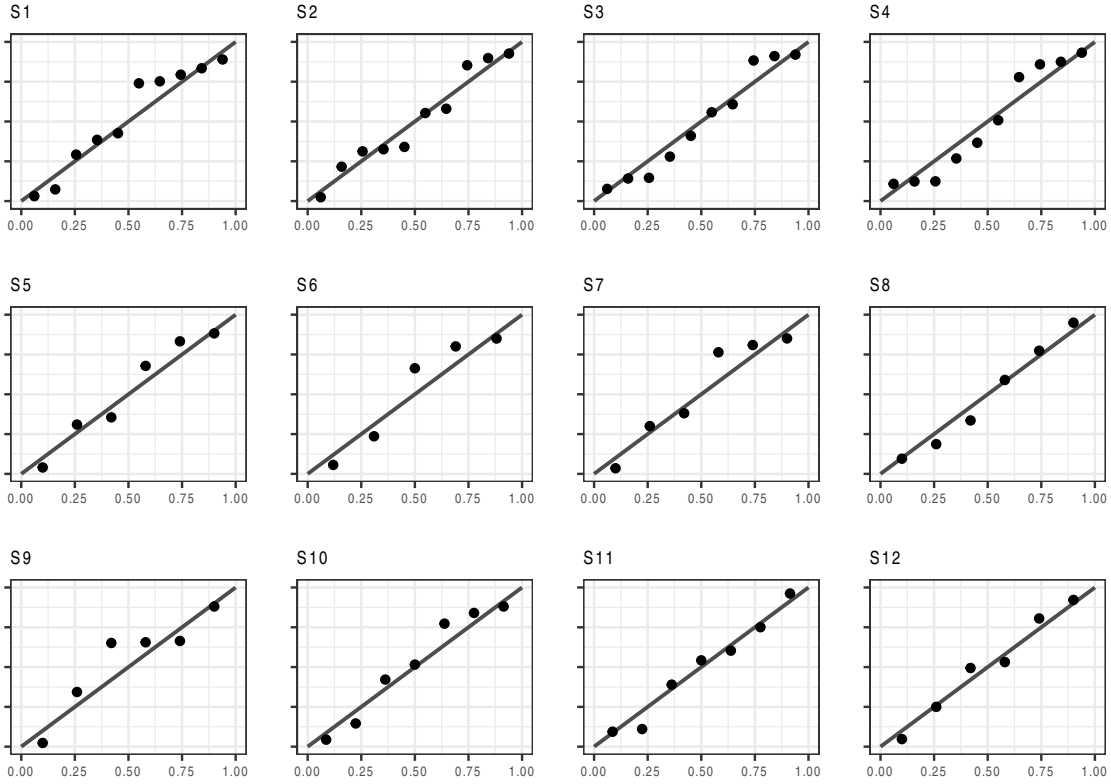

Figure 3.9.: P-P plots for the observed variables.

Figure 3.9 displays the P-P plot for the real-world data set considered in this thesis. Looking back, the boxplots in Figure 3.5 suggest that sample 9 might have an outlier. However, this is not as intuitively represented in the P-P plots in Figure 3.9, but rather in the Q-Q plots.

Due to the limited usefulness, as described in [7], P-P plots may be less user-friendly and less intuitively understandable than Q-Q plots.

3.2.4. Detrended Probability Plots

Since the discrepancy between the sample distribution and the null hypothesis in Q-Q plots is measured by the vertical distance of the points to the line, the slope of the line could distort the visual representation of these deviations. Therefore, the idea of detrended Q-Q plots is to plot these vertical distances on the y-axis instead of the observed values.

Henry Thode explicitly described that detrended Q-Q plots are plotting

$$x_{(i)} - \hat{\sigma}\Phi^{-1}(p_i)$$

against the expected value of the quantiles $\Phi^{-1}(p_i)$ or against the plotting positions p_i, with an estimation $\hat{\sigma}$ for the standard deviation. Under the null hypothesis of normality, the plot is supposed to show a straight relation along the x-axis [20, p. 25].

As described in [7, p. 29], there are multiple ways to construct the line in a Q-Q plot: (a) with the quantiles, shown in red in the Figure 3.10, (b) with linear regression of the model $x_{(i)} = \hat{\mu} + \hat{\sigma}z_{(i)}$, where the $z_{(i)}, i = 1, \cdots, n$ are estimated by $\mathbb{E}(Z_{(i)}) = \Phi^{-1}(p_i)$ in green, and (c) by sample estimates $\hat{\mu}$ and $\hat{\sigma}$ in blue.

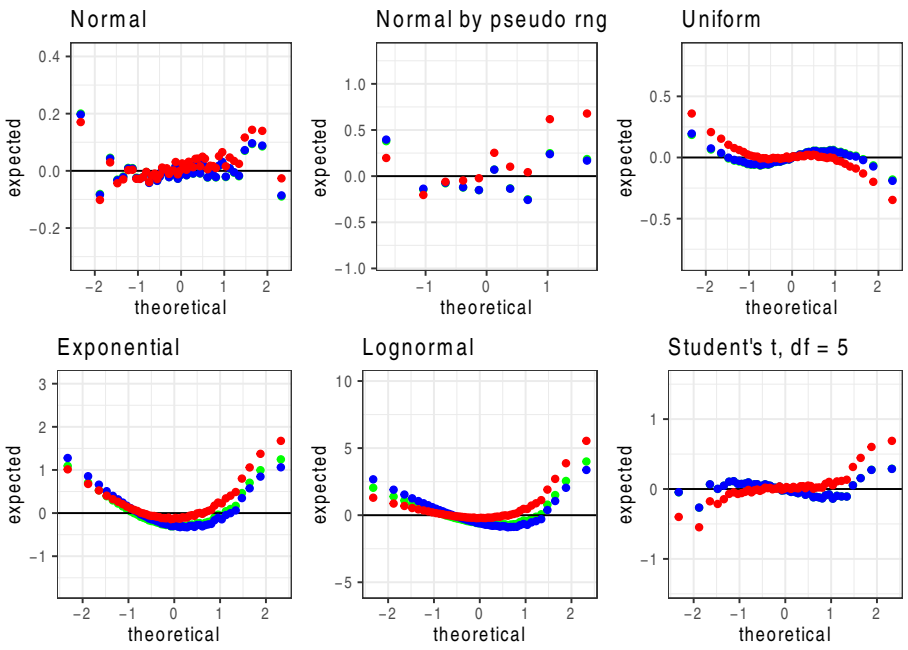

Figure 3.10.: Behavior of the normal distribution and of alternative distributions in detrended Q-Q plots.

Figure 3.10 shows expected behavior in detrended Q-Q plots for the normal distribution and some alternative distributions. While the two normal samples, one consisting of quasi-random numbers and the other of pseudo-random numbers, seem to be scattered around

$y = 0$, the alternative distributions show distinct patterns. The right-skewed samples, derived from an exponential and a log-normal distribution, show curves resulting from their lighter tail on the right and their heavier tail at the left end of the range. The uniform and the Student's t distributions each show different S-shapes, caused by their lighter resp. heavier tails compared to a normal distribution.

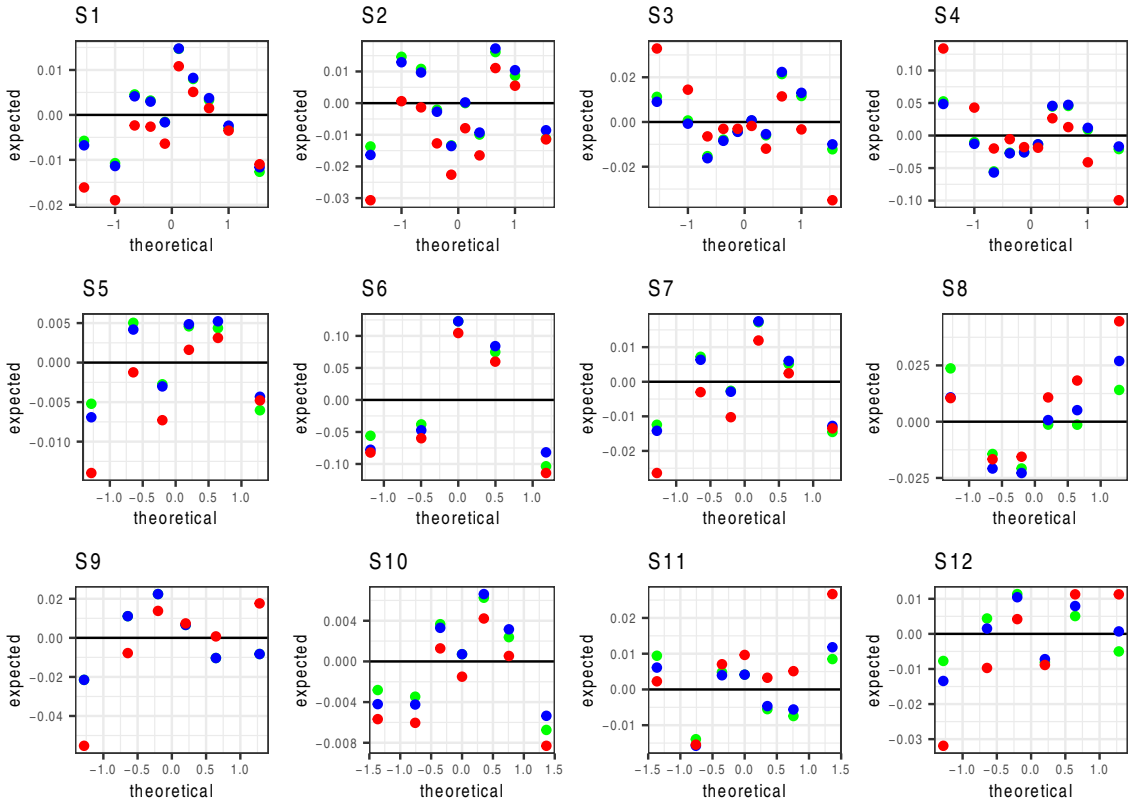

Figure 3.11.: Detrended Q-Q plots for the observed samples.

Figure 3.11 displays the real-world data set in detrended Q-Q plots. With such small sample sizes, it cannot be clearly identified if the deviation depicted in the detrended Q-Q plots is caused by randomness or by distinct patterns, as is the case for alternative distributions in Figure 3.10. However, the ranges of the distances plotted on the y-axis are much smaller than those in Figure 3.10.

4. A Regression Test, the Shapiro-Wilk Test

Although graphical methods, as described in Chapter 3, are valuable in identifying characteristics of the sample distribution, relying only on them can lead to an incorrect conclusion due to their sensitivity to random occurrences in the observed samples. Therefore, more objective methods are necessary to verify the distributional assumptions [20, p.26].

As an extension of probability plots, regression tests measure the linear relation displayed in the plots. However, in contrast to a subjective interpretation of probability plots, these tests provide a more objective procedure for assessing the suitability of the null hypothesis [20, p. 16].

Regression tests arise from a linear regression model, in which the expected values of the order statistics $\mathbb{E}(Z_{(i)}), i = 1, \cdots, n$ under the null hypothesis are explaining their observed values from the sample $X_{(1)} \leq \cdots \leq X_{(n)}$

$$X_{(i)} = \mu + \sigma \mathbb{E}(Z_{(i)}) + \epsilon_i$$

with the error term ϵ_i [7, p. 393].

As described in Section 3.2 about probability plots, the sorted sample $X_{(1)} \leq \cdots \leq X_{(i)} \leq \cdots \leq X_{(n)}$ serves as a set of empirical order statistics and is plotted against a specified function of i, such as transformations of the plotting points 3.4 and 3.5 suggested by Blom. The deviation of the data points from a fitted straight line can give indication of the fit of the null hypothesis for the underlying sample. Considering this graphical method, tests based on this approach can be constructed as well. One of these tests, the Shapiro-Wilk test, is presented in this chapter [7, p. 195].

An approach for measuring the fit of this model under the null hypothesis, which is used by the Shapiro-Wilk test, is based on estimating the scale parameter σ and comparing its squared value with another estimate, such as the sample variance of the sample [7, p. 197].

Shapiro-Wilk Test

Assume under H_0, that $(x_1, \cdots, x_n)$ is a realization of a random sample $(X_1, X_2, \ldots, X_n)$ that comes from a normal distribution $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. for $i \in \{1, \cdots, n\}$ and similarly $(z_1, \cdots, z_n)$ is a realization of a random sample $(Z_1, Z_2, \ldots, Z_n)$ following a standard normal distribution. In the following, the expectation of the i-th order statistic is denoted by $m_i := \mathbb{E}(Z_{(i)})$ with the vector $m' = (m_1, \cdots, m_n)$, where m' denotes the transpose of the $\mathbb{R}^d$-valued vector m, and the variance-covariance matrix of the order statistics denoted by V [20, p.27].

According to the Gauss Markov Theorem, the best linear unbiased estimator for the unknown coefficient σ in the linear model

$$x_{(i)} = \mu + \sigma\, m_i + \epsilon_i, \quad i = 1, \cdots, n$$

is given by the least square solution

$$b := \hat{\sigma} = a'x = \sum_{i=1}^{n} a_i\, x_{(i)} \tag{4.1}$$

with exact coefficients

$$a = \frac{m'V^{-1}}{(m'V^{-1}V^{-1}m)^{1/2}} \in \mathbb{R}^n \quad \text{fulfilling} \quad a'a = 1.$$

The vector a has symmetric entries, i.e $a_i = -a_{n-i+1}$, resulting in

$$b = \sum_{i=1}^{n/2} a_{n-i+1}(x_{(n-i+1)} - x_{(i)}) \tag{4.2}$$

The test statistic W of the Shapiro-Wilk test is based on the relation between the estimated variance b^2 and the sample variance $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$. It is given by

$$W = \frac{b^2}{(n-1)s^2} \tag{4.3}$$

[20, p. 27] [7, p. 393].

The coefficients a are explicitly given for sample sizes $n < 20$. However, since for each n the test statistic W depends on a different set of a-values, the computation becomes more cumbersome the larger the sample, therefore estimates for the coefficients were developed [20, p. 27].

As a suggestion in connection with the test by Shaprio and Francia in [20, p. 29,30] it was mentioned to use the plotting points by Blom for approximations of the order statistics

$$m_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right). \tag{4.4}$$

Patrick Royston recommended using a transformation of the test statistic W that is approximately standard normally distributed. Furthermore, he presented approximations for the coefficients a and estimated parameters for the transformation of W [20, p. 28]. Later on, Royston improved this algorithm by using polynomials in $u = 1/\sqrt{n}$ and weights $c_i = m_i/(m'm)^{-1/2}$ derived from the plotting points $m' = (m_1, \cdots, m_n)$ defined as in (4.4) [17, p. 38].

The algorithm as described in [17] works as follows. For $n \geq 4$, define

$$\hat{a}_n = -2.706056\,u^5 + 4.434685\,u^4 - 2.071190\,u^3 - 0.147981\,u^2 + 0.221157\,u + c_n$$
$$\hat{a}_{n-1} = -3.582633\,u^5 + 5.682633\,u^4 - 1.752461\,u^3 - 0.293762\,u^2 + 0.042981\,u + c_{n-1}$$

For $n \leq 5$ resp. $n > 5$, let

$$\epsilon_n = \frac{m'm - 2m_n^2}{1 - \hat{a}_n^2} \quad \text{resp.} \quad \epsilon_n = \frac{m'm - 2m_n^2 - 2m_{n-1}^2}{1 - 2\hat{a}_n^2 - 2\hat{a}_{n-1}^2} \tag{4.5}$$

With that, define the rest of the coefficients as

$$\hat{a}_i = \frac{m_i}{\epsilon_n} \quad \text{for} \quad \begin{cases} i = 2, \cdots, n-1 & \text{if } n \leq 5 \\ i = 3, \cdots, n-3 & \text{if } n > 5 \end{cases}$$

with ϵ_n calculated according to the sample size n, as in (4.5). With these approximations of the coefficients $\hat{a}$, the value of the test statistic for the underlying sample $(x_1, x_2, \ldots, x_n)$ can be calculated as shown in (4.2) and (4.3) [17, p. 38].

For the calculation of the p-value, Royston used the approach of finding a normalizing transformation $g(W)$ that approximately follows a normal distribution. This is a common method for continuous distributions and typically uses the Box-Cox transformation

$$g(W) = (W^\lambda - 1)/\lambda$$

or a logarithmic shift

$$g(W) = \ln(W - \gamma)$$

with suitable parameters λ and γ [17, p. 39].

With simulations of samples of size $5 \leq n \leq 2000$, the distribution of the underlying statistic W was determined to find a suitable transformation by trial and error. The results for the approximating parameters of the normal distribution μ and σ, as well as the parameters used in the transformation $g(W)$ are smoothed and expressed as functions of the sample size n or of $\ln n$, typically in polynomials [17, p. 38, 39].

Royston also suggested two normalizing transformations depending on the range of the sample size n, one for $4 \leq n \leq 11$ and another for $12 \leq n \leq 2000$. The transformations and parameters are listed in Table 4.1.

Range of n	Transformation $g(W)$	u	Parameter		Coefficients
$4-11$	$-\ln\big(\gamma-\ln(1-W)\big)$	n	γ	0	-2.273
				1	0.459
			μ	0	0.5440
				1	-0.39978
				2	0.025054
				3	-0.0006714
			$\ln(\sigma^2)$	0	1.3822
				1	-0.77857
				2	-0.062767
				3	-0.0020322
$12-2000$	$\ln(1-W)$	$\ln(n)$	μ	0	-1.5861
				1	-0.31082
				2	-0.083751
				3	0.0038915
			$\ln(\sigma^2)$	0	-0.4803
				1	-0.082676
				2	0.0030302

Table 4.1.: Normalizing transformation $g(W)$ and corresponding parameter estimates by Royston, as seen in [17, p. 40]

The table is read in the following way. For $4 \leq n \leq 11$, the transformation

$$g(W) = -\ln(\gamma - \ln(1-W)) \approx \mathcal{N}(\mu, \sigma^2)$$

is approximately normal distributed $g(W) \sim \mathcal{N}(\mu, \sigma^2)$ with

$$\begin{aligned}\gamma &= 0.459\,n - 2.273\\ \mu &= -0.0006714\,n^3 + 0.025054\,n^2 - 0.39978\,n + 0.5440\\ \sigma^2 &= \exp(-0.0020322\,n^3 - 0.062767\,n^2 - 0.77857^{\,n} + 1.3822)\end{aligned}$$

In case of $12 \leq n \leq 2000$ the smoothing parameters are

$$\begin{aligned}\mu &= 0.0038915\,\ln(n)^3 - 0.083751\,\ln(n)^2 - 0.31082\,\ln(n) - 1.5861\\ \sigma^2 &= \exp(0.0030302\,\ln(n)^2 - 0.082676\,\ln(n) - 0.4803)\end{aligned}$$

With that, the transformation

$$\frac{g(W) - \mu}{\sigma} = \frac{\ln(1 - W) - \mu}{\sigma} \approx N(0, 1)$$

approximately follows the standard normal distribution.

With these transformations, the p-value can be calculated from the value of the test statistic w for the observed sample.

$$p = \mathbb{P}(\, g(W) > w \,) = 1 - \mathbb{P}(\, g(W) \leq w \,)$$

The entire procedure of the test, implemented in R code, is displayed in the Appendix C.

5. A Test based on Moments, the Jarque-Bera Test

Another way to differ between a normal distribution and a non-normal distribution is to focus on the skewness and kurtosis of a random variable X with expectation $\mathbb{E}(X) = \mu$. These are given by the following expressions in the terms of the central moments, as described in [7, p. 42]:

$$\text{Skewness:} \quad \beta_1 = \frac{\mu_3'}{{\mu_2'}^{3/2}} = \frac{\mathbb{E}(X-\mu)^3}{\{\mathbb{E}(X-\mu)^2\}^{3/2}}$$
$$\text{Kurtosis:} \quad \beta_2 = \frac{\mu_4'}{{\mu_2'}^{2}} = \frac{\mathbb{E}(X-\mu)^4}{\{\mathbb{E}(X-\mu)^2\}^{2}}\,.$$

Any normally distributed random variable has skewness $\beta_1 = 0$ and kurtosis $\beta_2 = 3$. The Jarque-Bera test, also knows as the D'Agostino-Pearson test or the Bowman-Shenton test, uses this fact, see e.g. [7, p. 42] [10, p. 1].

Let X be a random variable with expectation $\mathbb{E}(X) = \mu$ and the k-th central moment $\mu_k' = \mathbb{E}(X-\mu)^k$. For the realization $(x_1, x_2, \ldots, x_m)$ of a random sample $(X_1, \cdots, X_n)$ of i.i.d. random variables, that are all distributed like X, the k-th central sample moment is estimated as

$$m_k' = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k\,.$$

The estimators for the central sample moments can be used to estimate skewness and kurtosis of the sample [5, p. 20].

As stated by Gel and Gastwirth in [10, p. 1], the sample skewness and kurtosis asymptotically follow a normal distribution and are asymptotically independent:

$$\sqrt{n}\begin{pmatrix} \dfrac{m_3'}{{m_2'}^{3/2}} \\ \dfrac{m_4'}{m_2'} - 3 \end{pmatrix} \approx \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}6 & 0\\ 0 & 24\end{pmatrix}\right)$$

Both statistics could be used separately to test the deviance of the sample's skewness from $\beta_1 = 0$ and the sample's kurtosis from $\beta_2 = 3$ by using the standard normal distribution by normalizing the statistics with their asymptotic variances accordingly [5, p. 20].

However, as mentioned in [5, p. 21], Carlos Jarque and Anil K. Bera showed, that under a normality assumption the sum of the statistics

$$JB = \frac{n}{6}\left(\frac{m_3'}{{m_2'}^{3/2}}\right)^2 + \frac{n}{24}\left(\frac{m_4'}{m_2'} - 3\right)^2 \approx \chi_2^2$$

is asymptotically Chi-squared distributed with two degrees of freedom.

With a value z of the test statistic for an underlying sample, the p-value is calculated as

$$p = \mathbb{P}(JB > z) = 1 - F_{\chi_2^2}(z)\,,$$

where $F_{\chi_2^2}$ is the cumulative distribution function of a Chi-squared distribution χ_2^2 with two degrees of freedom.

Gel and Gastwirth [10] criticize this original version of the Jarque-Bera test and suggest a more robust version. Since the sample moments can be sensitive regarding outliers, specifically the variance, they present a modification that is using the so-called "average absolute deviation from the median" (MAAD). As the median is more robust to extreme values, the MAAD replaces the sample variance in the denominators of the Jarque-Bera statistic for an alternative measure of the spread in the underlying data.

With the sample median M, as defined in Section 3.1.2, the MAAD is defined as

$$J_n = \frac{C}{n}\sum_{i=1}^{n}|X_i - M| \quad \text{with} \quad C = \sqrt{\pi/2}$$

and sets the robust test statistic of Jarque-Bera to

$$RJB = \frac{n}{C_1}\left(\frac{m_3'}{J_n^3}\right)^2 + \frac{n}{C_2}\left(\frac{m_4'}{J_n^4} - 3\right)^2$$

with positive constants C_1 and C_2.

Under the null hypothesis, that the random sample $(X_1, \cdots, X_n)$, with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, it holds that the modified estimators for the sample skewness and kurtosis asymptotically follow a joint normal distribution.

$$\sqrt{n}\begin{pmatrix} \frac{m_3'}{J_n^3} \\ \frac{m_4'}{J_n^4} - 3 \end{pmatrix} \approx \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}\right) \tag{5.1}$$

For the proof of this, Gastwirth [10, p. 31] first proved that J_n is a consistent estimator for the sample's standard deviation σ and asymptotically follows a normal distribution, given that the sample follows a normal distribution $\mathcal{N}(\mu, \sigma)$. Secondly, with the strong law

of large numbers, it holds that J_n converges almost surely to σ. Finally, the multivariate Slutsky Theorem is applicable and implies that the robust estimators m_3'/J_n^3 and m_4'/J_n^4 asymptotically have a joint normal distribution as in (5.1).

The values for the constants C_1 and C_2 were derived from Monte Carlo simulations and can depend on the significance level α. For $\alpha = 5\%$, the recommended values are $C_1 = 6$ and $C_2 = 64$. [10, p. 31]

Similarly to the original statistic, the robust version of the Jarque-Bera test asymptotically follows a Chi-squared distribution χ_2^2 with two degrees of freedom.

$$RJB \sim \chi_2^2 .$$

The p-value is calculated as

$$p = \mathbb{P}(\, RJB > z \,) = 1 - F_{\chi_2^2}(z).$$

For more details, see [10].

For both tests the procedures, implemented in R code, are displayed in the Appendix C.

6. Goodness-of-Fit Tests based on the Empirical Distribution Function

Unlike the tests introduced in previous chapters, which were specifically constructed for identifying non-normal samples, a goodness-of-fit test can work with any simple null hypothesis, i.e. the sample is assumed to be distributed after a specified distribution function $F_0(x)$. For the derivation of the test statistics and the p-value, the sample is transformed with the assumed null distribution F_0. This transformed sample follows a uniform distribution under the null hypothesis, resulting in the ability to test for a general distribution F_0 [20, p. 99].

For statistical tests discussed here, the p-value appears uniformly if the null distribution is completely specified with all parameters and given that the test assumptions are fulfilled. However, this does not hold any longer if at least one parameter of the null distribution is unknown. For example, the null hypothesis assumes that a sample is coming from a normal distribution with known variance $\sigma^2 = 1$, but unknown mean [20, p. 99].

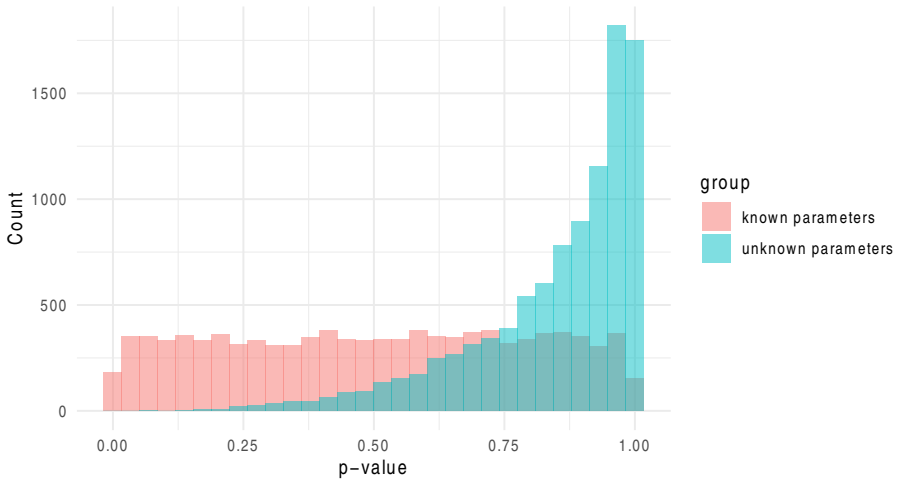

Figure 6.1.: Histogram of p-values of the Anderson-Darling test in a simulation of 10000 random samples of a standard normal distribution. For one group, the parameters were assumed to be known. For the other group, the parameters were assumed to be unknown and were estimated by sample man and sample standard deviation.

For Figure 6.1, a simulation was carried out, where 10000 samples of size $n = 100$ from a standard normal distribution $\mathcal{N}(0, 1)$ were generated and tested with the Anderson-Darling test, which is explained in detail in Section 6.4. For each sample, the test was performed twice. First, it was assumed that the parameters are known ($\mu = 0, \sigma^2 = 1$). Second, it was assumed that the parameters are unknown. Thus, for the testing procedure, the parameters were estimated by the sample mean and the sample standard deviation. The resulting p-values are shown in Figure 6.1. In the figure, it is visible that the uncertainty about the real parameters and their estimation with sample estimates can have an impact on the distribution of the test statistic. This can cause the p-values to not appear uniformly, making the testing procedure unreliable. However, under the null hypothesis this should be the case. Assume that the test statistic T follows an invertible distribution function F, that is, $F(t) = \mathbb{P}(T \leq t)$. With the Probability Integral Transformation, the transformation $F(T)$ follows a uniform distribution $U = F(T) \sim \mathcal{U}[0, 1]$ with the distribution function $\mathbb{P}(U \leq u) = u$ for $u \in [0, 1]$, see Chapter 1. With this, the following holds under the null hypothesis.

$$\mathbb{P}(U \leq u) = \mathbb{P}(F(T) \leq u) = \mathbb{P}(T \leq F^{-1}(u)) = F(F^{-1}(u)) = u\,, \quad u \in [0, 1]\,.$$

Therefore, when the p-value does not follow a uniform distribution under the null hypothesis, it could indicate that something in the testing procedure is affecting the assumed distribution of the test statistic.

In the following section, all goodness-of-fit tests are presented in the general case where all parameters in the null hypothesis are unknown. In addition, the presented goodness-of-fit tests are based on the empirical distribution function, defined in Chapter 1 or in (3.1). These so-called "EDF tests" depend on the discrepancy between the empirical distribution of the observed sample $F_n(x)$ and the hypothetical distribution of the null hypothesis $F(x)$ by measuring the vertical distances [20, p. 99].

For a normality assumption in the null hypothesis (μ, σ^2 unknown), the values of the empirical distribution function $F_n(x_{(i)}) = i/n$ are compared to the theoretical values

$$p_{(i)} = \Phi\big(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\big)\,,$$

where $\hat{\mu}$ and σ^2 are estimators.

Since the tests reject the null if the discrepancy between these two sets of values is significantly high, EDF tests are upper-tail tests, i.e. only high values of the test statistic lead to a rejection of the null hypothesis [20, p. 101].

A disadvantage of methods based on the empirical cumulative distribution function is their sensitivity toward random occurrences in the samples, such as outliers or clusters. Thus, relying only on these could influence the final decision-making. This becomes especially important for small sample sizes, which are of interest in this thesis. For more details, see e.g. [7].

Following [20] and [7], in this section, the EDF tests are grouped into two categories. The first contains tests that consider the maximum distance between the two distribution functions. The second category contains tests that consider a quadratic measure of the distance.

Supremum statistics measure the largest vertical distance between the sample's empirical cumulative distribution function and the null distribution function. The statistic D^+

$$D^+ = \sup_x (F_n(x) - F(x))$$

measures their distance on the intervals, where F_n is greater than F and the statistic D^-

$$D^- = \sup_x (F(x) - F_n(x))$$

is the measure when F is greater than F_n. [7, p. 100]

The most popular supremum statistic D introduced by Kolmogorov takes the maximum value of D^+ and D^-.

$$D = \max\{D^+, D^-\}$$

Another statistic K in this class, introduced by Kuiper, takes their sum.

$$K = D^+ + D^-$$

The second category, the quadratic statistics, are derived from the Cramér-von Mises class of statistics Q, which are specified by a function $\psi(x)$, that is weighing the squared difference between F_n and F:

$$Q = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \psi_Q(F(x)) dF(x) \tag{6.1}$$

The Cramér-von Mises statistic W^2 is derived from Q by choosing the weighing function

$$\psi_{W^2}(p) = 1\,,$$

while the Anderson-Darling statistic A^2 is constructed with

$$\psi_{A^2}(p) = (p(1-p))^{-1}\,.$$

Based on the definitions of the supremum and quadratic statistics, suitable formulas for the calculation of the statistics values are derived with the Probability Integral Transformation. Assume F is the true distribution of the sample X. Then $Z = F(X)$ follows a standard uniform distribution $Z \sim U(0,1)$ with the cumulative distribution function $F^*(z) = z$ for $0 \leq z \leq 1$. With the transformation

$$z_i = F(x_i), i = 1, \cdots, n\,,$$

the observed sample $(x_1, \cdots, x_n)$ is transformed into the uniformly distributed sample $(z_1, \cdots, z_n)$ with the corresponding empirical distribution function $F_n^*(z)$ of the standard uniform distribution with $U(0,1)$, for which the cumulative distribution function is denoted by F^* [7, p. 101].

The test statistics of the EDF test are based on a comparison of $F_n^*(z)$ with the distribution function of the standard uniform distribution. With $z = F(x)$ and the property of the uniform distribution $F(x) = F^*(F(x)) = F^*(z) = z$, it holds:

$$F_n(x) - F(x) = F_n^*(z) - F^*(z) = F_n^*(z) - z$$

Under the null hypothesis, the equality of $F_n(x) = F_n^*(z)$ is true, because of the definition of the empirical distribution function as a step functions, see (3.1).

Thus, the value of the test statistic is the same whether the empirical cumulative distribution function F_n^* of the sample $z = F(x)$ is compared to the uniform distribution or whether the empirical cumulative distribution function F_n of the observed sample x is compared to the null distribution F. Further details are available in [7, p. 101].

This results in the class of quadratic statistics of the form:

$$Q = n \int_0^1 (F_n^*(z) - z)^2 \psi_Q(z) dz \,. \tag{6.2}$$

Since the empirical distribution function $F_n^*(z)$ of the transformed sample $z = F(x)$ is a step function, the integral in (6.2) can be solved and expressed by a sum, depending on the choice of the weighing function ψ.

These formulas for the values of the supremum and quadratic test statistics, based on the transformed sample $z_{(1)} \leq \cdots \leq z_{(n)}$, as given in [7, p. 101], are:

$$D^+ = \max_{i=1,\cdots,n} \{\frac{i}{n} - z_{(i)}\} \tag{6.3}$$

$$D^- = \max_{i=1,\cdots,n} \{z_{(i)} - \frac{i-1}{n}\} \tag{6.4}$$

$$D = \max\{D^+, D^-\} \tag{6.5}$$

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(z_{(i)} - \frac{2i-1}{2n}\right)^2 \tag{6.6}$$

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)(\log z_{(i)} + \log z_{(n-i+1)}) \tag{6.7}$$

The next subsection comments on how to deal with the case that the null hypothesis only assumes a specific family of distributions, but with unknown parameters. In further subsections, the goodness-of-fit tests that are used for the power simulation in Chapter 7

are presented. The theory behind them is displayed in a high-level overview to present the formulas that are used internally in R functions, that are implemented in various R packages.

6.1. General Comments

For a goodness-of-fit test, the null hypothesis H_0 has the following form.

$$H_0 : \text{the sample } (X_1, \cdots, X_n) \text{ comes from the distribution } F(x; \theta),$$

where $F(x; \theta)$ is a cumulative distribution function with parameters θ from a certain parameter space Θ. The null H_0 states the distribution with its distributional parameters that is being tested. In the case of testing whether the sample comes from a normal distribution, θ has the form $\theta = (\mu, \sigma^2)$ with F being part of the family of cumulative distribution functions of normal distributions [7, p. 102].

A null hypothesis, where the parameters are fully specified in values, is referred to as "Case 0" or "simple hypothesis". Here, for the ordered sample X, the values $Z_i = F(X_{(i)}; \theta)$ are uniformly distributed and the null hypothesis and the formulas in (6.3) - (6.7) are used to calculate the values of the EDF statistics. However, the parameters could also be unknown and the null hypothesis only states that the sample comes from a specified family of distributions, e.g. it will be tested if the sample comes from a normal distribution with unknown mean μ and unknown variance σ^2. These testing situations are resp. referred to "Case 1", "Case 2" or "Case 3" when only μ or only σ^2 or both are unknown [7, p. 102].

For "Case 0" tests, the theory behind the EDF statistics is well studied and developed If at least one component of θ is unknown and replaced by sample estimates, the derived test statistics in (6.3) - (6.7) can still be used, but the resulting sample Z is not exactly uniform anymore, even if the null is true, as seen from the simulation in Figure 6.1. The distributions of the EDF statistics will also be different since they highly depend on the null hypothesis, the estimated parameters, the type of estimation used and the sample size [7, p. 102].

If unknown location and scale parameters are appropriately estimated, the distribution of the EDF statistic is independent of the true parameter values and only depends on the family of distributions tested as well as the sample size. However, finding these modifications for the distributions of test statistics is still a complex task and has been studied with Monte Carlo simulations for finite sample size n [7, p. 102].

The following tests focus on the general "Case 3", where both location μ and scale σ^2 parameters are unknown. The estimators in the testing procedures are given resp. by the sample mean and the sample standard deviation $\sqrt{s^2}$ [7, p. 122].

For this case, Stephens has investigated modifications of quadratic statistics as functions of the test statistic and the sample size n to account for finite samples in front of the

asymptotic background. These modifications result from the way the percentage points converge to the asymptotic points for $n \to \infty$ at a fixed significance level α. The modifications give good results for $\alpha < 0.2$, but are most accurate for a significance level of around $\alpha = 0.05$ [7, p. 103]. In particular, for the modified versions of the Anderson-Darling test and Cramér-von Mises test, both presented in detail in the following section, the percentage points of the asymptotic test statistic distribution are given in Table 4.7 in [7, p. 123].

Table 4.9 in [7, p. 127] gives formulas for the logarithmic p-values, resp. for $q := 1 - p$. They are supposedly more accurate on the upper tail ($p > 0.5$), but give good approximations on the lower tail as well. The relevant parts of the tables are shown in Sections 6.3.2 and 6.4.2 about the adjusted versions of the Cramér-von Mises and the Anderson-Darling tests [7, p. 126].

6.2. The Kolmogorov-Smirnov Test

The test statistic D of the Kolmogorov-Smirnov test considers the maximum distance between the empirical distribution function and the order statistics of the transformed sample $z = F(x)$, see (6.3) - (6.5):

$$\begin{aligned} D^+ &= \max_{i=1,\cdots,n} \{i/n - z_{(i)}\} \\ D^- &= \max_{i=1,\cdots,n} \{z_{(i)} - (i-1)/n\} \\ D &= \max\{D^+, D^-\} \end{aligned}$$

Marsaglia, Tsang and Wang provide in [15] a method for computing the distribution function of the test statistic D dependent on the sample size $n \in [2, 160000]$. Their calculation method also provides a quicker approximation that requires a shorter calculation time with the downside of a maximum accuracy up to the 7-th digit for larger samples, instead of up to the 13-th to 15-th digit.

Namely, for the probability $\mathbb{P}(D < d)$ of the event that the test statistic D_n takes a value smaller then the derived value d for the underlying sample, an expression with an $h \in [0, 1)$ and a positive integer k is used for d:

$$d = \frac{k - h}{n}\,.$$

The starting point of this method is to view the steps of the empirical cumulative distribution function of the sample as a Poisson process with intensity n and jump size $\frac{1}{n}$, similarly as Kolmogorov did, see [18, p. 66].

In accordance with results of a study by Durbin, the algorithm is based on calculating the n-th power of a matrix H with dimensions $m \times m$, that has a simple form. The value for m is $m := 2k - 1$. The approximation of the probability is then given by the element t_{kk} in the resulting matrix in row k and column k, and multiplied by a constant that depends

on n

$$\mathbb{P}(D \leq d) = \frac{n!}{n^n} t_{kk}, \text{ where } t_{kk} := (T)_{kk} = (H^n)_{kk} \text{ and } T := H^n \,.$$

For example, let $m = 6$, then the matrix H, whose n-th power results in the desired value t_{kk}, has the form:

$$H = \begin{pmatrix} (1-h^1)/1! & 1 & 0 & 0 & 0 & 0 \\ (1-h^2)/2! & 1/1! & 1 & 0 & 0 & 0 \\ (1-h^3)/3! & 1/2! & 1/1! & 1 & 0 & 0 \\ (1-h^4)/4! & 1/3! & 1/2! & 1/1! & 1 & 0 \\ (1-h^5)/5! & 1/4! & 1/3! & 1/2! & 1/1! & 1 \\ \tilde{h} & (1-h^5)/5! & (1-h^4)/4! & (1-h^3)/3! & (1-h^2)/2! & (1-h^1)/1! \end{pmatrix}$$

with

$$\tilde{h} = \begin{cases} (1-2h^m)/m! & ,0 \leq h \leq 1/2 \\ (1-2h^m+(2h-1)^m)/m! & ,1/2 < h < 1 \,. \end{cases}$$

For further details see [15].

The entire procedure of the Kolmogorov-Smirnov test, implemented in R code, is displayed in the Appendix C.

6.3. The Cramér-von Mises Test

The Cramér-von Mises test belongs to the class of quadratic EDF tests and is derived from the statistic in (6.1) with the choice of the weighting function $\psi(x) = 1$. Its test statistic is calculated with the following formula.

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(z_{(i)} - \frac{2i-1}{2n}\right)^2$$

For additional details, see [20, p. 103].

In this section, an approximation of the distribution function of the Cramér-von Mises test statistic is presented. Furthermore, a modified version based on an adjusted value of the test statistic is introduced, which depends on the sample size n.

6.3.1. Approximation of the p-value

In their paper [6], Csörgő and Faraday present an approximation for the distribution function of the Cramér-von Mises test statistic. This approximation assumes that the null distribution is fully specified with location and scale parameters and, therefore, does not consider effects from estimating the parameters. However, if the parameters need to be estimated, Braun's method (1980) can be used to adjust these effects, which supposedly works well for larger samples [8, p. 5, 7]. This method will not be investigated in further detail here.

As described at the beginning of Chapter 6, the test statistic of Cramér-von Mises is given by

$$Q = n \int_{-\infty}^{\infty} \big(F_n(x) - F(x)\big)^2 dF(x) = n \int_0^1 \big(F_n^*(z) - z\big)^2 dz$$

which can be rewritten with the empirical cumulative distribution function $F_n^*(z)$ of the transformed sample $z = F(x)$. Under the null hypothesis the sample z follows a standard uniform distribution $U(0,1)$. Therefore, the statistic is based on the comparison of $F_n^*(z)$ and $F^*(z) = z$, for $z \in [0,1]$ [6, p. 221].

Because the empirical cumulative distribution function $F_n^*(z)$ is a step function, the integral in w_n^2 can be solved as a sum based on the ordered transformed sample z and has the following boundaries.

$$\frac{1}{12n} \leq \frac{1}{12n} + \sum_{i=1}^{n} \left(z_{(i)} - \frac{2i-1}{2n}\right)^2 \leq \frac{n}{3} \tag{6.8}$$

The upper boundary $n/3$ is attained in the extreme case, that all sample values take on the highest or the lowest possible value, i.e., $z_{(n)} = 0$ or $z_{(1)} = 1$, and therefore results from integrating over z^2 or $(1-z)^2$ instead of $(F_n^*(z) - z)^2$. The other boundary results for the

other extreme case, that the sample is perfectly uniformly distributed, in which the sum in (6.8) is zero. These boundaries lead to the first approximation of the distribution function

$$V_n(x) := \mathbb{P}(\, Q^2 \leq x\,) = \begin{cases} 0 & , x \leq \frac{1}{12n} \\ 1 & , x \geq \frac{n}{3}\,. \end{cases}$$

Smirnov showed for the characteristic function of Q^2 with the imaginary unit i and $t \in \mathbb{R}$:

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}(\exp(itw_n^2)) = \lim_{n\to\infty} \int_{1/(12n)}^{n/3} \exp(itx)\, dV_n(x) &= \left(\frac{(-2it)^{1/2}}{\sinh(-2it)^{1/2}}\right)^{1/2} = \varphi(t) \\ &= \int_0^\infty \exp(itx)\, dV(x)\,. \end{aligned}$$

Further by inversion of the characteristic function $\varphi(.)$, he obtained the following formula for the asymptotic distribution V of the test statistic for $x > 0$:

$$\begin{aligned} V(x) &= 1 - \frac{2}{\pi}\sum_{k=1}^{\infty}(-1)^{k+1} \int_{(2k-1)\pi}^{2k\pi} \frac{\exp(-u^2x/2)}{(-u\sin u)^{1/2}}\, du = \\ &= \frac{1}{\pi^{3/2}x^{1/2}} \sum_{k=0}^{\infty} \frac{\Gamma(k+1/2)}{k!}(4k+1)^{1/2} \exp\Big(-\frac{(4k+1)^2}{16x}\Big)\, K_{1/4}\Big(\frac{(4k+1)^2}{16x}\Big) = \\ &= \frac{2}{\pi^{1/2}x^{1/4}} \sum_{k=0}^{\infty}(-1)^k \binom{-1/2}{k} \exp\Big(-\frac{(4k+1)^2}{16x}\Big)\, D_{-1/2}\left(\frac{4k+1}{2x^{1/2}}\right). \end{aligned} \tag{6.9}$$

The appearing identities are:

$$\begin{aligned} D_{-1/2}(2u^{1/2}) &= u^{1/4}K_{1/4}(u)/\pi^{1/2} \\ D_{1/2}(2u^{1/2}) &= \frac{u^{3/4}}{\pi^{1/2}}(K_{1/4}(u) + K_{3/4}(u)) \\ D_{3/2}(2u^{1/2}) &= \frac{u^{5/4}}{\pi^{1/2}}(2K_{1/4}(u) + 3K_{3/4}(u) - K_{5/4}(u)) \\ (-1)^k \binom{-l/2}{k} &= \frac{\Gamma(k+l/2)}{k!\Gamma(l/2)}, \quad k \in \mathbb{N}_{\geq 0}, l \in \mathbb{N}_{>0} \end{aligned} \tag{6.10}$$

with the usual gamma function $\Gamma(.)$, a modified Bessel function

$$K_\nu(y) = \frac{\pi^{1/2}y^\nu}{2^\nu\Gamma(\nu+1/2)} \int_0^\infty \exp(-y\cosh u)\sinh^{2\nu}(u)\, du, \quad y > 0 \tag{6.11}$$

and the parabolic cylinder function $D_\nu(u)$ with $u, \nu \in \mathbb{R}$ as defined in (6.10).

Additionally, Götze proposed an approximation of $V_n(x)$ by the expansion

$$V_n(x) = V(x) + \sum_{j=1}^{r} \psi_j(x) n^{-j} + O(n^{-r-1}), \quad x \in \mathbb{R}$$

for every integer $r \geq 1$ and coefficient functions $\psi_i, i = 1, \cdots, r$.

To make the computation numerically feasible, the expansion was cut to the first term

$$V_n(x) = V(x) + \frac{\psi_1(x)}{n} + O(n^{-2}), \quad x \in \mathbb{R}\,. \tag{6.12}$$

Csörgő and Faraway state, that Götze's inversion to obtain ψ_1, which is based on the Fourier-Stieltjes tranformation $\theta_1(t) = \int_0^\infty \exp(itx)\, d\psi_1(t)$ derived as

$$\theta_1(t) = \frac{1}{12}\varphi(1) - \frac{-2it}{144}\varphi(t) - \frac{1}{36}\varphi^3(t) - \frac{1}{32}\varphi^5(t) - \frac{7\cosh(-2it)^{1/2}}{288}\varphi^2(t) \tag{6.13}$$

is not correct and recalculated it. By using a method shown by Anderson and Darling and applying it to (6.13) and an inversion formula for simplification, they obtained for $x > 0$:

$$\begin{aligned}
\psi_1(x) =& \frac{1}{12}V(x) - \frac{1}{9\pi^{1/2}x^{3/4}} \sum_{k=0}^{\infty} (-1)^k \binom{-3/2}{k} \exp\Big(-\frac{(4k+3)^2}{16x}\Big) D_{1/2}\Big(\frac{4k+3}{2x^{1/2}}\Big) \\
&- \frac{7}{144\pi^{1/2}x^{3/4}} \sum_{k=0}^{\infty} (-1)^k \binom{-3/2}{k} \exp\Big(-\frac{(4k+1)^2}{16x}\Big) D_{1/2}\Big(\frac{4k+1}{2x^{1/2}}\Big) \\
&- \frac{7}{144\pi^{1/2}x^{3/4}} \sum_{k=0}^{\infty} (-1)^k \binom{-3/2}{k} \exp\Big(-\frac{(4k+5)^2}{16x}\Big) D_{1/2}\Big(\frac{4k+5}{2x^{1/2}}\Big) \\
&- \frac{1}{72\pi^{1/2}x^{5/4}} \sum_{k=0}^{\infty} (-1)^k \binom{-1/2}{k} \exp\Big(-\frac{(4k+1)^2}{16x}\Big) D_{3/2}\Big(\frac{4k+1}{2x^{1/2}}\Big) \\
&- \frac{1}{4\pi^{1/2}x^{5/4}} \sum_{k=0}^{\infty} (-1)^k \binom{-5/2}{k} \exp\Big(-\frac{(4k+5)^2}{16x}\Big) D_{3/2}\Big(\frac{4k+5}{2x^{1/2}}\Big),
\end{aligned} \tag{6.14}$$

with the parabolic cylinder functions D_ν as in (6.10).

To make their approximations more accessible for numerical computation, Csörgo and Faraway provided another form of (6.14) using the Bessel function K_ν as in (6.11) instead of $D_\nu(.)$:

$$\psi_1(x) = \frac{1}{12}V(x) + \frac{1}{\pi^{3/2}} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\Gamma(k+3/2)A_k(x)}{576x^{3/2}} + \frac{B_k(x)}{2304x^{5/2}} \right), \quad x > 0 \tag{6.15}$$

with help functions for $y > 0$:

$$\begin{aligned}
G(y) &= -\exp(-y)\big(K_{1/4}(y) + K_{3/4}(u)\big) \\
H(y) &= \exp(-y)\big(K_{5/4}(y) - 3K_{3/4}(y) - 2K_{1/4}(y)\big)
\end{aligned} \tag{6.16}$$

and

$$\begin{aligned} A_k(x) =& 7(4k+1)^{3/2}\, G\Big(\frac{(4k+1)^2}{16x}\Big) + 16(4k+3)^{3/2}\, G\Big(\frac{(4k+3)^2}{16x}\Big) + \\ &+ 7(4k+5)^{3/2}\, G\Big(\frac{(4k+5)^2}{16x}\Big) \\ B_k(x) =& \Gamma(k+1/2)(4k+1)^{5/2}\, H\Big(\frac{(4k+1)^2}{16x}\Big) + \\ &+ 24\Gamma(k+5/2)(4k+5)^{5/2}\, H\Big(\frac{(4k+5)^2}{16x}\Big)\,. \end{aligned} \tag{6.17}$$

For more details see [6, p. 220-224].

The algorithm, which was translated to R code by Adrian Baddeley [1], is based on approximating the cumulative distribution by its first order expansion, given in (6.12)

$$V_n(x) = V(x) + \varphi_1(x)/n$$

and additionally by implementing functions for the asymptotic distribution function V as in (6.9), for the first term ψ_1 as in (6.15) and its help functions shown in (6.17) and (6.16).

This version of the Cramér-von Mises test was not implemented in Appendix C, however, it is accessible in [1]. For the power simulation, an R function of the package "goftest" is used, which uses the code in [1].

6.3.2. Adjusted Test Statistic

The modification of the Cramér-von Mises test by Stephens, mentioned in Section 6.1

$$W^{2*} = (1.0 + 0.5/n)W^2$$

is supposed to account for the sample size n and was intended for $n \geq 8$. The following table is an excerpt of Table 4.9 in [7, p. 127] and contains the approximations of the p-value for this test statistic.

		W^2, Case 3
$z < z_1$	$\log q$	$= -13.953 + 775.5\,z - 12542.61\,z^2$
z_1		0.0275
$z_1 < z < z_2$	$\log q$	$= -5.903 + 179.546\,z - 1515.29\,z^2$
z_2		0.051
$z_2 < z < z_3$	$\log p$	$= 0.886 - 31.62\,z + 10.897\,z^2$
z_3		0.092
$z > z_3$	$\log p$	$= 1.111 - 34.242\,z + 12.832\,z^2$

Table 6.1.: Approximations for the p-value, resp. for $q := 1-p$, depending on the magnitude of the value of the adjusted test statistic $z = W^{2*}$ of the Cramér-von Mises test, as seen in [7, p. 127].

The table is read as follows. Assume the value of the adjusted test statistic W^{2*} takes a value in $0.0275 < z < 0.051$, then using $q := 1 - p$, the p-value is approximated as:

$$p = \mathbb{P}(W^{2*} > z) = 1 - \exp(-5.903 + 179.546\,z - 1515.29\,z^2)$$

The entire procedure of the test, implemented in R code, is displayed in the Appendix C.

6.4. The Anderson-Darling Test

The Anderson-Darling test comes from the same class of EDF tests as the Cramér-von Mises test described in (6.1), but with the weighing function $\psi(p) = (p(1-p))^{-1}$. This results in the following formula for the test statistic

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\big(\log(z_{(i)}) + \log(1-z_{(n-i+1)})\big)\,.$$

Compared to $\psi(p) = 1$, used for the Cramér-von Mises test, the weighing function of the Anderson-Darling test places higher weights on the tails of the distribution, see [20, p. 104].

In the following two sections, an approximation of the distribution function of the Anderson-Darling test statistic and a modified version based on an adjusted value of the test statistic that depends on the sample size are presented.

6.4.1. Approximation of the p-value

In their paper, Marsaglia and Marsaglia present an evaluation method for the asymptotic distribution function $\lim_{n\to\infty}\mathbb{P}(A^2 < z)$ of the Anderson-Darling test statistic with accuracy up to machine precision and provide the algorithm in a C code. In addition, they give a method to determine the distribution function for finite n, which includes an adjustment term derived from simulation. This approximation is only precise up to the 4-5th digit but requires less calculation power. [14, p. 1,2]

As mentioned in [14], Anderson and Darling expressed the asymptotic distribution of A^2 by

$$\lim_{n\to\infty}\mathbb{P}(\,A^2 < z\,) = \frac{\sqrt{2\pi}}{z}\sum_{j=0}^{\infty}\left\{\binom{-1/2}{j}(4j+1)\exp\Big(-\frac{(4j+1)^2\pi^2}{8z}\Big)\right.$$
$$\left.\int_0^{\infty}\exp\Big(\frac{z}{8(1+w^2)} - \frac{w^2(4j+1)^2\pi^2}{8z}\Big)\,dw\right\} \tag{6.18}$$

and gave values for quantiles by numerical integration.

In the C code provided by Marsaglia and Marsaglia, the asymptotic distribution is denoted as $ADinf(z)$ and is derived from the form that was achieved by Anderson and Darling in (6.18):

$$\lim_{n\to\infty}\mathbb{P}(\,A^2 < z\,) = ADinf(z) = \frac{1}{z}\sum_{j=0}^{\infty}\binom{-1/2}{j}(4j+1)f(z,j) \tag{6.19}$$

with the function $f(z,j)$ for $z\in\mathbb{R}, j\in\mathbb{N}$ defined as in [14, p. 2]:

$$f(z,j) = \sqrt{2\pi}e^{-t_j}\int_0^{\infty}\exp\Big(\frac{z}{8(1+w^2)} - w^2 t_j\Big)\,dw, \quad t_j = (4j+1)^2\pi^2/(8z)\,.$$

For the evaluation of $f(z,j)$ the term $\exp\big(z/(8(1+w^2))\big)$ is expanded into a series

$$f(z,j) = c_0 + c_1\frac{z}{8} + c_2\big(\frac{z}{8}\big)^2\frac{1}{2!} + c_3\big(\frac{z}{8}\big)^3\frac{1}{3!} + c_4\big(\frac{z}{8}\big)^4\frac{1}{4!} + \cdots \tag{6.20}$$

with the following recursion for the coefficients $c_n = \sqrt{2\pi}e^{-t_j}\int_0^\infty \frac{e^{-w^2 t_j}}{(1+w^2)^n}\,dw$, which is using the complementary error function denoted as erfc:

$$\begin{aligned} c_0 &= \pi e^{-t}(2t)^{-1/2} \\ c_1 &= \pi(\pi/2)^{1/2}\,\mathrm{erfc}(t^{1/2}) \\ c_{n+1} &= \frac{(n-1/2-t)c_n + tc_{n-1}}{n} \end{aligned} \tag{6.21}$$

With equations (6.19) - (6.21) the value of $ADinf(z)$ can be evaluated as the limit of a series:

$$ADinf(z) = \frac{1}{z}\big(f(z,0) - \frac{1}{2}\frac{5}{1!}f(z,1) + \frac{1}{2}\frac{3}{9}\frac{9}{2!}f(z,2) - \frac{1}{2}\frac{3}{2}\frac{5}{2}\frac{13}{3!}f(z,3) + \frac{1}{2}\frac{3}{2}\frac{5}{2}\frac{7}{2}\frac{17}{4!}f(z,4) - \cdots\big)\,,$$

This algorithm can be found in the C code file `ADinf.c` in [14]. In the second C file, `AnDarl.c`, the authors provide a faster but less accurate method to evaluate the asymptotic distribution function, which is called $adinf(z)$ and consists of an easier formula with two cases depending on the magnitude of z.

The simplification for $adinf(z)$ is constructed from the leading term $2z^{-1/2}e^{\pi^2/(8z)}$ in $ADinf(z)$, with $\pi^2/8 \approx 1.23370055$ exchanged by 1.2337141 to guarantee continuity in $z = 2$, and a polynomial of degree 5 for $0 < z \le 2$. This results in:

$$adinf(z) = \begin{cases} \text{for } 0 < z < 2, \text{ with } |\,\text{error}\,| < .000002 : \\ \quad z^{-1/2}e^{-1.2337141/z}\Big(2.00012 + (0.247105 - .0649821 + \\ \qquad -(.0347962 - (.0116720 - .00168691z)z)z)z)z\Big) \\ \text{for } 2 \le z < \infty, \text{ with } |\,\text{error}\,| < .0000008 : \\ \quad \exp\Big(-\exp\big(1.0776 - (2.30695 - (.43424 - (.082433 + \\ \qquad -(.008056 - .0003146z)z)z)z)z\big)\Big) \end{cases} \tag{6.22}$$

For increasing z, the authors worked with an extreme-value term $\exp(-\exp(a+bz))$. This was visually verified by plotting $\log(-\log(ADinf(z))$, which shows an approximately linear behavior. However, to keep the precision high, a polynomial with degree 5 is used, instead of a linear term. [14, p. 3]

As a last step in the construction of an approximation for the distribution of the test statistics A^2 for finite n, the function (6.22) was compared with a simulated distribution function $\mathbb{P}(\, A^2 < z\,)$. This comparison yielded an error term that depends on the sample size n and the asymptotic distribution function $ADinf(z)$. [14, p. 4]

In [14] the error term was obtained as follows. Assume that the sample Z follows the random variable A^2, then because the distribution of A^2 is close to the asymptotic distribution A_∞, the transformed sample

$$x = ADinf(Z)$$

should be close to a uniform distribution $U(0,1)$. To derive the error term, the unit interval $(0,1)$ is divided into 1000 intervals and the number of points of $ADinf(Z)$ is counted per cell. The average cell counts in the simulation were around 10^7. The error function is derived such that

$$ADinf(z) + errfix(n, ADinf(z))$$

is closely following a uniform distribution.

Figure 6.2 displays the error curves $errfix(n,x)$ for five different sample sizes n.

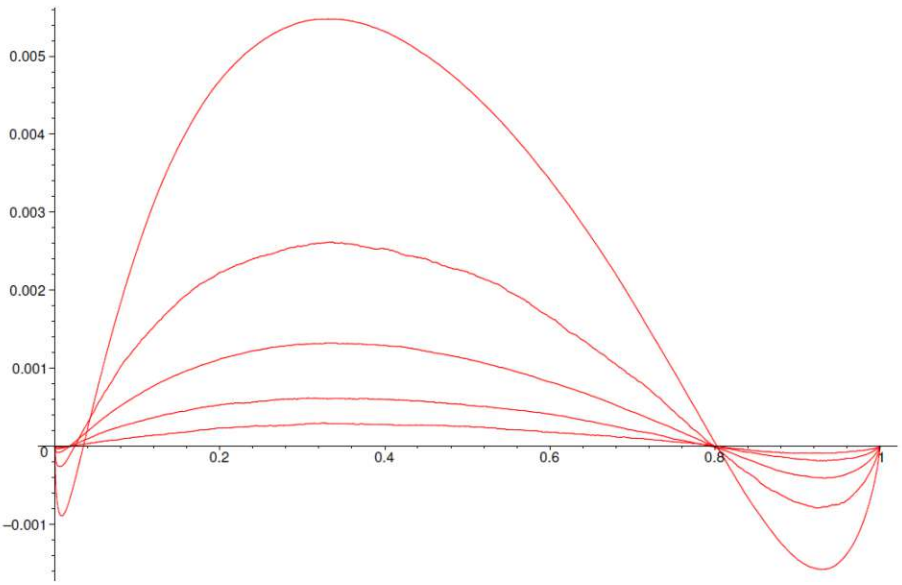

Figure 6.2.: Error terms $errfix(n,x)$ for $n = 8, 16, 32, 64, 128$ and $x \in [0,1]$, as seen in [14].

The authors achieved a piece-wise representation of the error curves with the following form:

$$errfix(n,x) = \begin{cases} (0.0037/n^3 + 0.00078/n^2 + 0.00006/n)g_1(x/c(n)) & \text{if } x < c(n) \\ (0.04213/n + 0.01365/n^2)g_2((x-c(n))/(0.8-c(n))) & \text{if } c(n) \leq x < 0.8 \\ g_3(x)/n & \text{if } 0.8 < x \end{cases} \tag{6.23}$$

with the stretch factors g_1, g_2 and g_3:

$$g_1(x) = \sqrt{x}(1-x)(49x-102)$$
$$g_2(x) = -0.00022633 + (6.54034 - (14.6538 - (14.458 - (8.259 - 1.91864x)x)x)x)x$$
$$g_3(x) = -130.2137 + (745.2337 - (1705.091 - (1950.646 - (1116.360 - 255.7844x)x)x)x)x$$

After the initial dip that can be seen in Figure 6.2 close to $x = 0$, the point where the curve is crossing the x-axis is indicated by $c(n)$ depending on the sample size n:

$$c(n) = 0.01265 + \frac{0.1757}{n}$$

The other crossing point on the upper end of the range is approximately at 0.8 for all observed sample sizes. For mroe details, see [14].

With these approximations, the distribution function of the Anderson-Darling test statistic A_n for finite n can be approximated with (6.22) and (6.23) as:

$$\mathbb{P}(A_n < z) = adinf(z) + errfix(n, adinf(z))$$

The procedure of the fast version of the test as presented in [14], implemented in R code, is displayed in the Appendix C.

6.4.2. Adjusted Test Statistic

As mentioned in Section 6.1, Stephens suggested an adjusted test statistics A^{2*} to account for the sample size n

$$A^{2*} = (1.0 + \frac{0.75}{n} + \frac{2.25}{n^2})A^2 \,.$$

It was originally intended for $n \geq 8$. [20, p. 104]

The following excerpt of Table 4.9 in [7, p. 127] shows the approximations for its p-value.

		A^2, Case 3
$z < z_1$	$\log q$	$= -13.436 + 101.14\,z - 223.73\,z^2$
z_1		0.200
$z_1 < z < z_2$	$\log q$	$= -8.318 + 42.796\,z - 59.938\,z^2$
z_2		0.340
$z_2 < z < z_3$	$\log p$	$= 0.9177 - 4.279\,z + 1.38\,z^2$
z_3		0.600
$z > z_3$	$\log p$	$= 1.2937 - 5.709\,z + 0.0186\,z^2$

Table 6.2.: Approximations for the p-value, resp. for $q := 1-p$, depending on the magnitude of the value of the adjusted test statistic $z = A^{2*}$ of the Anderson-Darling test, as seen in [7, p. 127].

The table is read as follows. Assume the value of the adjusted test statistic A^{2*} is in $0.200 < z < 0.340$, then using $q := 1 - p$, the p-value is approximates by:

$$p = \mathbb{P}(A^{2*} > z) = 1 - \exp(-8.318 + 42.796\,z - 59.938\,z^2)$$

The entire procedure of the test, implemented in R code, is displayed in the Appendix C.

7. Power Simulation

The simulation of the power of tests that was carried out for this study is described in this chapter. It is supposed to provide a basis for drawing a conclusion which of the tests, introduced in the previous chapters, are more appropriate for statistical testing for normality of samples with small sizes. Besides the process of the simulation, the set of alternative distributions is presented, which were used to simulate the rejection rate of the tests on non-normal samples. At the end of this chapter the results for the power simulation are displayed. This chapter presents original work, with code and simulations based on a real data set, and is a significant part of the contribution of this thesis.

7.1. Procedure of the Power Simulation

As mentioned in Chapter 2, the power of the test, denoted by $1 - \beta$ is defined as the probability of rejecting the null hypothesis H_0, given that it is not true,

$$1 - \beta = \mathbb{P}(\text{ reject } H_0 \,|\, H_0 \text{ is not true }) .$$

All tests consider testing the null hypothesis that the underlying sample was derived from a normal distribution. Thus, the higher the power, the better is the test in detecting non-normal samples.

Using this definition, the power of the test for a fixed sample size will be estimated by evaluating the tests on a series of non-normal distributions and deriving what proportion of tested samples were rejected. The random samples are generated from these non-normal distributions, which are fitted onto the observed data with maximum likelihood estimators.

For each of the nine alternative distributions presented in the following, 12 sets of distributional parameters are fitted to the real-world data set, mentioned in Chapter 1. From each of these 108 non-normal distributions, 1000 random samples with sample size n are drawn, on which each test is evaluated. The power of a test is then estimated by the proportion of these 1000 samples that are rejected. This process was repeated for the sample sizes $n \in \{5, 6, \ldots, 15\}$.

The results for some of the sets of distributional parameters are displayed in Section 7.3 in individual plots per alternative distribution.

Since the power simulation focuses on small sample sizes and therefore cannot be increased, the level of significance α is fixed at a predefined level, to simulate the power. As mentioned

in Chapter 2, some of the tests are specifically calibrated to a significance level of $\alpha = 0.05$. Therefore, this level is used for all tests.

For pseudo-random sampling the R function `runif` was used to produce uniform samples, which were then transformed by the inverse distribution function to create a sample of a desired distribution.

7.2. Alternative Distributions

This section presents all of the alternative distributions that were used in the power simulation. These were selected since they are well-known continuous distributions apart from the normal distribution.

The Cauchy distribution is often used in power simulation as an indicator of how sensitive a test is to detect deviations from a normal assumption due to heavy tails [11].

In the following, the density function for each distribution is displayed, based on which the likelihood and the log-likelihood functions are constructed. These were used for the maximum likelihood estimation of the distributional parameters, as explained in Section 2.2. The maximum likelihood estimators are obtained by solving the system of nonlinear equations, the so-called normal equations

$$\frac{\partial}{\partial\theta} \ln L(\theta \,|\, x_1, \dots x_n) = 0\,,$$

which for some distributions reduce to a system of linear equations. However, in general the system of equations do not have a closed-form solution and must be solved using numerical methods, such as the optimization method of Newton-Raphson, used by the R function `optim()`.

The definition of the distributions are standard, but the notation of [21] is used.

Student's t-Distribution with a small Degree of Freedom

The probability density function of the Student's t-distribution $T \sim t_\nu$ with ν-degrees of freedom, $\nu > 0$, $\nu \in \mathbb{N}$, is

$$f(x \,|\, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}(1 + \frac{x^2}{\nu})^{-\frac{\nu+1}{2}}\,, \quad x \in \mathbb{R}\,,$$

where Γ denotes the Euler Gamma function.
The distribution has an expected value of $\mathbb{E}(T) = 0$. It is symmetric and has heavier tails than the normal distribution. Since the real-world data set considered in this thesis is not located around zero, the observed samples are fit to a Student's t random variable with location parameter $m \in \mathbb{R}$ and scale parameter $s > 0$, with the transformation

$$X \overset{d}{=} m + s\,T\,.$$

The distribution function of X becomes

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(m + s\,T \leq x) = \mathbb{P}(T \leq \frac{x-m}{s}) = F_T(\frac{x-m}{s})$$

and with the chain rule for differentiating the density is

$$f_X(x) = \frac{d}{dx}F_X = \frac{d}{dx}F_T(\frac{x-m}{s}) = \frac{1}{s}f_T(\frac{x-m}{s})\,.$$

For the maximum likelihood estimation, the R function `optim` minimizes the negative log-likelihood function regarding m, s and ν, which is given by

$$-\sum_{i=1}^{n} \ln\left(\frac{1}{s}f_T(\frac{x-m}{s})\right).$$

When fitting a Student's t-distribution to an underlying sample, it might happen that the degrees of freedom become very large, as was the case with most of the observed samples in the real-world data set.

Since the Student's t-distribution approximates the Normal distribution with such a high degree of freedom, the fitted distribution might not be representative for a Student's t-distribution [21, p.143]. Therefore, the degree of freedom is set to the sample size of the observed samples minus one, i.e., $\nu = n - 1$. The other two parameters are estimated with the maximum likelihood method.

Log-Normal Distribution

The probability density function of the log-normal distribution $T \sim Lognormal(\mu, \sigma^2)$ with variables $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$ is given by:

$$f(x \,|\, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} \quad , x \in \mathbb{R}_+\,.$$

According to the R documentation of the package `MASS`, the function `fitdistr` can calculate the maximum likelihood estimators for a number of distributions, which also includes the log-normal distribution. It uses either the closed form of the corresponding likelihood function or estimates the parameters by optimizing the log-likelihood function with `optim()` similar to the approach shown for the Student's t-distribution. [16]

Uniform Distribution

The probability density function of the uniform distribution $T \sim U(a, b)$ with lower bound $a \in \mathbb{R}$ and upper bound $b \in \mathbb{R}$ is given by:

$$f(x) = \frac{1}{b-a} \quad \text{for } \; a \leq x < b$$

Following the approach of maximizing the log-likelihood function, the estimators for a and b are given by the minimum and maximum value of the underlying sample X, i.e. the first and last order statistic

$$a = \min(X_1, \cdots, X_n) = X_{(1)} \text{ and } b = \max(X_1, \cdots, X_n) = X_{(n)}$$

Weibull Distribution

The probability density function of the Weibull distribution $T \sim Weibull(\eta, \sigma)$ with shape parameter $\eta \in \mathbb{R}_+$ and scale parameter $\sigma \in \mathbb{R}_+$ is:

$$f(x \,|\, \eta, \sigma) = \frac{\eta}{\sigma}\big(\frac{x}{\sigma}\big)^{\eta-1} e^{-(\frac{x}{\sigma})^{\eta}} \quad , x \in \mathbb{R}.$$

Its likelihood function is

$$L(\eta, \sigma \,|\, x_1, \cdots, x_n) = \Pi_{i=1}^{n} \big(\frac{\eta}{\sigma}\big)\big(\frac{x_i}{\sigma}\big)^{\eta-1} \exp\big(-\big(\frac{x_i}{\sigma}\big)^{\eta}\big)$$

and the log-likelihood function is

$$\ln L(\eta, \sigma \,|\, x_1, \cdots, x_n) = n \ln \eta - n\eta \ln \sigma + (\eta - 1) \sum_{i=1}^{n} \ln x_i - \frac{1}{\sigma^{\eta}} \sum_{i=1}^{n} x_i^{\eta}$$

Laplace (Double-Exponential) Distribution

The density function of the Laplace distribution $T \sim Laplace(\mu, \lambda)$ with parameters $\mu \in \mathbb{R}$ and $\lambda > 0$ is given by

$$f(x \,|\, \mu, \lambda) = \frac{\lambda}{2} e^{-\lambda |x - \mu|} \,.$$

Its likelihood function is given by

$$L(\mu, \lambda \,|\, x_1, \cdots, x_n) = \frac{\lambda^n}{2^n} \exp(-\sum_{i=1}^{n} \lambda |x_i - \mu|)$$

and the log-likelihood function is

$$\ln L(\mu, \lambda \,|\, x_1, \cdots, x_n) = n \ln \lambda - n \ln 2 - \lambda \sum_{i=1}^{n} |x_i - \mu| \,.$$

Gumbel Distribution

The density function of a Gumbel distribution $T \sim Gumbel(\mu, \beta)$ with the location parameter $\mu \in \mathbb{R}$ and the scale parameter $\beta > 0$ is given by

$$f(x \,|\, \mu, \beta) = \frac{1}{\beta} e^{\frac{x-\mu}{\beta}} e^{-\exp(\frac{x-\mu}{\beta})} \quad , x \in \mathbb{R} \,.$$

Its likelihood function is given by

$$L(\mu, \beta \,|\, x_1, \ldots, x_n) = \frac{1}{\beta^n} \exp(\sum_{i=1}^{n} \frac{x_i - \mu}{\beta}) \exp(-\sum_{i=1}^{n} \exp(\frac{x_i - \mu}{\beta}))$$

and the log-likelihood function is

$$\ln L(\mu, \beta \,|\, x_1, \cdots, x_n) = -n \ln \beta - \sum_{i=1}^{n} \frac{x_1 - \mu}{\beta} - \sum_{i=1}^{n} \exp(-\frac{x_i - \mu}{\beta}) \,.$$

Cauchy Distribution

The Cauchy distribution $T \sim Cauchy$ has a density function of the form

$$f(x) = \frac{1}{\pi}(1 + x^2)^{-1}$$

It has a median at zero, its expected value and its variance do not exit. [21, p. 26]

Similar to the Student's t-distribution, for the simulation, a Cauchy distribution with location $l \in \mathbb{R}$ and scale $s > 0$ is used for fitting with the maximum likelihood method. Then the density becomes

$$f(x) = \frac{1}{\pi s}\Big(1 + (\frac{x - l}{s})^2\Big)^{-1} \quad , x \in \mathbb{R} \,.$$

Its likelihood function is given by

$$L(l, s \,|\, x_1, \cdots, x_n) = \frac{1}{(\pi s)^n} \Pi_{i=1}^{n} \Big(1 + (\frac{x_i - l}{s})^2\Big)^{-1}$$

and its log-likelihood function is

$$\ln L(l, s \,|\, x_1, \cdots, x_n) = -n \ln(\pi s) - \sum_{i=1}^{n} \ln(1 + (\frac{x_i - l}{s})^2) \,.$$

Gamma Distribution

The probability density function of the gamma distribution with positive shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ is

$$f(x \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\beta x} \text{ for } x > 0 \,.$$

The likelihood function is given by

$$L(\alpha, \beta \,|\, x_1, \cdots, x_n) = \frac{\beta^{\alpha n}}{\Gamma(\alpha)^n} \prod_{i=1}^{n} x_i^{\alpha - 1} e^{-\beta x_i}$$

and the log-likelihood function is

$$\ln L(\alpha, \beta \,|\, x_1, \cdots, x_n) = \alpha n \ln \beta - n \ln \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \ln x_i - \beta \sum_{i=1}^{n} x_i \,.$$

Triangular Distribution

The probability density function of a triangular distribution with range $a \leq x \leq b$ for $a, b \in \mathbb{R}$ and shape parameter c, such that $a \leq c \leq b$ is

$$f(x \mid a, b, c) = \begin{cases} \dfrac{(x-a)^2}{(b-a)(c-a)} & \text{if } a \leq x \leq c \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-c)} & \text{if } c \leq x \leq b\,. \end{cases}$$

The log-likelihood function used for the maximum likelihood estimation is

$$\sum_{i=1}^{n} \ln f(x_i \mid a, b, c)\,.$$

7.3. Results of the Power Simulation

The figures in the following section show the results of the power simulation. Each plot contains the results for one distribution and one of the 12 sets of distributional parameters that were used to generate 1000 samples. Of the 12 sets of distribution parameters fitted to the data set, four representatives were selected for display:

- sample 4 behaves like sample 8 for all the alternative distributions,
- sample 6 shows unique patterns, when the results for all alternative distributions are considered as one combination,
- sample 9 shows unique patterns as well and
- sample 11 behaves mostly like samples 1, 2, 3, 5, 7, 10, 12, except for the triangular distribution, where most of these samples show results similar to samples 4 or 6.

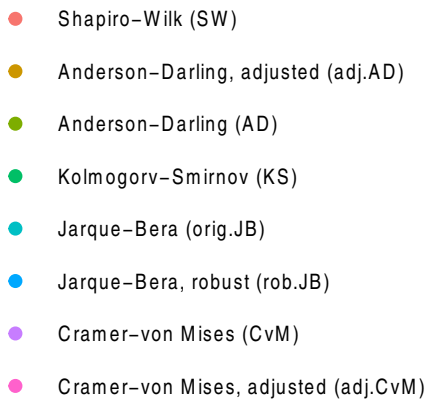

Figure 7.1.: Legend for identifying the normality tests.

One set of distributional parameters is referred to as a sample, as they were fitted onto one of the samples of the data set, described in Chapter 1. The value on the y-axis in the graphs is the proportion of 1000 samples for which the normality assumption is rejected, that is, the calculated p-value was below the α-level of 5%. These are plotted against the sample size of the generated random samples on the x-axis.

In addition, in the upper left corner of each plot, the mean and standard deviation (in the form $mean +/- std.dev.$) of the estimated sample skewness and kurtosis values of all the random samples used for each plot are displayed. The formulas for these are presented in Section 2.3.

For the analysis of the simulation results, the alternative distributions are divided into four groups. Each group is described separately below. Furthermore, the original versions of the Anderson-Darling test, Cramér-von Mises test and the Jarque-Bera test did not perform nearly as well as their adjusted, more robust versions. Therefore, the following discussion will not focus on them. Similarly, the Kolmogorov-Smirnov test was unable to reject nearly as many samples. Further discussion of their performance is given in the next Section 7.4. A prominent feature of these tests, that distinguishes them from the others adjusted tests, is their dependence on asymptotic assumption for determining the test statistic. For instance, the original and the robust versions of the Jarque-Bera test both depend on the asymptotic distribution of their test statistics. When this assumption is not satisfied for finite samples, or due to the interference of the estimation of the distributional parameters of the goodness-of-fit tests as mentioned in Section 6.1, the resulting

distribution of the test statistic differs from the theoretical distribution, resulting in a low power.

The triplet of the adjusted Anderson-Darling test, the adjusted Cramér-von Mises test and the Shapiro-Wilk test performed better and show a similar behavior in each graph, whereas the robust version of the Jarque-Bera test has a different performance on heavy-tailed and light-tailed distributions than this triplet.

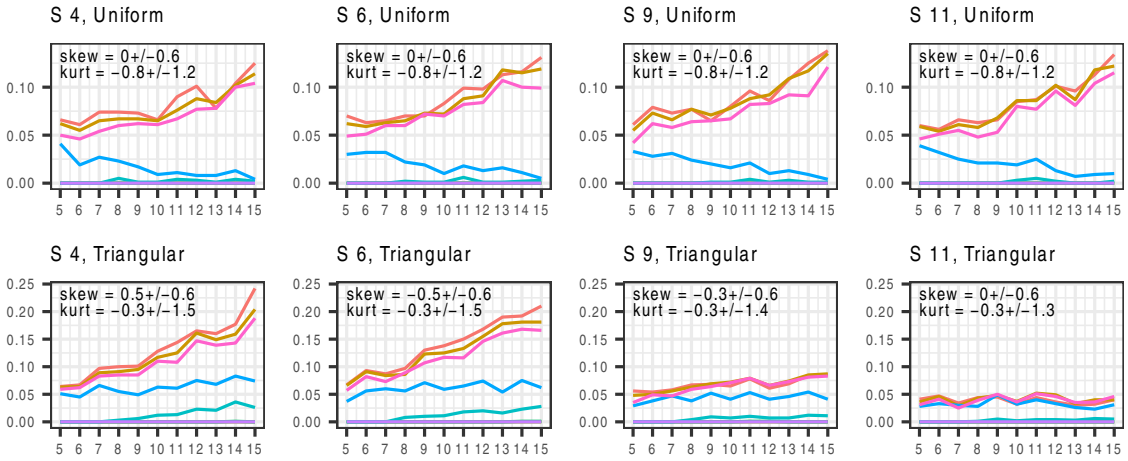

Figure 7.2.: Results of the simulation on group 1, the light-tailed distributions.

The first group consists of light-tailed distributions, represented by the uniform and the triangular distribution. The results in Figure 7.2 seem to indicate that the robust version of the Jarque-Bera test does not perform well on light-tailed distributions at all, unlike the triplet of the Shapiro-Wilk test, the adjusted Cramér-von Mises test and the adjusted Anderson-Darling test. These show a low power for only two of the triangular distributions, where the mean sample skewness and kurtosis values are close to the values of a normal distribution.

Figure 7.3 displays the results for the second group of alternative distributions, which are the log-normal, Weibull, and gamma distributions. Depending on the distributional parameters, these distributions can be asymmetric with heavier tails than a normal distribution. Regarding the distributions in this figure, the higher the mean sample kurtosis and the more the mean sample skewness differs from zero, the higher is the rejection rate. However, when the sample skewness and sample kurtosis values are around the values of the normal distribution, the tests seem to be unable to detect non-normal samples.

The third group of distributions contains distributions that are heavy-tailed distributions independent of their distributional parameters, which are the Laplace, the Gumbel and the Cauchy distributions. Overall, the four tests in Figure 7.4 show good power even for the smallest sample sizes. For Laplace distributions, the robust version of the Jarque-Bera test

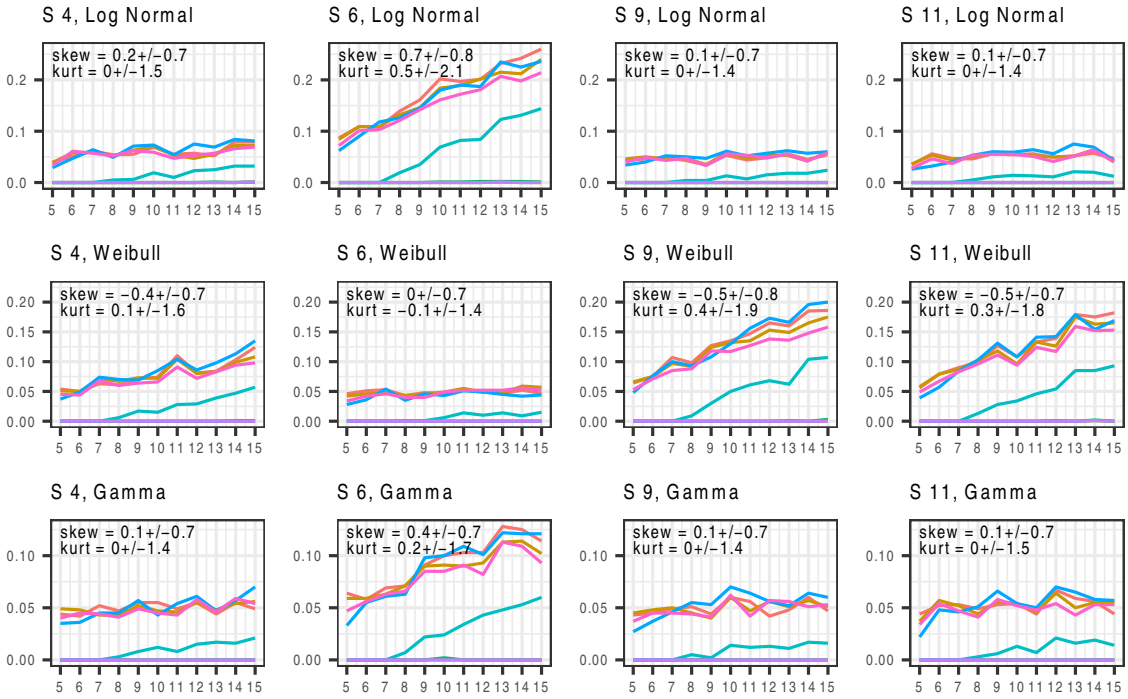

Figure 7.3.: Results of the simulation on group 2, distributions that can have heavy tails depending on the parameters.

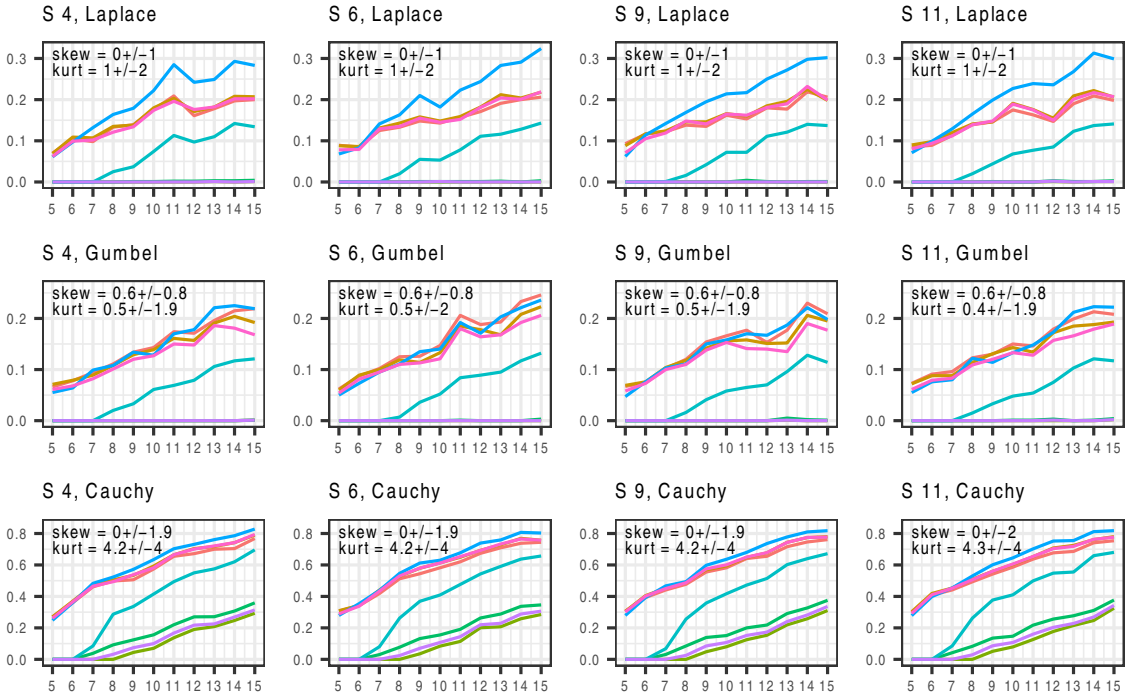

Figure 7.4.: Results of the simulation for group 3, the heavy-tailed distributions.

has even higher power than all other tests.

The Cauchy distribution is the only distribution for which even the other statistical tests, i.e. the original version of the Anderson-Darling test and Cramér-von Mises test, as well as the original Jarque-Bera test and the Kolmogorov-Smirnov test, show a power that is noticable higher than zero, but still they are not nearly as powerful as their adjusted and more robust versions.

As stated by David Harris in [11], the Cauchy distribution is described as an important example of a distribution that "looks similar to a normal distribution" but has much heavier tails. This has turned the Cauchy distribution into an indicating distribution that can reveal how sensitive a normality test is to heavy-tail departures. Furthermore, the Cauchy distribution can test how robust a test's assumptions are since its mean and standard deviation are undefined. Practically, this means that a large sample of data points, collected from a Cauchy distribution, will not provide more accurate estimates for the moments than a single data point.

Considering this key figure, that the Cauchy distribution has in power simulations, and considering the results displayed in the above figure, the original versions of the normality tests, i.e. the Anderson-Darling, Cramér-von Mises, Kolmogorov Smirnov and the original version of the Jarque-Bera test might not be sensitive enough to detect departure from the normal distribution in samples with small sizes.

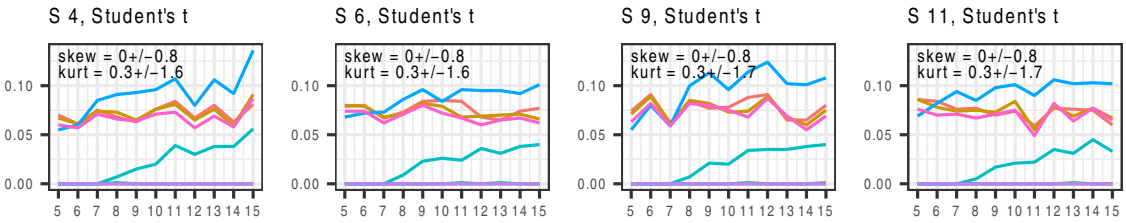

Figure 7.5.: Results of the simulation for group 4, the Student's t distribution.

Lastly, in group 4, containing the samples of the Student's t distribution, which have a degree of freedom $\nu \in \{9, 4, 5, 6\}$, the triplet of the tests were unable to reject more than 10% of the samples and even the robust version of the Jarque-Bera test was barely able to achieve a power significantly higher than 10%, as displayed in Figure 7.5. Although, with a degree of freedom in this range, the Student's t-distribution is considered to be a heavier-tailed distribution, the tests did not perform notably well.

In Figure 7.6 the mean of the estimated skewness and kurtosis values of the random samples generated for each of the alternative distributions and the 12 sets of distribution parameters are shown to give an overview of the alternative distributions. The values are estimated using the formulas for G_1 and G_2 presented in Section 2.3, which consider an adjustment for the sample size. Because of this, the values might differ from the theoretical skewness

and kurtosis values of the distributions calculated with the central moments. Furthermore, the kurtosis used here is the so-called "excess kurtosis", which is derived as:

$$\text{excess kurtosis} \;=\; \text{kurtosis} - 3$$

Therefore, the normal distributions here have an estimated kurtosis of around 0, instead of around 3, as described in Chapter 5 for the Jarque-Bera test.

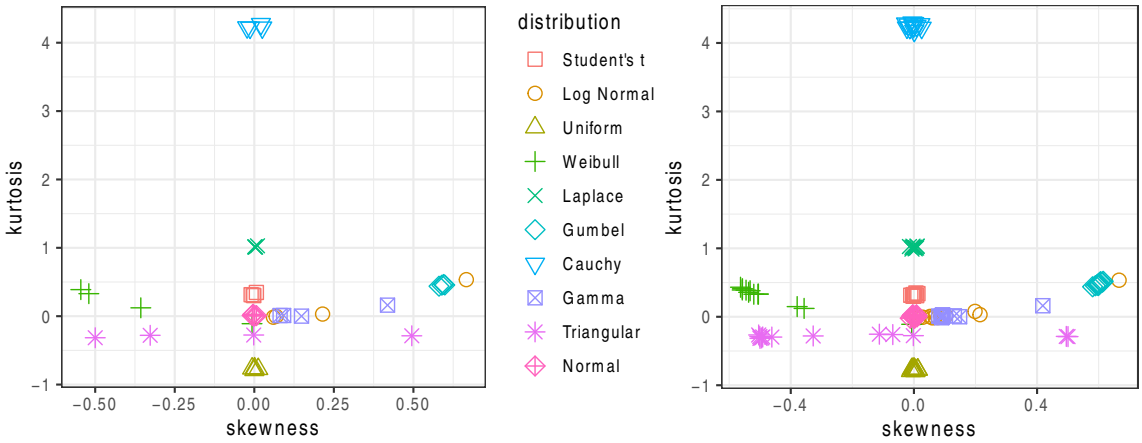

Figure 7.6.: Mean of the estimated skewness and kurtosis values for the generated random samples for each of the alternative distributions and the 4 displayed (on the left) resp. all 12 (on the right) sets of distributional parameters.

For distributions for which the mean of the estimated skewness and kurtosis values are close to those of the normal distribution, the tests did not show a power significantly higher than the alpha level of $\alpha = 5\%$. This affects all of the displayed Student's t-distributions, part of the log-normal, one of the Weibull, part of the gamma and one of the triangular distributions. In Figure 7.6, these affected distributions can be identified through the symbols close to the origin $(0, 0)$.

The investigation of the results continues in Section 7.4.

7.4. Performance of the Tests on Normal Distributions

Considering the behavior of the tests both for alternatively distributed samples and normally distributed samples can provide a more comprehensive comparison or even an insight into the behavior of certain tests.

Figure 7.7 displays the simulation results of the tests for four of the 12 fitted normal distributions. Since the probability of making a Type I error is fixed at a significance level of 5%, it is expected that about 5% of the tested samples are rejected by the tests.

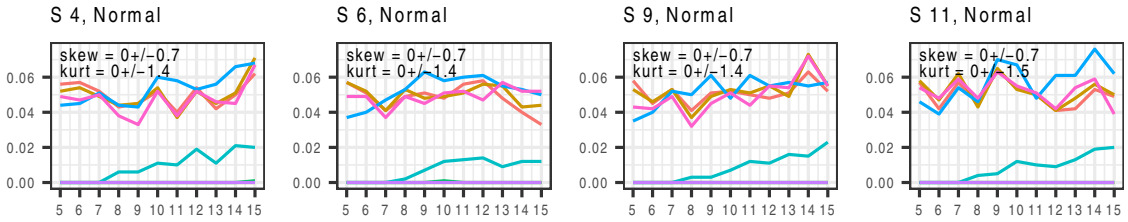

Figure 7.7.: Simulation of proportion of rejected normal samples.

For the adjusted tests of Anderson-Darling and Cramér-von Mises, as well as the Shapiro-Wilk test and the robust version of the Jarque-Bera test, the rejection rates of normally distributed samples seem to fluctuate around $\alpha = 0.05$.

For the original version of the Jarque-Bera test, the rejection rate is able to rise above a rate of 0%, however, it cannot reach a level of around 5%. The other three tests, i.e. the Kolmogorov-Smirnov test and the original versions of the Anderson-Darling and Cramér-von Mises tests, do not seem to be able to reject even a small proportion of the normal samples.

As mentioned in Chapter 2, one of the key elements of a hypothesis test is the distribution that the test statistics follow given that the null hypothesis is true. Some tests, as is the case for the Kolmogorov-Smirnov test, rely on asymptotic assumptions for the approximation of the distribution. When these tests are applied to finite samples, the p-value is still approximated assuming asymptotic assumptions, which might not be appropriate for small samples. This can result in a lower than expected rejection rate when the null hypothesis is actually true. If this is the case, the test is called a conservative test. For further details on conservative tests, see [9]

For conservative hypothesis tests, the decreased rejection rate of a true null hypothesis can affect the power of the test, i.e. its ability to detect a sample for which the null is actually not true, might be reduced as well. Fisher and Robbins mention in [9] that there are methods and transformations for conservative tests to improve the probability of making a false decision.

Examples of such corrections are the adjusted versions of the Anderson-Darling test and the Cramér-von Mises test, as described in [7].

Figure 7.8 shows the histograms of the p-values that were calculated in the evaluation of the normality tests for one of the normal distributions. For the Shapiro-Wilk test and the adjusted versions of the Anderson-Darling test and the Cramér-von Mises test, the distribution of the p-values approximate a uniform distribution. Under the null hypothesis and when the test assumptions are fulfilled, this is expected, as explained in Chapter 6.

However, for the original versions of the Anderson-Darling test and Cramér-von Mises test, as well as the Kolmogorov-Smirnov test and the Jarque-Bera test, the p-values do not appear to be distributed uniformly anymore. This could be a result of the asymptotic assumption of the tests, which is not satisfied for testing finite or small samples, as mentioned by Fisher and Robbins in [9].

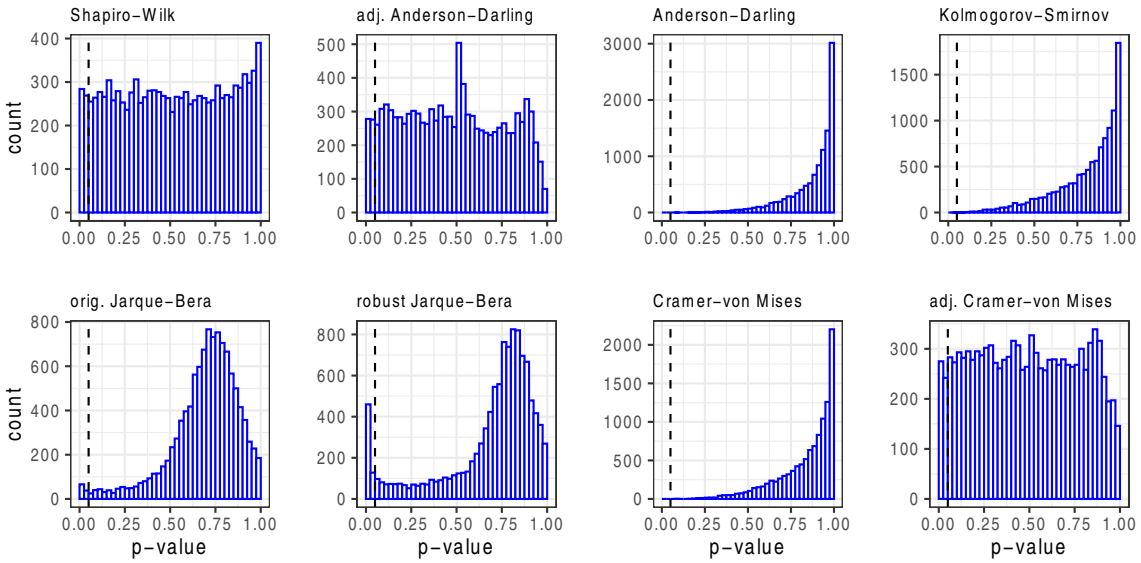

Figure 7.8.: Histograms of the p-values that were calculated for the evaluations of each test on the 4^{th} set of distributional parameters (S 4)

Although the robust version of the Jarque-Bera test performed similar to the adjusted tests of Anderson-Darling or Cramér-von Mises, or the Shaprio-Wilk test in the simulation of the alternative distribution (Section 7.3), the test shows a distinct distribution of the p-values. The distribution is skewed, but has almost a point probability for the event that the p-value is below $\alpha = 5\%$. This could be the case because the test statistic depends on asymptotic assumptions, but the constants that appear in the test statistic are specially calibrated towards a level of significance of $\alpha = 5\%$ in Monte Carlo simulations.

In addition, as mentioned in Chapter 6, having to estimate the mean and standard deviation for the null hypothesis can also interfere with the distribution of the test statistic and therefore cause a deviation in the distribution of the p-value and in the distribution of the test statistic, making the original version of the goodness-of-fit tests unreliable.

Therefore, it is important to validate whether the α-value is kept. If the simulation results differ too much, it might indicate that the test does not behave as assumed under the null hypothesis, since the underlying assumptions are not satisfied. This can indicate that the test is not appropriate for testing small samples as is but requires some kind of correction, as mentioned in [9].

8. Evaluation of the Tests on the Real-World Data Set

Following the simulation-based comparison of the normality tests, in this chapter, the tests are evaluated on real-world data, consisting of 12 samples with sizes varying in $n \in \{5, \ldots, 10\}$, as described in Chapter 1.

The real-world data set it visualized in several diagnostic plots in Chapter 3, in histograms (see Figures 3.2 - 3.3) and boxplots (see Figure 3.5), as well as in Q-Q plots (see Figure 3.7) and P-P plots (see Figure 3.9).

Table 8.1 displays the results of the tests, when evaluated on the observed data samples.

	SW	adj.AD	AD	KS	orig.JB	rob.JB	CvM	adj.CvM
S_1	0.2781	0.2794	0.8375	0.5937	0.6261	0.7015	0.7940	0.2950
S_2	0.5989	0.5148	0.9367	0.9495	0.8902	0.9118	0.9068	0.5258
S_3	0.3110	0.4233	0.9052	0.8866	0.6559	0.7513	0.9213	0.5694
S_4	0.1811	0.2456	0.8158	0.9025	0.5968	0.7011	0.8064	0.3140
S_5	0.6430	0.6134	0.9745	0.9876	0.7996	0.8354	0.9651	0.6999
S_6	0.4096	0.3913	0.9280	0.8538	0.7431	0.8302	0.8869	0.4144
S_7	0.3325	0.3268	0.8904	0.7930	0.7360	0.7932	0.8635	0.3822
S_8	0.9102	0.9015	0.9989	0.9957	0.8120	0.8490	0.9986	0.9316
S_9	0.1848	0.1137	0.7268	0.5323	0.5960	0.1450	0.6329	0.1073
S_{10}	0.5501	0.5889	0.9660	0.9377	0.7339	0.8127	0.9541	0.6652
S_{11}	0.7327	0.7312	0.9849	0.9833	0.8772	0.9338	0.9856	0.8274
S_{12}	0.9492	0.8869	0.9981	0.9896	0.8758	0.9408	0.9950	0.8825

Table 8.1.: p-values of the tests for each observed sample.

The resulting p-values are also visualized in Figure 8.1. In the left plot, the p-values of all tests are displayed, and in the right plot only the p-values of the Shapiro-Wilk test, the adjusted Anderson-Darling test, the robust version of the Jarque-Bera test, and of the adjusted Cramér-von Mises test are shown. These four tests showed a higher power in the simulation from the previous chapter. The significance level of $\alpha = 5\%$ is indicated by a horizontal line.

None of the samples are rejected by any of the hypothesis tests at the α-level of 5%. Initially, this means that there is no sufficiently strong evidence speaking against the normality assumption that was detected by the tests, which would justify rejecting the null hypothesis. However, this does not confirm that the samples actually come from a normal distribution, only that assuming normality is plausible under the testing conditions.

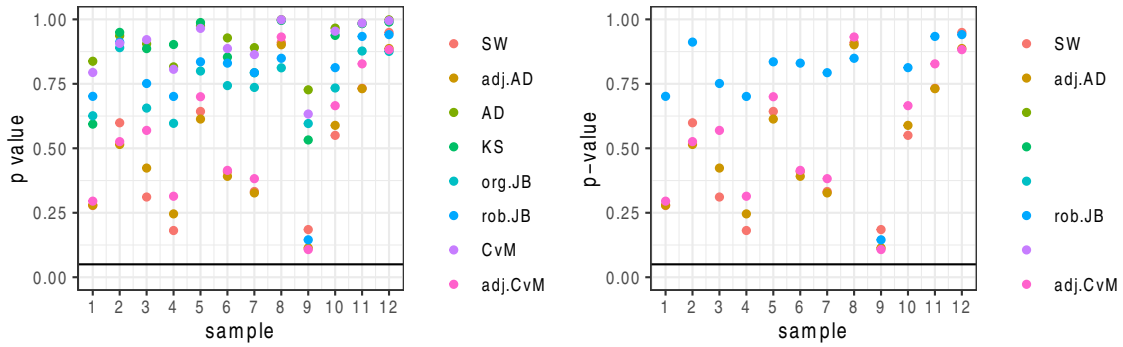

Figure 8.1.: Visualization of the p-values in Table 8.1, in the left plot for all tests, in the right plot for the four best performing tests considering the previous section.

Especially considering the small sample sizes and the insight into the performance of the tests provided in Sections 7.3 and 7.4, the results have to be interpreted with caution. The simulation suggested that the tests have a rather low power for small sample sizes, meaning that the tests may not be able to detect moderate deviations from normality.

From a technical perspective, the results and the power simulation imply that the tests do not provide strong evidence against the null hypothesis, maybe because for such small samples the tests are not sensitive enough or maybe because there is actually no strong evidence that would suggest that assuming normality is questionable. However, the hypothesis of normality can still be considered reasonable, provided that no other concerns arise, for instance through other methods, like graphical evaluations with plots, as presented in Chapter 3.

Graphical methods, such as QQ-plots or boxplots, can provide a more comprehensive assessment. They might be able to give more insight into the test's results and reveal statistical occurrences in the sample, such as asymmetry, clusters or outliers.

In the case of the observed data samples, the plots given in Chapter 3 were either not suitable for such small sizes, such as the histograms, or largely supported the results of the hypothesis tests, i.e., they did not show substantial visual departures from normality or distinct patterns hinting at alternative distributions. However, for a few samples, there can be mild skewness or potential outliers observed in the diagnostics plots.

The interpretation of the results from the statistical tests and the graphical methods has to be done with careful consideration, since multiple points can affect the relevance of the visual representation or the performance of the tests.

- Underlying assumptions, such as asymptotic assumptions when applied to finite samples, might not be satisfied when testing the considered samples. This can affect the testing procedure in such a way that the test statistic of the hypothesis test no longer

follows its assumed distribution. Therefore, the outcome of the test might not be reliable.

- Some of the presented methods are not able to test for a normal distribution in general but need specific parameters for the mean and the standard deviation. Therefore, these must be estimated first to be able to test for a specific normal distribution, bringing additional uncertainty into the process. The estimates might not give an appropriate representation of the true location and scale parameters. Furthermore, the step of estimating the distributional parameter as part of the testing procedure might affect the test statistic of a test such that it deviates from its assumed distribution, which might also make the test results unreliable.

- Most importantly, a small sample size can greatly affect how suitable or sensitive the methods are for representing departures from normality. For small samples, graphical methods might show curvature in the tails, even if the null hypothesis is actually true, as mentioned in Section 3.2. For hypothesis tests, the smaller the sample, the less information it provides for the test to detect the departure from normality.

9. Conclusion

Evaluating if a dataset follows a normal distribution is an important part in statistical analysis, especially when the analysis or a model depends on this assumption. This thesis has shown multiple techniques for evaluating a normality assumption, ranging from graphical methods to hypothesis testing. In addition, a way of comparing the hypothesis tests was presented, where the power of the tests was simulated on a series of alternative distributions. Since the motivation for this thesis was a real-world dataset, all methods focus on samples of small sizes. The purpose of the thesis is to provide insight into how tests can be compared and to identify which of these are more suitable for the testing environment of small samples.

Graphical approach

Graphical methods can provide valuable information on the characteristics of a sample. These methods are particularly useful for identifying patterns, clusters, and outliers that may influence the results of formal statistical tests. In this sense, graphical approaches can be more informative than statistical tests, offering a more detailed view on the structure of a sample.

However, graphical methods are inherently subjective since they do not follow a standardized procedure that provides a decision on the investigated null hypothesis, i.e. the normality assumption. Their interpretation can be influenced by user experience, expectations, or visual biases. Especially for small samples, where random variation can greatly influence the representation through the plots, the interpretation needs to be done with caution.

For instance, the empirical distribution function can be sensitive to random occurrences in a sample, especially in small samples, where even one individual data point can distort the visual representation of the underlying distribution. Thus, only relying on it could influence the decision making.

Histograms, which show the frequency of observations in bins, can approximate the sample distribution and thereby reflect the density function of the underlying distribution. However, the appearance of a histogram strongly depends on the number of bins. For small samples, it is often not possible to find a suitable number, which can make histograms inappropriate.

Boxplots also become less reliable in small samples. Their spread, indicated by the box and whiskers, is highly affected by random occurrences. While they can be useful for identifying potential outliers, they are not that well suited for assessing normality in small datasets, since the whole sample is summarized by a set of five statistics.

Probability plots, such as Q-Q plots, compare observed values of the order statistics with their theoretically expected values under the null hypothesis. When the null hypothesis is plausible, these graphs display a roughly linear pattern, making the evaluation easier. Small deviations from this linearity are typically caused by random variation, while a distinct curvature may indicate the presence of an alternative distribution. For a sample of a small size, the sample might not provide enough information about how the data behaves at the tails. Therefore, these plots might show curvature in the tails even when the null hypothesis is true.

The detrended versions of the Q-Q or P-P plots shows the vertical distances of the displayed data points to the linear relation. Since these plots are not as well known, they might be harder to interpret.

While graphical methods are valuable for identifying structural characteristics of the data, relying only on them can lead to incorrect conclusions, especially because of their sensitivity to random variation in small samples. However, graphical methods can prevent misinterpretations of test results that might lead to an inappropriate conclusion. Therefore, it is important to complement them with more objective testing procedures.

Hypothesis testing

While graphical methods can provide understanding of distributional characteristics through visual representation, statistical tests offer a more objective process to assess normality. These tests quantify the deviation of the observed sample from the null hypothesis and result in a decision about whether the hypothesis is plausible enough for the sample.

The presented normality test are divided into three categories, depending on their approach for detecting departures from normality: regression-based tests, moment-based tests and goodness-of-fit tests.

Regression tests, such as the Shapiro-Wilk test, extend the idea of Q-Q plots and measure the linear relationship between expected and observed values of the order statistics. Unlike the inherently subjective interpretation of Q-Q plots, regression tests offer a more objective decision rule for the evaluation of the null hypothesis. In the case of the Shapiro-Wilk test, the expected values of the order statistics are used as explanatory variables in a linear model to predict the observed values. The test statistic is a function of the regression coefficients and the sample variance. Its distribution under the null hypothesis is specifically calibrated for sample sizes between 5 and 2000.

Normality tests based on moments, such as the Jarque-Bera test, rely on skewness and kurtosis to assess the deviation from normality. The test statistic is derived under an asymptotic condition. In addition to the original test, a robust version of the test has been proposed, where the sample variance is replaced by the average absolute deviation from the median, making it less sensitive to outliers. The constants appearing in the robust version were calibrated using Monte Carlo simulations at a significance level of 5%.

Goodness-of-fit tests are designed to test whether a sample comes from a specified distribution, which is not limited to the normal distribution. These tests operate by transforming the observed data using the assumed null distribution, resulting in a sample that follows a uniform distribution under the null hypothesis. Based on this transformation, a test statistic and the corresponding p-value can be calculated.

The presented goodness-of-fit tests are based on measuring the vertical distances between the empirical distribution function and the theoretical distribution. However, a disadvantage of these methods is their sensitivity to random occurrences, such as outliers or clusters.

The tests can be categorized into two classes.

- Surpremum tests based on the maximum distance between the empirical and cumulative distribution functions, e.g. the Kolmogorov-Smirnov test.
- Quadratic tests based on the squared difference between the distributions, e.g. the Cramér-von Mises test and the Anderson-Darling test, which rely on the choice of the weighting function.

The Cramér-von Mises test uses a constant weighting function, while the Anderson-Darling test applies a weighting function that gives more weight to the tails of the distribution. For these two tests, adjusted versions were presented as well that account for small-sample behavior. The adjusted test statistics and their distributions were calibrated in simulations that included small samples. Additionally, these tests were specifically indented for the testing situation in which the distributional parameters of the assumed normal distribution have to be estimated from the sample.

Power Simulation

As statistical testing is part of statistical inference, performing a hypothesis test intends to make a statement about the underlying distribution of the statistical model. Since only an observed sample is available through which the underlying model can be observed, and maybe some additional limited information, the whole process of decision making is connected with uncertainty. This can result in false decisions about the null hypothesis that is being tested. In a testing environment, where all aspects can be controlled, the probabilities of rejecting the null hypothesis, given that it is true or that it is actually not true, are called the level of significance (α) and resp. the power of test ($1 - \beta$). However, if the sample size cannot be increased, the α-level is usually fixed at a predefined level and the

power of test can be observed in a simulation.

This is the idea behind the simulation that was carried out in this thesis. The results can be used to compare hypothesis tests and identify which tests are more powerful at the specified level of significance, i.e., which tests have a higher rejection rate of samples derived from distributions for which the null hypothesis is not fulfilled.

A power simulation provides a controlled and repeatable experiment to evaluate how well each test performs under various conditions, such as testing samples of small sizes. The simulation also allows for a systematic comparison on a series of alternative distributions that have different features, such as lighter or heavier tails. This can help to identify which tests are sensitive to specific departures from normality and which are more robust.

The results presented in Sections 7.3 and 7.4 indicate that

- the Shaprio-Wilk test,
- the adjusted version of the Anderson-Darling test,
- the adjusted version of the Cramér-von Mises test and
- the robust version of the Jarque-Bera test

are more effective and appropriate in detecting non-normal samples of small sample sizes than

- the Kolmogorov-Smirnov test,
- the original version of the Anderson-Darling test,
- the original version of the Cramér-von Mises test and
- the original version of the Jarque-Bera test.

All the tests that showed a higher power are in some way calibrated or modified for application on finite samples. The reasons why some of the tests show such a low power, as shown in Section 7.3, could be a combination of the following reasons.

- In the settings of the tested distributions, mainly concerning the sample size and alternative distributions, the tests might not be sensitive enough to detect deviations from normality. As mentioned in [11], the Cauchy distribution can be used to measure how sensitive normality tests are in detecting a heavy-tailed alternative. The lower power here, as shown in Figure 7.4, could indicate that the small samples do not provide enough information for these tests to make a reliable decision.
- Observing the distribution of the p-values on normally distributed samples can also give indication whether asymptotic assumptions of the test are satisfied. When tests based on asymptotic assumptions are evaluated and the p-value does not appear

uniformly, the underlying assumption might not be appropriate in the testing environment, such as testing samples of small sizes. Such asymptotic tests that have a lower rejection rate on normally distributed samples than the fixed level of significance α, are also called conservative tests, as described in [9]. Some of the tests presented here show this behavior, as shown in Figure 7.8. This behavior can also result in low power when evaluating the tests on alternatively distributed samples of small sizes. In order to improve the performance of a conservative test, some kind of correction might be necessary, for which the adjusted Anderson-Darling test or the adjusted Cramér-von Mises test are examples. These account for the sample size.

- As mentioned in Chapter 6, when testing a normality assumption with both expectation and standard deviation being unknown, as is the case here, the parameters are estimated from the sample to complete the null hypothesis. Having to estimate these parameters can interfere with the test statistic and influence its theoretical distribution. Simulating the performance of the tests on normally distributed samples and looking at the distribution of the p-values can indicate whether the test statistic still follows the assumed distribution.

Although the robust version of Jarque-Bera is one of the tests that showed a higher power, from Figure 7.8 it is still visible that this test still relies on an asymptotic assumption for the test statistic.

Evaluation on the Data Set

Since the motivation for this thesis came from a real-world application, Chapter 8 discusses the results of the tests when applied to this dataset.

None of the samples are rejected by any of the hypothesis tests at the α-level of 5%. Initially, this means that there is no sufficiently strong evidence speaking against the normality assumption that was detected by the tests, which would justify rejecting the null hypothesis. However, this does not confirm that the samples are actually normally distributed, only that assuming normality is plausible under the testing conditions. However, the results need to be interpreted with caution, as the simulation suggested that the tests have a rather low power on small smaple sizes.

The hypothesis of normality can still be considered reasonable, provided that no other concerns arise, for instance, through other methods, such as graphical evaluations with graphs, as presented in Chapter 3.

In the case of the observed data samples, the plots given in Chapter 3 were either not suitable for such small sizes, such as the histograms, or largely supported the results of the hypothesis tests, i.e. they did not show substantial visual departures from normality or distinct patterns hinting at alternative distributions.

The interpretation of the results from the statistical tests and the graphical methods has

to be done with careful consideration, since multiple points can affect the relevance of the visual representation or the performance of the tests.

- Underlying assumptions, such as asymptotic assumptions when applied to finite samples, might not be satisfied when testing the considered samples. This can affect the testing procedure in such a way that the test statistic of the hypothesis test no longer follows its assumed distribution. Therefore, the outcome of the test might not be reliable.

- Some of the presented methods are not able to test for a normal distribution in general but need specific parameters for the mean and the standard deviation. Therefore, these must be estimated first to be able to test for a specific normal distribution, bringing additional uncertainty into the process. The estimates might not give an appropriate representation of the true location and scale parameters. Furthermore, the step of estimating the distributional parameter as part of the testing procedure might affect the test statistic of a test such that it deviates from its assumed distribution, which might also make the test results unreliable.

- Most importantly, a small sample size can greatly affect how suitable or sensitive the methods are for representing departures from normality. For small samples, graphical methods might show curvature in the tails, even if the null hypothesis is actually true, as mentioned in Section 3.2. For hypothesis tests, the smaller the sample, the less information it provides for the test to detect the departure from normality.

As an important part of the thesis, all test, including the calculation of the test statistic and the p-value, are implemented in R. The codes for simulations, graphical methods and application of the tests for the real-world dataset are given in Appendix A, B and C.

A. Appendix - Code for the Simulation

```
library(dplyr)
library(knitr)
library(extraDistr)
library(MASS)
library(glogis)
library(fitdistrplus)
library(extraDistr)
library(lhs)
library(EnvStats)
library(tibble)
library(dplyr)
library(reshape2)
library(ggplot2)
library(cowplot)
library(nortest)
library(goftest)
library(matrixcalc)
library(qqplotr)
library(gridExtra)

# the variable data_list contains the 12 samples in a list format

#####################################################
##### 1. Fitting of distributions onto the samples
#####    Maximum Likelihood estimation
#####################################################

# fitting student't distributions
loglik <- function(x,par){
  t <- length(x) -1
  sigma <- par[2]
  mu <- par[1]
  if(sigma>0){
    #return(-sum(log(dt((x-mu)/sigma,df=t)/sigma)))
    return(-sum(log(extraDistr::dlst(x,df=t,mu = mu, sigma = sigma))))
    # same results with either one
  } else{
    Inf
  }
}

call_optim <- function(x){
  optim(c(median(x),IQR(x)/2),loglik, x = x)$par
}
loc_scale <- sapply(data_list, call_optim)
est_st_t <- rbind(lengths(data_list) - 1, loc_scale)
```

```
47
48
49  # fitting lognormal distributions
50  call_lognormal <- function(x){
51    fitdistr(x, "lognormal")$estimate
52  }
53  est_lognormal <- sapply(data_list, call_lognormal)
54
55
56  # fitting uniform distributions
57  est_unif <- matrix(0,ncol=12,nrow=2)
58  est_unif[1,] <- sapply(data_list, min)
59  est_unif[2,] <- sapply(data_list, max)
60
61
62  # fitting Weibull distributions
63  loglik_weibull <- function(x, par){
64    n <- length(x)
65    eta <- par[1]
66    sigma <- par[2]
67    if(sigma > 0 & eta > 0){
68      -(n*log(eta) - n*eta*log(sigma) + (eta-1)*sum(log(x)) - (1/sigma^eta)*
       sum(x^eta))
69    } else {
70      Inf
71    }
72  }
73
74  call_optim_weibull <- function(x){
75    optim(par = c(5,1), loglik_weibull, x = x)$par
76  }
77  est_weibull <- sapply(data_list, call_optim_weibull)
78
79
80  # fitting Laplace distributions
81  loglik_laplace <- function(x, par){
82    n <- length(x)
83    mu <- par[1]
84    lambda <- par[2]
85    if(lambda > 0){
86      -(n*log(lambda) - n*log(2) - lambda*sum(abs(x-mu)))
87    } else {
88      Inf
89    }
90  }
91
92  call_optim_laplace <- function(x){
93    optim(par = c(1,1), loglik_laplace, x = x)$par
94  }
95  est_laplace <- sapply(data_list, call_optim_laplace)
96
97
98  # fitting Gumbel distributions
99  loglik_gumble <- function(x, par){
100   n <- length(x)
```

```
  mu <- par[1]
  beta <- par[2]
  if(beta > 0){
    -(-n*log(beta) - sum((x-mu)/beta) - sum(exp(-(x-mu)/beta)) )
  } else {
    Inf
  }
}

call_optim_gumble <- function(x){
  optim(par = c(1,1), loglik_gumble, x = x)$par
}
est_gumbel <- sapply(data_list, call_optim_gumble)


# fitting Cauchy distributions
loglik_cauchy <- function(x,par){
  n <- length(x)
  l <- par[1]
  s <- par[2]
  if(s > 0){
    -(-n*log(pi*s) - sum(log(1 + ((x-l)/s)^2)) )
  } else{
    Inf
  }
}

call_optim_cauchy <- function(x){
  optim(par = c(1,1), loglik_cauchy, x = data_list[[1]])$par
}
est_cauchy <- sapply(data_list, call_optim_cauchy)


# fitting Gamma distributions
loglik_gamma <- function(x,par){
  n <- length(x)
  alpha <- par[1]
  beta <- par[2]
  if(alpha > 0 & beta > 0){
    -(alpha*n*log(beta) - n*log(gamma(alpha)) + (alpha-1)*sum(log(x)) - beta
    *sum(x))
  } else {
    Inf
  }
}

call_optim_gamma <- function(x){
  optim(par = c(1,1), loglik_gamma, x = x)$par
}
est_gamma <- sapply(data_list, call_optim_gamma)


# fitting triangular distrubtions
loglik_tri <- function(x,par){
  n <- length(x)
```

```
155   a <- par[1]
156   b <- par[2]
157   m <- par[3]
158   if(a < m & m < b){
159     -sum(log(dtri(x, min = a, max = b, mode = m)))
160   } else {
161     Inf
162   }
163 }
164
165 call_optim_tri <- function(x){
166   optim(par = c(0,2,1), loglik_tri, x = x)$par
167 }
168 est_triangular <- sapply(data_list, call_optim_tri)
169
170
171 # fitting Normal distributions
172 mean <- sapply(1:12, function(x) sum(data_list[[x]])/length(data_list[[x]]))
173 var <- sapply(1:12, function(x) sum((data_list[[x]] - mean[x])^2)/length(
      data_list[[x]]) )
174 est_normal <- rbind(mean, sqrt(var))
175
176
177
178
179 ########################################################
180 ##### 2. Preparaton for the Power-of-Test Simulation
181 ########################################################
182
183 n <- 5:15
184 number_of_tests <- 8
185 RF <- 12
186 number_of_distributions <- 9
187 number_simulations <- 1000
188 name_of_test <- c("SW", "adj.AD", "AD", "KS", "org.JB", "rob.JB", "CvM", "
      adj.CvM")
189 alpha <- 0.05
190
191 # distribution indication
192 dist <- c( "st_t",         "lognormal",  "unif",    "weibull", "laplace", "
      gumbel", "cauchy", "gamma", "triangular")
193 distr <- c("Student's t", "Log Normal", "Uniform","Weibull", "Laplace", "
      Gumbel", "Cauchy", "Gamma", "Triangular")
194
195 ## generate data: a given uniform sample is transform into a sample from the
       specified distribution (dist_word) by applying the inverse cumulative
      distribution function
196 generate_data2 <- function(dist_word, unif_sample, par){
197
198   stopifnot(dist_word %in% dist)
199
200   if(dist_word == "st_t"){data_ <- extraDistr::qlst(unif_sample,
201                                                     df = length(unif_sample)
      -1, mu = par[2],
202                                                     sigma = par[3])}
```

```
203   if(dist_word == "lognormal"){ data_ <- qlnorm(unif_sample, meanlog = par
        [1],
204                                                 sdlog = par[2])}
205   if(dist_word == "unif"){data_ <- qunif(unif_sample, min = par[1], max =
        par[2])}
206   if(dist_word == "weibull"){data_ <- qweibull(unif_sample, shape = par[1],
207                                                scale = par[2])}
208   if(dist_word == "erlang"){data_ <- qgamma(unif_sample, shape = par[2],
        scale = par[1])}
209   if(dist_word == "logistic"){data_ <- qglogis(unif_sample, location = par
        [1],
210                                                scale = par[2], shape = par
        [3])}
211   if(dist_word == "laplace"){data_ <- extraDistr::qlaplace(unif_sample,
212                                                            mu = par[1],
        sigma = 1/par[2])}
213   if(dist_word == "gumbel"){data_ <- extraDistr::qgumbel(unif_sample, mu =
        par[1],
214                                                          sigma = par[2])}
215   if(dist_word == "cauchy"){data_ <- qcauchy(unif_sample, location = par[1],
216                                              scale = par[2])}
217   if(dist_word == "gamma"){data_ <- qgamma(unif_sample, shape = par[1], rate
        = par[2])}
218   if(dist_word == "triangular"){data_ <- qtri(unif_sample, min = par[1], max
        = par[2],
219                                               mode = par[3])}
220   return(data_)
221 }
222
223
224 ## function definitions for tests
225
226 # adjusted version of the Anderson-Darling test
227 adj.AD.test <- function(x){
228
229   # calculate the test statistic
230   n <- length(x)
231   sort_data <- sort(x)
232   std_data <- (sort_data - mean(sort_data))/sd(sort_data)
233   p_i <- pnorm(std_data)
234   A_sigma <- sum( (2*c(1:n)-1)*(log(p_i) + log(1-p_i[n:1])) )
235   A <- -n-(1/n)*A_sigma
236
237   # p value
238   Z <- A*(1+ 0.75/n + 2.25/n^2)
239
240   p_val <- 0
241   if(Z >= 0.6){
242     p_val <- exp(1.2937 - 5.709*Z + 0.0186*Z^2)
243   } else if (0.34 < Z & Z < 0.6){
244     p_val <- exp(0.9177 - 4.279*Z - 1.38*Z^2)
245   } else if (0.2 < Z & Z < 0.34){
246     p_val <- 1- exp(-8.318 + 42.796*Z - 59.938*Z^2)
247   } else if (Z <= 0.2){
248     p_val <- 1 - exp(-13.436 + 101.14*Z - 223.73*Z^2)
```

```
  }

  result <- data.frame(test.statistic = A, p.value = p_val)
}


# adjusted version of the Cramer-von Mises test
adj.CvM.test <- function(data){
  n <- length(data)
  std_data <- (sort(data)-mean(data))/sd(data)
  sorted_data <- sort(std_data)
  summands <- (2*(1:n)-1)/(2*n) - pnorm(sorted_data)
  Test_val <- 1/(12*n) + sum(summands^2)
  Test_val

  # adjusted test statistic for p value
  Z_adj <- Test_val*(1+0.5/n)

  if(Z_adj < 0.0275){
    p <- 1 - exp(-13.953 + 775.5*Z_adj - 12542.61*Z_adj^2)
  } else if(0.0275 < Z_adj & Z_adj < 0.051){
    p <- 1 - exp(-5.903 + 179.546*Z_adj -1515.29*Z_adj^2)
  } else if(0.051 < Z_adj & Z_adj < 0.092){
    p <- exp(0.886 - 31.62*Z_adj + 10.897*Z_adj^2)
  } else if(Z_adj > 0.092){
    p <- exp(1.111 - 34.242*Z_adj + 12.832*Z_adj^2)
  }

  res <- data.frame(Test_val,p)
  names(res) <- c("statistic", "p.value")

  return(res)
}


# original version of the Jarque-Bera test
JB.test <- function(data){

  n <- length(data)
  m2 <- sum((data-mean(data))^2)/n # second moment
  m3 <- sum((data-mean(data))^3)/n # third moment
  m4 <- sum((data-mean(data))^4)/n # fourth moment

  skewness_ <- m3/(m2)^(3/2) # skewness
  kurtosis_ <- m4/(m2)^2     # kurtosis

  # test statistic
  S <- (n/6)*skewness_^2 + (n/24)*(kurtosis_ - 3)^2

  # p-value
  p_S <- 1 - pchisq(S, df=2)

  return(list(statistic = S, p.value = p_S))
}

```

```r

# robust version of the Jarque-Bera test
robust.JB.test <- function(data){
  C1 <- 6
  C2 <- 64
  C <- sqrt(pi/2)
  M <- median(data)
  n <- length(data)
  Jn <- (C/n)*sum(abs(data-M))

  m3 <- sum((data-mean(data))^3)/n # third moment
  m4 <- sum((data-mean(data))^4)/n # fourth moment

  # value of test statistic
  RJB <- (n/C1)*(m3/Jn^3)^2 + (n/C2)*(m4/Jn^4 - 3)^2

  # p value
  p <- 1 - pchisq(RJB,df=2)

  return(list(statistic = RJB, p.value = p))
}


## calculate p-value of all tests applied to a given data sample
evaluate_tests <- function(data){
  # be aware of possible data pre-processing and additional input parameter!
  p_values <- rep(NA, number_of_tests)
  std_data <- (sort(data) - mean(data))/sd(data)

  # shapiro wilk
  p_values[1] <- shapiro.test(data)$p.value

  # anderson darling, adjusted, own version
  p_values[2] <- adj.AD.test(data)$p.value

  # anderson darling
  p_values[3] <- goftest::ad.test(std_data, null = "pnorm")$p.value

  # kolmogorov smirnov
  p_values[4] <- ks.test(std_data, "pnorm")$p.value

  # Jarque Bera, original version
  p_values[5] <- JB.test(data)$p.value

  # Jarque Bera, robust modification
  p_values[6] <- robust.JB.test(data)$p.value

  # Cramer von Mises
  p_values[7] <- goftest::cvm.test(std_data, "pnorm")$p.value

  # Cramer von Mises,adjusted, own version
  p_values[8] <- adj.CvM.test(data)$p.value

  return(p_values)
}
```

```
359
360
361 # adjusted sample estimates for the skewness and kurtosis values
362
363 adj_skewness <- function(data){
364   len <- length(data)
365   skewness_value <- len/((len-1)*(len-2)) * sum((data-mean(data))^3) / sd(
      data)^3
366
367   return(skewness_value)
368 }
369
370 adj_kurtosis <- function(data){
371   len <- length(data)
372   kurtosis_value <- len*(len+1)/((len-1)*(len-2)*(len-3)) * sum((data-mean(
      data))^4)/sd(data)^4 -
373     3*(len-1)^2/((len-2)*(len-3))
374
375   return(kurtosis_value)
376 }
377
378 ##############################################################################
379 ##### 3. Code for the Power-of-Test Simulation on Alternative Distributions
380 ##############################################################################
381
382 # The random numbers can also be generated by a quasi-random number
      generator. In the end, this option was not used for the simulation.
383
384 random.numbers <- "pseudo"
385 my.seed <- 1904
386
387 stopifnot(random.numbers %in% c("quasi", "pseudo"))
388 set.seed(my.seed)
389 print(my.seed)
390
391 # result object to store the results
392 results <- array(0, dim=c(length(n),
393                           number_of_tests,
394                           RF,
395                           number_of_distributions),
396                  dimnames = list(paste0("n ",n),
397                                  name_of_test,
398                                  paste0("RF",1:12),
399                                  distr))
400
401 store_skewness <- array(NA, dim = c(length(n),
402                                     number_simulations,
403                                     RF,
404                                     number_of_distributions + 1),
405                         dimnames = list(paste0("n ", n),
406                                         paste0("sim",1:number_simulations),
407                                         paste0("RF", 1:12),
408                                         c(distr,"Normal")))
409
410 store_kurtosis <- array(NA, dim = c(length(n),
```

```
                                    number_simulations,
                                    RF,
                                    number_of_distributions + 1),
                          dimnames = list(paste0("n ", n),
                                          paste0("sim",1:number_simulations),
                                          paste0("RF", 1:12),
                                          c(distr, "normal")))

##### simulation
for(n_ in n){
  print(n_) # for progress
  for(d in 1:number_of_distributions){

    for(RF_ in 1:RF){
      # get fitted parameters
      par <- eval(parse(text = paste0("est_", dist[d],"[ ,RF_]")))

      # reset counter (how often the tests are rejecting)
      counter <- rep(0, number_of_tests)

      for(j in 1:number_simulations){

        # generate random numbers: uniform sample
        if(random.numbers == "quasi"){
          unif_sample <- as.numeric(randomLHS(n = n_,k = 1))
        } else if (random.numbers == "pseudo"){
          unif_sample <- runif(n_)
        }

        # transform uniform sample into sample from alternative distribution
        data <- generate_data2(dist[d], unif_sample, par)

        # calculate adjusted sample skewness and kurtosis
        store_skewness[n_-4, j, RF_, d] <- adj_skewness(data)
        store_kurtosis[n_-4, j, RF_, d] <- adj_kurtosis(data)

        # pvalues for generated data for each test
        p_values <- evaluate_tests(data)

        counter <- counter + ifelse(p_values <= alpha, 1, 0)
      }

      # store the simulated power in the result object
      results[n_-4, , RF_,d] <- counter/number_simulations
    }
  }
}

### calculate mean and sd of skewness/kutosis for each plot
# mean and sd over the sizes n and the number_simulation
calculated_mean_skewness <- matrix(NA,nrow = number_of_distributions + 1,
    ncol = RF) # nrow +1 for normal distr.
calculated_sd_skewness <- matrix(NA,nrow = number_of_distributions + 1, ncol
     = RF)
calculated_mean_kurtosis <- matrix(NA,nrow = number_of_distributions + 1,
```

```
    ncol = RF)
calculated_sd_kurtosis <- matrix(NA,nrow = number_of_distributions + 1, ncol
    = RF)
for(d in 1:number_of_distributions){
  for(RF_ in 1:RF){
    calculated_mean_skewness[d,RF_] <- mean(store_skewness[,,RF_,d])
    calculated_sd_skewness[d,RF_] <-     sd(store_skewness[,,RF_,d])
    calculated_mean_kurtosis[d,RF_] <- mean(store_kurtosis[,,RF_,d])
    calculated_sd_kurtosis[d,RF_] <-     sd(store_kurtosis[,,RF_,d])
  }
}

### results per distr and RF, plotted in grid

name_of_test <- c("SW", "adj.AD", "AD", "KS", "org.JB", "rob.JB", "CvM", "
    adj.CvM")
k_ <- 1 # counter to fill the list with the plots
plot_list <- list()

for(d in 1:number_of_distributions){

  plot_list_temp <- list()
  counter2 <- 1
  range.y.max <- 0


  for(RF_ in 1:RF){
    # create all plots at once and later take subset for printing

    tmp_results <- results[ , ,RF_, d]

    # processing
    tmp_results <- tmp_results %>% as_tibble %>%
      setNames(name_of_test) %>% mutate(sample_size = n)

    # plot
    plot_list_temp[[counter2]] <- tmp_results %>% melt(id = c("sample_size")
    )  %>%
      ggplot(aes(x = sample_size, y = value, group = variable,
                 col = factor(variable))) +
      geom_line(size = 0.5) + theme_bw() +
      scale_x_continuous(breaks= n) +
      ggtitle(paste0("S ", RF_, ", ",distr[d])) +
      xlab(" ") +
      ylab(" ") +
      theme(legend.position = "none") +
      theme(text = element_text(size = 6),
            plot.title = element_text(size = 7),
            plot.margin = margin(t = 1, r = 1, b = 0, l = 0, unit = "pt"))

    range.y.max <- pmax(range.y.max,
                        layer_scales(plot_list_temp[[counter2]])$y$range$
    range[2])

    counter2 <- counter2 + 1
```

```
514   }
515
516   # add annotations about kurtosis and skewness values
517   for(counter3 in 1:RF){
518     plot_list[[k_]] <- plot_list_temp[[counter3]] + ylim(0, range.y.max) +
519       annotate("text", x = n[1], y = range.y.max*4.8/5, hjust = 0, size = 2,
520                label = paste0("skew = ",
521                               round(calculated_mean_skewness[d, counter3],1)
    ,
522                               "+/-",
523                               round(calculated_sd_skewness[d, counter3],1)))
    +
524       annotate("text", x = n[1], y = range.y.max*4.2/5, hjust = 0, size = 2,
525                label = paste0("kurt = ",
526                               round(calculated_mean_kurtosis[d, counter3],1)
    ,
527                               "+/-",
528                               round(calculated_sd_kurtosis[d, counter3],1)))
529
530     k_ <- k_ + 1
531   }
532
533 }
534
535
536 ### printing the plots
537
538 # cluster representatives
539 cluster_set_up <- c(FALSE, # 1
540                     FALSE, # 2
541                     FALSE, # 3
542                     TRUE,  # 4
543                     FALSE, # 5
544                     TRUE,  # 6
545                     FALSE, # 7
546                     FALSE, # 8
547                     TRUE,  # 9
548                     FALSE, # 10
549                     TRUE,  # 11
550                     FALSE) # 12
551 n_col <- sum(cluster_set_up)
552
553 # Figure 7.2
554 # group 1: light-tailed, uniform (3), triangular (9)
555 subset_plots <- matrix(FALSE,12,9)
556 subset_plots[,c(3,9)] <- cluster_set_up
557 group1 <- plot_list[as.logical(subset_plots)]
558 plots_group1 <- do.call("grid.arrange", c(group1, ncol = n_col))
559 ggsave("plots_group1.pdf",plots_group1, height = 10/9*2, width = 8/6*n_col)
560
561 # Figure 7.3
562 # group 2: gamma (8), weibull (4), lognormal (2)
563 subset_plots <- matrix(FALSE,12,9)
564 subset_plots[,c(2,4,8)] <- cluster_set_up
565 group2 <- plot_list[as.logical(subset_plots)]
```

```
plots_group2 <- do.call("grid.arrange", c(group2, ncol = n_col))
ggsave("plots_group2.pdf",plots_group2, height = 10/9*3, width = 8/6*n_col)

#Figure 7.4
# group 3: laplace (5), gumbel (6), cauchy (7)
subset_plots <- matrix(FALSE,12,9)
subset_plots[,c(5,6,7)] <- cluster_set_up
group3 <- plot_list[as.logical(subset_plots)]
plots_group3 <- do.call("grid.arrange", c(group3, ncol = n_col))
ggsave("plots_group3.pdf",plots_group3, height = 10/9*3, width = 8/6*n_col)

# Figure 7.5
# group 4: student's t (1)
subset_plots <- matrix(FALSE,12,9)
subset_plots[,1] <- cluster_set_up
group4 <- plot_list[as.logical(subset_plots)]
plots_group4 <- do.call("grid.arrange", c(group4, ncol = n_col))
ggsave("plots_group4.pdf",plots_group4, height = 10/9*1, width = 8/6*n_col)




################################################################################
##### 4. Code for the Power-of-Test Simulation on the Normal Distribution
################################################################################

stopifnot(random.numbers %in% c("quasi", "pseudo"))
set.seed(my.seed)

##### result object to store the results
results_check <- array(0, dim=c(length(n),
                                number_of_tests,
                                RF),
                       dimnames = list(paste0("n ",n),
                                       name_of_test,
                                       paste0("RF",1:12)))

# collect the p-values for one of the RF, e.g. for RF 1
collect_p_value <- matrix(NA, ncol = number_of_tests, nrow = number_
    simulations*length(n))
count_row <- 1

##### simulation on normal data
for(n_ in n){
  print(n_) # for progress
  for(RF_ in 1:RF){
    # collect fitted parameters
    par <- eval(parse(text = paste0("est_normal","[ ,RF_]")))

    # reset counter (how often the tests are rejecting)
    counter <- rep(0, number_of_tests)
    for(j in 1:number_simulations){

      if(random.numbers == "quasi"){
```

```
620         unif_sample <- as.numeric(randomLHS(n = n_,k = 1))
621       } else if(random.numbers == "pseudo"){
622         unif_sample <- runif(n_)
623       }
624       # transform to normal data
625       data <- qnorm(unif_sample,mean = par[1], sd = par[2])
626 
627       # calculate adjusted sample skewness and kurtosis
628       store_skewness[n_-4, j, RF_, number_of_distributions + 1] <- adj_
    skewness(data)
629       store_kurtosis[n_-4, j, RF_, number_of_distributions + 1] <- adj_
    kurtosis(data)
630 
631 
632       # p-values for generated data for each test
633       p_values <- evaluate_tests(data)
634 
635       if(RF_ == 4){
636         collect_p_value[count_row,] <- p_values
637         count_row <- count_row + 1
638       }
639 
640       counter <- counter + ifelse(p_values <= alpha, 1, 0)
641     }
642 
643     # store the simulated power in the result object
644     results_check[n_-4, , RF_] <- counter/number_simulations
645   }
646 }
647 
648 # calculate mean and sd of these skewness and kurtosis values
649 # mean and sd over the sizes n and the number_simulation
650 
651 # filled for normal distributions
652 d <- number_of_distributions+1
653 for(RF_ in 1:RF){
654   calculated_mean_skewness[d,RF_] <- mean(store_skewness[,,RF_,d])
655   calculated_sd_skewness[d,RF_] <-     sd(store_skewness[,,RF_,d])
656   calculated_mean_kurtosis[d,RF_] <- mean(store_kurtosis[,,RF_,d])
657   calculated_sd_kurtosis[d,RF_] <-     sd(store_kurtosis[,,RF_,d])
658 }
659 
660 
661 
662 # printing the output
663 
664 name_of_test <- c("SW", "adj.AD", "AD", "KS", "org.JB", "rob.JB", "CvM", "
    adj.CvM")
665 name_of_test_long <- c("Shapiro-Wilk",
666                        "adj. Anderson-Darling",
667                        "Anderson-Darling",
668                        "Kolmogorov-Smirnov",
669                        "orig. Jarque-Bera",
670                        "robust Jarque-Bera",
671                        "Cramer-von Mises",
```

```
                          "adj. Cramer-von Mises")
k_ <- 1 # counter to fill the list with the plots
plot_list <- list()
counter2 <- 1
range.y.max <- 0


for(RF_ in 1:RF){
  # create all plots at once and later take subset for printing

  tmp_results <- results_check[ , ,RF_]

  # processing
  tmp_results <- tmp_results %>% as_tibble %>%
    setNames(name_of_test) %>% mutate(sample_size = n)

  # plot
  plot_list[[counter2]] <- tmp_results %>% melt(id = c("sample_size"))  %>%
    ggplot(aes(x = sample_size,
               y = value,
               group = variable,
               col = factor(variable))) +
    geom_line(size = 0.5) + theme_bw() +
    scale_x_continuous(breaks= n) +
    ggtitle(paste0("S ", RF_, ", Normal")) +
    xlab(" ") +
    ylab(" ") +
    theme(legend.position = "none") +
    theme(text = element_text(size = 6),
          plot.title = element_text(size = 7),
          plot.margin = margin(t = 1, r = 1, b = 0, l = 0, unit = "pt"))

  range.y.max <- pmax(range.y.max,
                      layer_scales(plot_list[[counter2]])$y$range$range[2])

  counter2 <- counter2 + 1
}

# add annotations about skewness and kurtosis
for(counter3 in 1:RF){
  plot_list[[counter3]] <- plot_list[[counter3]] + ylim(0, range.y.max) +
    annotate("text", x = n[1], y = range.y.max*4.8/5, hjust = 0, size = 2,
             label = paste0("skew = ",
                            round(calculated_mean_skewness[number_of_
    distributions+1, counter3],1),
                            "+/-",
                            round(calculated_sd_skewness[number_of_
    distributions+1, counter3],1))) +
    annotate("text", x = n[1], y = range.y.max*4.2/5, hjust = 0, size = 2,
             label = paste0("kurt = ",
                            round(calculated_mean_kurtosis[number_of_
    distributions+1, counter3],1),
                            "+/-",
                            round(calculated_sd_kurtosis[number_of_
    distributions+1, counter3],1)))
```

```
723
724 }
725
726 # Figure 7.7
727 # group 5: normal distribution
728
729 plots_group5 <- do.call("grid.arrange", c(group5, ncol = n_col))
730 ggsave("plots_group5.pdf",plots_group5, height = 10/9*1, width = 8/6*n_col)
731
732
733 # Figure 7.8
734 # histograms for the p_values
735
736 plot_list_hist <- list()
737
738 for(j in 1:number_of_tests){
739   data_temp <- data.frame(x = collect_p_value[,j])
740   plot_list_hist[[j]] <- ggplot(data = data_temp, aes(x=x)) +
741     geom_histogram(color = "blue", fill = "white", breaks = seq(0,1,0.025))
      +
742     theme_bw() +
743     ggtitle(name_of_test_long[j]) +
744     ylab("") +
745     xlab("") +
746     theme(plot.title = element_text(size = 9)) +
747     geom_vline(aes(xintercept = 0.05), linetype = "dashed")
748
749   if(j %in% c(1,5)){
750     plot_list_hist[[j]] <- plot_list_hist[[j]] + ylab("count")
751   }
752   if(j %in% 5:8){
753     plot_list_hist[[j]] <- plot_list_hist[[j]] + xlab("p-value")
754   }
755 }
756
757 p_value_histogram_normal <- do.call("grid.arrange", c(plot_list_hist, ncol =
       4))
758 ggsave("p_values_histogram_normal.pdf", p_value_histogram_normal, height =
     4, width = 8)
759
760
761
762 ############################################################
763 ##### 4. Evaluation of Tests on the real-world Data Set
764 ############################################################
765
766 # The following code also contains a process that collects the values of the
       test statistics and the p-values and generates a plot for the p-values,
       which is included in the thesis. The values of the test statistics are
     not shown in this master thesis.
767
768 ##### function to calculate the values of the test statistics
769 evaluate_test_statistics <- function(data){
770   # be aware of possible data pre-processing and additional input parameter!
771   statistics <- rep(NA, number_of_tests)
```

```
772   std_data <- (sort(data) - mean(data))/sd(data)
773
774   # shapiro wilk
775   statistics[1] <- shapiro.test(data)$statistic
776
777   # anderson darling, adjusted, own version
778   statistics[2] <- adj.AD.test(data)$test.statistic
779
780   # anderson darling
781   statistics[3] <- goftest::ad.test(std_data, null = "pnorm")$statistic
782
783   # kolmogorov smirnov
784   statistics[4] <- ks.test(std_data, "pnorm")$statistic
785
786   # Jarque Bera, original version
787   statistics[5] <- robust.JB.test(data)$statistic
788
789   # Jarque Bera, robust modification
790   statistics[6] <- JB.test(data)$statistic
791
792   # Cramer von Mises
793   statistics[7] <- goftest::cvm.test(std_data, "pnorm")$statistic
794
795   # Cramer von Mises,adjusted, own version
796   statistics[8] <- adj.CvM.test(data)$statistic
797
798   return(statistics)
799 }
800
801
802 ##### collect the values of the test statistics
803 stats_res <- matrix(0, ncol = number_of_tests, nrow = RF)
804
805 for(RF_ in 1:RF){
806   statistics <- evaluate_test_statistics(data_list[[RF_]])
807   stats_res[RF_,] <- round(statistics,4)
808 }
809 test_stats_res <- stats_res %>% as_tibble %>% setNames(name_of_test)
810 tmp_test_stats_res <- test_stats_res %>% mutate(data = data_names) %>%
      relocate(data)
811 kable(tmp_test_stats_res, caption = "Value of test statistics of all tests
      and risk factor samples")
812
813
814 ##### collect the p values
815 pval_res <- matrix(0, ncol = number_of_tests, nrow = RF)
816
817 for(RF_ in 1:RF){
818   p_values <- evaluate_tests(data_list[[RF_]])
819   pval_res[RF_,] <- round(p_values,4)
820 }
821 test_p_res <- pval_res %>% as_tibble %>% setNames(name_of_test)
822 tmp_test_p_res <- test_p_res %>% mutate(data = data_names) %>% relocate(data
      )
823 kable(tmp_test_p_res, caption = "P-value of all tests and risk factor
```

```
    samples")

############ visualizing the p-values

##### plot p-values for all tests and all risk factor samples
plot_data <- test_p_res %>% mutate(sample = 1:12) %>% melt(id = "sample")
evaluate_p_1 <- plot_data %>%
  ggplot(aes(x = sample, y = value, group =  variable, col = factor(variable
    )),size = 1.2) +
  geom_point() + theme_bw() + ylab("p value") + xlab("sample") +
  scale_x_continuous(breaks = 1:12) + labs(color = "") + ylim(0,1) +
  geom_hline(yintercept = 0.05)

##### only plot p-value for the tests: adj.AD, adj.CvM, robust.JB, SW
plot_data3 <- test_p_res %>% mutate(sample = 1:12) %>% melt(id = "sample")

# set value regarding other tests to high value and limit the range on y
    axis.
plot_data3$value[plot_data$variable %in% c("AD", "CvM", "org.JB", "KS")] <-
    2
evaluate_p_3 <- plot_data3 %>%
  ggplot(aes(x = sample, y = value, group =  variable, col = factor(variable
    ))) +
  geom_point() + theme_bw() + ylab("p-value") + xlab("sample") +
  scale_x_continuous(breaks = 1:12) + labs(color = "") + ylim(0,1) +
  geom_hline(yintercept = 0.05) +
  scale_color_discrete(labels = c("SW", "adj.AD", "", "", "", "rob.JB", "",
    "adj.CvM"))

plot_grid(evaluate_p_1, evaluate_p_3, nrow = 1, ncol = 2)
ggsave("evaluate_RF_p_values.pdf", width = 8, height = 2.5)


# Figure 7.1
### extract the legend

get_legend_plot <- evaluate_p_3 +
  scale_color_discrete(labels = c("Shapiro-Wilk (SW)",
                                  "Anderson-Darling, adjusted (adj.AD)",
                                  "Anderson-Darling (AD)",
                                  "Kolmogorv-Smirnov (KS)",
                                  "Jarque-Bera (orig.JB)",
                                  "Jarque-Bera, robust (rob.JB)",
                                  "Cramer-von Mises (CvM)",
                                  "Cramer-von Mises, adjusted (adj.CvM)"))

my_leg <- get_legend(get_legend_plot + labs(color = "") +
                       theme(legend.text = element_text(size = 7),
                             legend.title = element_blank()) +
                       guides(color = guide_legend(ncol = 1)))

just_legend <- plot_grid(my_leg, ncol = 1) + theme(plot.margin = unit(c(0,
    0, 0, 0), "cm"))

ggsave(plot = just_legend, file = "normal_tests_legend.pdf", width = 2.1,
```

```
    height = 2)


############################
###### 5. Additional plots
############################

# Figure 6.1

set.seed(1)
n <- 100
p_AD <- c()
p_AD_estimate <- c()

for(j in 1:10000){

  data <- rnorm(n,0,1)
  std_data <- std_data <- (sort(data) - mean(data))/sd(data)
  p_AD_estimate[j] <- goftest::ad.test(std_data, null = "pnorm")$p.value
  p_AD[j] <- goftest::ad.test(data, null = "pnorm")$p.value
  #p_AD[j] <- adj.AD.test(data)$p.value

}

plot_data <- data.frame(value = c(p_AD_estimate, p_AD),
                        group = rep(c("unknown parameters", "known
    parameters"),c(length(p_AD), length(p_AD_estimate))))

ggplot(plot_data, aes(x = value, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 30) +
  labs(title = " ", x = "p-value", y = "Count") +
  theme_minimal()

ggsave("p_value_not_uniform.pdf", width = 7, height = 4)



# Figure 7.6

library(ggplot2)
library(tidyr)


extract_skewness <- t(calculated_mean_skewness) %>% as_tibble() %>% setNames
    (c(distr, "Normal")) %>%
  mutate(row_number = 1:RF) %>%
  melt("row_number") %>%
  setNames(c("row_number","distribution", "skewness"))
extract_kurtosis <- t(calculated_mean_kurtosis) %>% as_tibble() %>% setNames
    (c(distr, "Normal")) %>%
  mutate(row_number = 1:RF) %>%
  melt("row_number") %>%
  setNames(c("row_number","distribution", "kurtosis"))

skew_kurt_all_data <- left_join(extract_skewness, extract_kurtosis, by = c("
```

```
    row_number", "distribution"))

skew_kurt_all_plot <- ggplot(skew_kurt_all_data,
                             mapping = aes(x = skewness,
                                           y = kurtosis,
                                           group = distribution,
                                           col = distribution,
                                           shape = distribution)) +
  scale_shape_manual(values = 0:(number_of_distributions)) +
  geom_point(size = 3 ) +
  theme_bw() +
  theme(legend.position = "none")



extract_skewness_short <- t(calculated_mean_skewness) %>% as_tibble() %>%
    setNames(c(distr, "Normal")) %>%
  mutate(row_number = 1:RF) %>%
  filter(cluster_set_up) %>%
  melt("row_number") %>%
  setNames(c("row_number","distribution", "skewness"))
extract_kurtosis_short <- t(calculated_mean_kurtosis) %>% as_tibble() %>%
    setNames(c(distr, "Normal")) %>%
  mutate(row_number = 1:RF) %>%
  filter(cluster_set_up) %>%
  melt("row_number") %>%
  setNames(c("row_number","distribution", "kurtosis"))

skew_kurt_short_data <- left_join(extract_skewness_short,
                                  extract_kurtosis_short, by = c("row_number
    ", "distribution"))

skew_kurt_short_plot <- ggplot(skew_kurt_short_data,
                               mapping = aes(x = skewness,
                                             y = kurtosis,
                                             group = distribution,
                                             col = distribution,
                                             shape = distribution)) +
  scale_shape_manual(values = 0:(number_of_distributions)) +
  geom_point(size = 3 ) +
  theme_bw()


plot_grid(skew_kurt_short_plot, skew_kurt_all_plot, nrow = 1, ncol = 2)
ggsave("skew_kurt_together.pdf", width = 8, height = 2.5)
```

B. Appendix - Code for the Graphical Methods

```
######################################################################
##### Empirical distribution function with fitted normal density
######################################################################

# Figure 3.1

plot_list <- vector(mode = "list", length = data_len)

for(i in 1:data_len) {
  RF_name <- paste0("S", i)

  df <- data.frame(RF = data_list[[i]])
  p_ecdf <- ggplot(df, aes(RF)) + stat_ecdf(geom = "step")

  range_ecdf <- layer_scales(p_ecdf)$x$range$range
  adj_range <- range_ecdf + rep(diff(range_ecdf) * 0.1, 2) * c(-1, 1)
  df_null <- data.frame(x_ = seq(adj_range[1], adj_range[2], 0.0001),
                        y_ = pnorm(seq(adj_range[1], adj_range[2], 0.0001),
                                   mean = mean(df[, 1]), sd = sd(df[, 1]) ))
  p <- p_ecdf + geom_line(data = df_null, aes(x = x_,y = y_,color = 3)) +
    theme_bw() + ggtitle(data_names[i]) + xlab("") + ylab("") +
    theme(legend.position = "none") +
    theme( plot.margin = margin(0.1,0.3,0.1,0,"cm"),
           plot.title = element_text(size = 9),
           axis.text.y = element_blank(),
           axis.text.x = element_blank())

  plot_list[[i]] <- p
}

plots <- do.call("grid.arrange", c(plot_list, ncol = 4, nrow = 3))
ggsave("01_ecdf_plots.pdf", plots, width = 10/1.2, height = 7/1.2)


###############
#### Histogram
###############

# Figure 3.3

# Struges: number of bins = ceiling(log2(n) + 1)
h <- vector(mode = "list", length = 12)
for(i in 1:12){
  df <- data.frame(RF = data_list[[i]])
```

```
  p_hist <- ggplot(df, aes(x = RF)) +
    geom_histogram(bins = ceiling(log2(length(data_list[[i]]))+1),
                   color="black", fill="white") + ggtitle(data_names[i]) +
    theme(text = element_text(size = 8),
          plot.title = element_text(size = 9),
          axis.text.x = element_blank()) +
    xlab(" ") + ylab(" ")

  h[[i]] <- p_hist
}

plots <- do.call("grid.arrange", c(h, ncol = 4, nrow = 3))
ggsave("02_hist_RF_sturges.pdf", plots, width = 10/1.2, height = 5.5)

# Figure 3.2

# number of bins = ceiling(sqt(n))
h <- vector(mode = "list", length = 12)
for(i in 1:12){
  df <- data.frame(RF = data_list[[i]])
  p_hist <- ggplot(df, aes(x = RF)) +
    geom_histogram(bins = ceiling(sqrt(length(data_list[[i]]))),
                   color="black", fill="white") + ggtitle(data_names[i]) +
    theme(text = element_text(size = 8),
          plot.title = element_text(size = 9),
          axis.text.x = element_blank()) +
    xlab(" ") + ylab(" ")
  h[[i]] <- p_hist
}

plots <- do.call("grid.arrange", c(h, ncol = 4, nrow = 3))
ggsave("03_hist_RF_sqrt.pdf", plots, width = 10/1.2, height = 5.5)



##############
##### Boxplot
##############

# Figure 3.5

# for the Data Set
df <- data.frame(x_ = unlist(data_list),
                 RF_class = rep(paste0("S", 1:12),lengths(data_list)))
ggplot(df, aes(x = reorder(RF_class,rep(1:12, lengths(data_list))), y = x_))
    +
  geom_boxplot() +
  xlab("values") +
  stat_boxplot(coef = 1.5, geom = "errorbar", width = 0.2) + ylab("Values")
   +
  theme_bw() +
  theme(axis.text.y = element_blank()) +
  xlab("Sample") +
  geom_jitter(width = 0.2, shape = 1, color = "blue")

```

```r
ggsave("04_boxplot_RF.pdf", width = 7, height = 5)

# Figure 3.4

# for quasi random samples to show distinct pattern of alternative
    distributions
set.seed(1)
n <- 10
dist <- c("Normal", "Normal by pseudo rng", "Uniform", "Exponential", "
    Lognormal", "Student's t")
p <- list()
sample <- randomLHS(n = n, k = 1)

dist_data <- c("qnorm(sample, mean = 0, sd = 1)",
               "rnorm(n, mean = 0, sd = 1)",
               "sample",
               "qexp(sample, rate = 1)",
               "qlnorm(sample, meanlog = 0, sdlog = 1)",
               "qt(sample, df = 2)")
for(i in 1:6){
  df <- data.frame(y_ = eval(parse(text = dist_data[[i]])), class = rep(1,n)
    )

  p[[i]] <- ggplot(df, aes(x = class, y = y_)) + geom_boxplot() + xlab(" ")
    +
    stat_boxplot(coef = 1.5, geom = "errorbar", width = 0.2) + ylab(" ") +
    theme_bw() +
    theme(axis.text.x = element_blank(),
          axis.ticks.x = element_blank()) +
    ggtitle(dist[i]) +
    geom_jitter(width = 0.1, shape = 1, color = "blue")
}
plots <- do.call("grid.arrange", c(p, ncol = 3, nrow = 2))
ggsave("05_example_boxplots.pdf", plots, width = 7, height = 5)


##############
##### QQ Plot
##############

# not used

# for the Data Set
p <- list()
for(i in 1:12){
  df <- data.frame(y_ = rev(data_list[[i]]))
  p[[i]] <- ggplot(df, aes(sample = y_)) + ggplot2::stat_qq() + ggplot2::
    stat_qq_line() +
    ggtitle(data_names[i]) +
    xlab(" ") + ylab(" ") +
    theme_bw() +
    theme(plot.title = element_text(size = 8),
          axis.title=element_text(size=6),
          axis.text=element_text(size=6),
          axis.text.y = element_blank())
```

```
}
plots <- do.call("grid.arrange", c(p, ncol = 4))
ggsave("07_sample_qq.pdf", plots, width = 7, height = 5)

# Figure 3.6

# for quasi random samples to show distinct pattern of alternative
    distributions
set.seed(1)
n <- 50
n2 <- 10
dist <- c("Normal", "Normal, pseudo-rng", "Uniform", "Exponential", "
    Lognormal",
          "Student's t, df = 5")
p <- list()
sample <- randomLHS(n = n, k = 1)
dist_data <- c("qnorm(sample, mean = 0, sd = 1)", "rnorm(n2, mean = 0, sd =
    1)",
               "sample", "qexp(sample, rate = 1)",
               "qlnorm(sample, meanlog = 0, sdlog = 1)", "qt(sample, df = 5)
    ")

for(i in 1:6){
  df <- data.frame(y_ = eval(parse(text = dist_data[[i]])))
  p[[i]] <- ggplot(df, aes(sample = y_)) +
    ggplot2::stat_qq() +
    ggplot2::stat_qq_line() +
    ggtitle(dist[i]) + xlab("expected") + ylab("observed") + theme_bw() +
    theme(plot.title = element_text(size = 9),
          axis.text = element_text(size = 8),
          axis.title = element_text(size = 8))
}



plots <- do.call("grid.arrange", c(p, ncol = 3))
ggsave("06_example_qq_plot.pdf", plots, width = 6, height = 4)


#####################################
##### Detrended QQ Plots and QQ Plots
#####################################

# These (detrended) QQ plots contain different variations of a fitted line.


plotting_points <- function(n){
  if(n <= 10){
    p <- (1:n - 0.375)/(n + 1/4)
  } else {
    p <- (1:n - 0.5)/n
  }
  return(p)
}

```

```
199
200 ## for quasi random samples to show distinct pattern of alternative
        distributions
201 set.seed(1)
202 n <- 50
203 n2 <- 10
204 dist <- c("Normal", "Normal by pseudo rng", "Uniform", "Exponential", "
        Lognormal", "Student's t, df = 5")
205 p <- list()
206 sample <- as.numeric(randomLHS(n = n, k = 1))
207
208 dist_data <- c("qnorm(sample, mean = 0, sd = 1)",
209                "rnorm(n2, mean = 0, sd = 1)",
210                "sample",
211                "qexp(sample, rate = 1)",
212                "qlnorm(sample, meanlog = 0, sdlog = 1)",
213                "qt(sample, df = 5)")
214 h <- list()
215
216 for(i in 1:6){
217
218   y_sample <- sort(eval(parse(text = dist_data[[i]])))
219   n <- length(y_sample)
220
221   pi <- (1:n - 0.5)/n
222   x_coord <- qnorm(pi)  # phi_pi
223
224   df <- data.frame(x_ = x_coord, y_ = y_sample)
225
226   # line 1: quantiles 0.25, 0.75
227   xq <- qnorm(c(0.25,0.75)) #### new
228   yq <- quantile(y_sample, c(0.25, 0.75))
229   # f(xp[1]) = a + b*xp[1] = yp[1]
230   # f(xp[2]) = a + b*xp[2] = yp[2]
231   # 1: b*(xp[1]-xp[2]) = yp[1]-yp[2]        b = (yp[1] - yp[2])/(xp[1] - xp
        [2])
232   # 2: a = yp[1] - b*xp[1]
233
234   b <- diff(yq)/diff(xq)
235   a <- yq[1] - b*xq[1]
236   line1 <- a + x_coord*b
237   # add line by: + geom_abline(slope = b, intercept = a)
238
239
240   # line 2: linear regression
241   sigma_hat <- sum((x_coord - mean(x_coord))*y_sample) / sum((x_coord - mean
        (x_coord))^2)
242   mu_hat <- mean(y_sample) - mean(x_coord)*sigma_hat
243   line2 <- mu_hat + x_coord*sigma_hat
244   # add line by: + geom_abline(slope = sigma_hat, intercept = mu_hat)
245
246
247   # line 3: sample estimates
248   sd_hat <- sd(y_sample)
249   mean_hat <- mean(y_sample)
```

```
  line3 <- mean_hat + x_coord*sd_hat
  # add line by: + geom_abline(intercept = mu_hat, slope = sd_hat)


  df1 <- data.frame(x_ = x_coord, y_ = y_sample - line1)
  df2 <- data.frame(x_ = x_coord, y_ = y_sample - line2)
  df3 <- data.frame(x_ = x_coord, y_ = y_sample - line3)

  p_de <- ggplot(df2, aes(x = x_, y = y_)) +
    geom_point(color = "green") + geom_hline(yintercept = 0) +
    geom_point(data = df3, mapping = aes(x = x_, y = y_), color = "blue") +
    geom_point(data = df1, mapping = aes(x = x_, y = y_), color = "red") +
    xlab("theoretical") + ylab("expected") + ggtitle(dist[i])
  # set ranges on x and y axis
  x_ran <- layer_scales(p_de)$x$range$range
  y_ran <- layer_scales(p_de)$y$range$range

  h[[i]] <- p_de + ylim(y_ran + diff(y_ran)*0.7*c(-1,1)) +
    theme(text = element_text(size = 8)) + theme_bw()
}

# Figure 3.10

plots <- do.call("grid.arrange", c(h, ncol = 3))
ggsave("08_example_detrended_qq_new.pdf", plots, width = 7, height = 5)




## for the Data Set
h <- list()  # for detrended qq plot
qq <- list() # for normal qq

for(i in 1: 12){

  y_sample <- sort(data_list[[i]])
  n <- length(data_list[[i]])

  #pi <- (1:n - 0.5)/n
  pi <- plotting_points(n)
  x_coord <- qnorm(pi)

  df <- data.frame(x_ = x_coord, y_ = y_sample)

  # line 1: quantiles 0.25, 0.75
  xq <- qnorm(c(0.25, 0.75))   ##### new
  yq <- quantile(y_sample, c(0.25, 0.75))
  # f(xp[1]) = a + b*xp[1] = yp[1]
  # f(xp[2]) = a + b*xp[2] = yp[2]
  # 1: b*(xp[1]-xp[2]) = yp[1]-yp[2]        b = (yp[1] - yp[2])/(xp[1] - xp
    [2])
  # 2: a = yp[1] - b*xp[1]

  b <- diff(yq)/diff(xq)
  a <- yq[1] - b*xq[1]
```

```
  line1 <- a + x_coord*b
  # add line by: + geom_abline(slope = b, intercept = a)


  # line 2: linear regression
  sigma_hat <- sum((x_coord - mean(x_coord))*y_sample) / sum((x_coord - mean
    (x_coord))^2)
  mu_hat <- mean(y_sample) - mean(x_coord)*sigma_hat
  line2 <- mu_hat + x_coord*sigma_hat
  # add line by: + geom_abline(slope = sigma_hat, intercept = mu_hat)


  # line 3: sample estimates
  sd_hat <- sd(y_sample)
  mean_hat <- mean(y_sample)
  line3 <- mean_hat + x_coord*sd_hat
  # add line by: + geom_abline(intercept = mu_hat, slope = sd_hat)

  p_qq <- ggplot(df, aes(x = x_, y = y_)) + geom_point(size = 2.5) +
    geom_abline(intercept = a, slope = b, color = "red", linewidth = 1) +
    geom_abline(intercept = mu_hat, slope = sigma_hat, color = "green",
                linetype = "dashed", linewidth =1)+
    geom_abline(intercept = mean_hat, slope = sd_hat, color = "blue",
                linewidth = 1, linetype = "dotted") +
    xlab("theoretical") + ylab("expected") + ggtitle(paste0("S",i)) +
    geom_point(data = df, mapping= aes(x = x_, y = y_)) + theme_bw()

  y_ran_ <- layer_scales(p_qq)$y$range$range
  qq[[i]] <- p_qq + ylim(y_ran_ + diff(y_ran_)*0.2*c(-1,1)) +
    theme(text = element_text(size = 8),
          axis.text.y + element_blank())

  df1 <- data.frame(x_ = x_coord, y_ = y_sample - line1)
  df2 <- data.frame(x_ = x_coord, y_ = y_sample - line2)
  df3 <- data.frame(x_ = x_coord, y_ = y_sample - line3)

  p_de <- ggplot(df2, aes(x = x_, y = y_)) + geom_point(color = "green") +
    geom_hline(yintercept = 0) +
    geom_point(data = df3, mapping = aes(x = x_, y = y_), color = "blue") +
    geom_point(data = df1, mapping = aes(x = x_, y = y_), color = "red") +
    xlab("theoretical") + ylab("expected") + ggtitle(paste0("S",i)) + theme_
    bw()
  # set ranges on x and y axis
  x_ran <- layer_scales(p_de)$x$range$range
  y_ran <- layer_scales(p_de)$y$range$range

  h[[i]] <- p_de + #ylim(y_ran + diff(y_ran)*0.5*c(-1,1)) +
    theme(text = element_text(size = 8))
}

# Figur 3.11

plots1 <- do.call("grid.arrange", c(h, ncol = 4))
ggsave("09_detrended_qq_RF_new.pdf", plots1, width = 7, height = 5)

```

```
# Figure 3.7

plots <- do.call("grid.arrange", c(qq, ncol = 4))
ggsave("10_qq_RF_new.pdf", plots, width = 7, height = 5)



##############
##### PP Plot
##############

## for quasi random samples to show distinct pattern of alternative
    distributions

set.seed(1)
n <- 50
n2 <- 10
dist <- c("Normal", "Normal, pseudo-random", "Uniform", "Exponential", "
    Lognormal",
          "Student's t, df = 5")p <- list()
sample <- as.numeric(randomLHS(n = n, k = 1))
dist_data <- c("qnorm(sample, mean = 0, sd = 1)", "rnorm(n2, mean = 0, sd =
    1)",
               "sample", "qexp(sample, rate = 1)",
               "qlnorm(sample, meanlog = 0, sdlog = 1)", "qt(sample, df = 5)
    ")

for(i in 1:6){
  df <- data.frame(y_ = eval(parse(text = dist_data[[i]])))
  p[[i]] <- ggplot(df, aes(sample = y_)) +
    stat_pp_line() +
    stat_pp_point() +
    ggtitle(dist[i]) +
    xlab("expected") + ylab("observed")  + theme_bw() +
    theme(plot.title = element_text(size = 7),
          axis.title=element_text(size=6),
          axis.text=element_text(size=6))

}

# Figure 3.8

plots <- do.call("grid.arrange", c(p, ncol = 3))
ggsave("11_example_pp.pdf", plots, width = 7/1.2, height = 4.5/1.2)


## for the Data Set

p <- list()
for(i in 1:12){
  df <- data.frame(y_ = data_list[[i]])
  p[[i]] <- ggplot(df, aes(sample = y_)) + stat_pp_line() + stat_pp_point()
    +
    ggtitle(data_names[i]) +
    xlab("") + ylab("") + theme_bw() +
```

```
    theme(plot.title = element_text(size = 8),
          axis.title=element_text(size=6),
          axis.text=element_text(size=6),
          axis.text.y = element_blank())
}

# Figure 3.9

plots <- do.call("grid.arrange", c(p, ncol = 4))
ggsave("12_pp_RF.pdf", plots, width = 7, height = 5)
```

C. Appendix - Code for the Tests

```
########################
##### Shapiro Wilk Test
########################

# Results compared with the function *stats::shapiro.test* were identical up to machine precision.

shapiro_wilk <- function(x){

  n <- length(x)

  ##### calculate the value of the test statistic:
  # 1 define values m1, ..,mn as quantiles from the standard normal distribution
  i <.- 1:n
  mi <- qnorm((i-0.375)/(n+0.25))
  # squared euklidean norm = sum of squares:
  m <- sum(mi^2)

  # 2. calculate the a values
  u <- 1/sqrt(n)
  a <- rep(NA,n) # construct empty vector and fill it

  # coefficient for a_n
  a_n_coef <- c(-2.706056, 4.434685, -2.071190, -0.147981, 0.221157)
  # coefficient for a_{n-1}
  a_n_1_coef <- c(-3.582633, 5.682633, -1.752461, -0.293762, 0.042981)

  a[n] <- -2.706056*u^5 + 4.434685*u^4 - 2.071190*u^3 - 0.147981*u^2 + 0.221157*u +
    mi[n]*m^(-1/2)

  if(n > 5){

    a[n-1] <- -3.582633*u^5 + 5.682633*u^4 - 1.752461*u^3 - 0.293762*u^2 + 0.042981*u +
      mi[n-1]*m^(-1/2)

    eps_n <- (m - 2*mi[n]^2 - 2*mi[n-1]^2)/(1 - 2*a[n]^2 - 2*a[n-1]^2)
    a[3:(n-2)] <- mi[3:(n-2)]/sqrt(eps_n)

    a[2] <- -a[n-1]

  } else if (n <= 5){

    eps_n <- (m - 2*mi[n]^2)/(1 - 2*a[n]^2)
```

```
43      a[2:(n-1)] <- mi[2:(n-1)]/sqrt(eps_n)
44
45    }
46
47    a[1] <- -a[n]
48
49    # sort the data in ascending order y_1 <= ... <= y_n
50    sorted_data <- sort(x)
51    # 3 calculate the test statistic:
52    W <- (sum(a*sorted_data))^2/sum((sorted_data-mean(sorted_data))^2)
53    W
54
55    #### p-value
56    if(n >= 12){
57      mu <- 0.0038915*(log(n)^3) - 0.083751*(log(n)^2) - 0.31082*(log(n)) -
      1.5861
58      sigma <- exp(0.0030302*(log(n)^2) - 0.082676*log(n) - 0.4803)
59
60      z <- (log(1-W)-mu)/sigma
61      p <- 1-pnorm(z)    # Z is approximated with the standard normal
      distribution
62    }
63
64    if(n < 12){
65      gamma <- -2.273 + 0.459*n
66      mu <- 0.5440 - 0.39978*n + 0.025054*n^2 - 0.0006714*n^3
67      sigma <- exp(1.3822 - 0.77857*n + 0.062767*n^2 - 0.0020322*n^3)
68
69      g <- -log(gamma-log(1-W))
70      z <- (g-mu)/sigma
71
72      p <- 1-pnorm(z)
73    }
74
75    result <- list(W,p)
76    names(result) <- c("statistic","p.value")
77
78    return(result)
79
80  }
81
82
83  ########################
84  ##### Jarque-Bera Test
85  ########################
86
87  ## original version
88  # Results compared with the R function *DescTools::JarqueBeraTest* (
      parameter robust = FALSE) were identical up to machine precision.
89
90  JB.test <- function(data){
91
92    n <- length(data)
93    m2 <- sum((data-mean(data))^2)/n # second moment
94    m3 <- sum((data-mean(data))^3)/n # third moment
```

```
  m4 <- sum((data-mean(data))^4)/n # fourth moment

  skewness_ <- m3/(m2)^(3/2) # skewness
  kurtosis_ <- m4/(m2)^2     # kurtosis

  # test statistic
  S <- (n/6)*skewness_^2 + (n/24)*(kurtosis_ - 3)^2

  # p-value
  p_S <- 1 - pchisq(S, df=2)

  return(list(statistic = S, p.value = p_S))
}


#######################################
##### robust version, Jarque-Bera Test
#######################################

# Results compared with the R function *DescTools::JarqueBeraTest* (
    parameter robust = TRUE) were identical up to machine precision.

robust.JB.test <- function(data){
  C1 <- 6
  C2 <- 64
  C <- sqrt(pi/2)
  M <- median(data)
  n <- length(data)
  Jn <- (C/n)*sum(abs(data-M))

  m3 <- sum((data-mean(data))^3)/n # third moment
  m4 <- sum((data-mean(data))^4)/n # fourth moment

  # value of test statistic
  RJB <- (n/C1)*(m3/Jn^3)^2 + (n/C2)*(m4/Jn^4 - 3)^2

  # p value
  p <- 1 - pchisq(RJB,df=2)

  return(list(statistic = RJB, p.value = p))
}


#############################
##### Kolmogorov-Smirnov Test
#############################

# Results compared with the R function *stats::ks.test*  were identical up
    to machine precision.

my.KS.test <- function(x){

  # standardize data x
  std_data <- (sort(x) - mean(x))/sd(x)

```

```r
  # value of test statistic
  y_data <- pnorm(std_data)
  n <- length(y_data)
  Dn <- max(y_data - (0:(n-1))/n, (1:n)/n - y_data)

  ##### p value
  # step 1: create matrix
  k <- floor(n*Dn) + 1
  m <- 2*k -1
  # h is determined such that: D = (k-h)/n, where k is positive integer and
    0 <= h < 1
  h <- k - Dn*n

  # construct matirx
  mat <- matrix(0, ncol=m, nrow=m)
  index <- col(mat) - row(mat) == 1
  mat[index] <- 1
  faktorial <- 1
  for(i in 1:m){
    faktorial <- faktorial*i
    index <- row(mat) - col(mat) == (i-1)
    mat[index] <- 1/faktorial
  }

  tmp <- ( 1 - h^(1:m) )
  mat[,1] <- tmp*mat[,1]
  mat[m,] <- rev(tmp)*mat[m,]
  mat[m,1] <- (1 - 2*h^m + (max(0, 2*h - 1))^m)/prod(1:m)

  # calculate power of matrix
  mat.power.n <- matrixcalc::matrix.power(mat,n)

  # p-value
  p <- 1 - (prod(1:n))/(n^(n)) * mat.power.n[k,k]

  res <- data.frame(Dn,p)
  names(res) <- c("statistic", "p.value")

  return(res)
}



############################
##### Cramer-von Mises Test
############################


# The calculation of the $p$-value by various functions in R, e.g. the *
    goftest::cvm.test*, is based on approximating the cumulative
    distribution by its first order expansion. The calculation of the $p$-
    value was not implemented, because the code is quite extensive and
    includes several functions. However, the code is available in the GitHub
     repository https://github.com/baddstats/goftest/blob/master/R/cramer.R

```

```
197
198
199 ####################################
200 ##### Adjusted Cramer-von Mises Test
201 ####################################
202
203 # Results compared with the R function *nortest::cvm.test* were identical up
      to machine precision, but is only applicable for samples with size 8 or
      more.
204
205 adj.CvM <- function(data){
206   n <- length(data)
207   std_data <- (sort(data)-mean(data))/sd(data)
208   sorted_data <- sort(std_data)
209   summands <- (2*(1:n)-1)/(2*n) - pnorm(sorted_data)
210   Test_stat <- 1/(12*n) + sum(summands^2)
211   Test_stat
212
213   # adjusted test statistic for p value
214   Z_adj <- Test_stat*(1+0.5/n)
215
216   # p-value
217   if(Z_adj < 0.0275){
218     p <- 1 - exp(-13.953 + 775.5*Z_adj - 12542.61*Z_adj^2)
219   } else if(0.0275 < Z_adj & Z_adj < 0.051){
220     p <- 1 - exp(-5.903 + 179.546*Z_adj -1515.29*Z_adj^2)
221   } else if(0.051 < Z_adj & Z_adj < 0.092){
222     p <- exp(0.886 - 31.62*Z_adj + 10.897*Z_adj^2)
223   } else if(Z_adj > 0.092){
224     p <- exp(1.111 - 34.242*Z_adj + 12.832*Z_adj^2)
225   }
226
227   res <- data.frame(Test_stat,p)
228   names(res) <- c("statistic", "p.value")
229
230   return(res)
231 }
232
233
234 ############################
235 ##### Anderson-Darling Test
236 ############################
237
238 # The R function *DescTool::AndersonDarlingTest* or the function *goftest::
      ad.test* are based on an approximation of the asymptotic distribution
      function $A_n$ that was published by Anderson and Darling.
239 # Additionally to this algorithm, the author also offered a faster
      calculation process which still has an accuracy to the 4th digit.
240 # This faster method uses a polynomial approximation of the implemented
      series expansion by Anderson and Darling and a correction term
      consisting of smoothed functions derived from simulations to account for
       the use of an approximation.
241
242 # The calculation algorithm is implemented below and compared to the
      mentioned R functions. The results for random sample were identical up
```

```
    to machine precision.
243
244 # calculates an approximation for Prob(A < a)
245 adinf <- function(A){
246   if(A<2){
247     return(exp(-1.2337141/A)/sqrt(A)*(2.00012+(.247105-(.0649821-(.0347962-
248
    (.011672-.00168691*A)*A)*A)*A)*A))
249   } else {
250     return(exp(-exp(1.0776-(2.30695-(.43424-(.082433-(.008056 -.0003146*A)*A
    )*A)*A)*A)))
251   }
252 }
253 # A is the value of the asymptotic test statistic
254
255 # errfix(n,x) corrects the error caused by using the asymptotic statistic
    for finite samples
256 # making the procedure accurate up to the 4th ot 5th digit
257 errfix <- function(n,x){
258   if(x>0.8){
259     return((-130.2137+(745.2337-(1705.091-(1950.646-(1116.360-255.7844*x)*x)
    *x)*x)*x)/n)
260   } else {
261
262     c=0.01265+0.1757/n
263     if(x < c){
264       t=x/c
265       t=sqrt(t)*(1.-t)*(49*t-102);
266       return( t*(.0037/(n*n)+.00078/n+.00006)/n )
267     } else{
268       t=(x-c)/(.8-c)
269       t=-.00022633+(6.54034-(14.6538-(14.458-(8.259-1.91864*t)*t)*t)*t)*t
270       return( t*(.04213+.01365/n)/n )
271     }
272   }
273 }
274
275
276 # the function AD(n,A) returns Prob(A < a)
277 AD <- function(n,A){
278   x <- adinf(A)
279   v <- errfix(n,x)
280   return(x+v)
281 }
282
283 AD.test <- function(x){
284
285   # calculate the test statistic
286   n <- length(x)
287   sort_data <- sort(x)
288   std_data <- (sort_data - mean(sort_data))/sd(sort_data)
289   p_i <- pnorm(std_data)
290   A_sigma <- sum( (2*c(1:n)-1)*(log(p_i) + log(1-p_i[n:1])) )
291   A <- -n-(1/n)*A_sigma
292
```

```
293   #calculate the p value
294   p <- 1 - AD(n,A)
295
296   # result object:
297   result <- data.frame(statistic = A, p.value = p)
298   return(result)
299 }
300
301
302
303
304 ######################################################
305 ##### Adjusted Version of the Anderson-Darling Test
306 ######################################################
307
308 # Results compared with the R function *nortest::ad.test* were identical up
      to machine precision, but only applicable for samples with size 8 or
      more.
309
310 adj.AD.test <- function(x){
311
312   # calculate the test statistic
313   n <- length(x)
314   sort_data <- sort(x)
315   std_data <- (sort_data - mean(sort_data))/sd(sort_data)
316   p_i <- pnorm(std_data)
317   A_sigma <- sum( (2*c(1:n)-1)*(log(p_i) + log(1-p_i[n:1])) )
318   A <- -n-(1/n)*A_sigma
319
320   # p-value
321   Z <- A*(1+ 0.75/n + 2.25/n^2)
322
323   if(Z >= 0.6){
324     p_val <- exp(1.2937 - 5.709*Z + 0.0186*Z^2)
325   } else if (0.34 < Z & Z < 0.6){
326     p_val <- exp(0.9177 - 4.279*Z - 1.38*Z^2)
327   } else if (0.2 < Z & Z < 0.34){
328     p_val <- 1- exp(-8.318 + 42.796*Z - 59.938*Z^2)
329   } else if (Z <= 0.2){
330     p_val <- 1 - exp(-13.436 + 101.14*Z - 223.73*Z^2)
331   }
332
333   # result object
334   result <- data.frame(statistic = A, p.value = p_val)
335
336 }
```

Bibliography

[1] Baddeley, A. (2021). *Distribution of the Cramér-von Mises test statistic.* From Matlab code written by Julian Faraway, Translated to R by Adrian Baddeley. https://github.com/baddstats/goftest/blob/master/R/cramer.R

[2] Casella, G. & Berger, R. L. (2002). *Statistical inference.* Second edition. Boca Raton, London, New York, CRC Press, Taylor & Francis Group.

[3] Chiang, C. L. (2003). *Statistical methods of analysis.* Singapore, World Scientific Publ., Co, Inc., River Edge, NJ.

[4] Cramér, H. (1999). *Theory of Estimation. In Mathematical Methods of Statistics (PMS-9).* 473–524. Princeton University Press.

[5] Cromwell, J. B., Labys, W. C. & Terraza, M. (1994). *Univariate Tests for Time Series Models.* 20-22. Sage Publications, Thousand Oaks, CA.

[6] Csörgő, S. & Faraway, J.J. (1996). *The exact and asymptotic distributions of Cramér-von Mises statistics.* Journal of the Royal Statistical Society, Series B 58, 221–234.

[7] D'Agostino, R. B. & Stephens, M. A. (1986). *Goodness-of-Fit Techniques.* New York. NY, Dekker.

[8] Faraway, J.,Marsaglia, G., Marsaglia, J. & Baddeley, A. (2021). *R Package 'goftest': Classical Goodness-of-Fit Tests for Univariate Distributions.* https://cran.r-project.org/web/packages/goftest/goftest.pdf

[9] Fisher, T. J & Robbins, M.W. (2019). *A Cheap Trick to Improve the Power of a Conservative Hypothesis Test.* The American Statistician 73(3), 232–242.

[10] Gel, Y. R. & Gastwirth, J. L.(2008). *A robust modification of the Jarque-Bera test of normality*, Economics Letters 99, 30-32.

[11] Harris, D. (2014). *Why heavy tails.* West Virginia University. https://swer.wtamu.edu/sites/default/files/Data/swer%2014%20Harris.pdf

[12] Joanes, D. N. & Gill, C. A. (1998). *Comparing Measures of Sample Skewness and Kurtosis.* Journal of the Royal Statistical Society. Series D (The Statistician), 47(1), 183–189.

[13] Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses.* Third Edition. Cham, Springer.

[14] Marsaglia, G. & Marsaglia, J. (2004). *Evaluating the Anderson-Darling Distribution.* Journal of Statistical Software 9(2), 1–5.

[15] Marsaglia, G., Tsang, W. W. & Wang, J. (2003). *Evaluating Kolmogorov's Distribution.* Journal of Statistical Software, 8(18), 1–4.

[16] Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A. & Firth, D. (2025). *R Package 'MASS': Support Functions and Datasets for Venables and Ripley's MASS.* https://cran.r-project.org/web/packages/MASS/MASS.pdf

[17] Royston, P. (1993). *A Toolkit for Testing for Non-Normality in Complete and Censored Samples.* Journal of the Royal Statistical Society. Series D (The Statistician), 42(1), 37–43.

[18] Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* Wiley Seri. Probab. Math. Statist., John Wiley & Sons, Inc., NY. New York, Wiley.

[19] Sprott, D. A. (2000). *Statistical inference in science.* Springer Ser. Stat. Stapringer-Verlag, New York.

[20] Thode, H. C. (2002). *Testing for Normality.* Statist. Textbooks Monogr. 164, Marcel Dekker, New York.

[21] Walck, C. (1996). *Hand-book on STATISTICAL DISTRIBUTIONS.* Version of Spetember 2007. University of Stockholm.