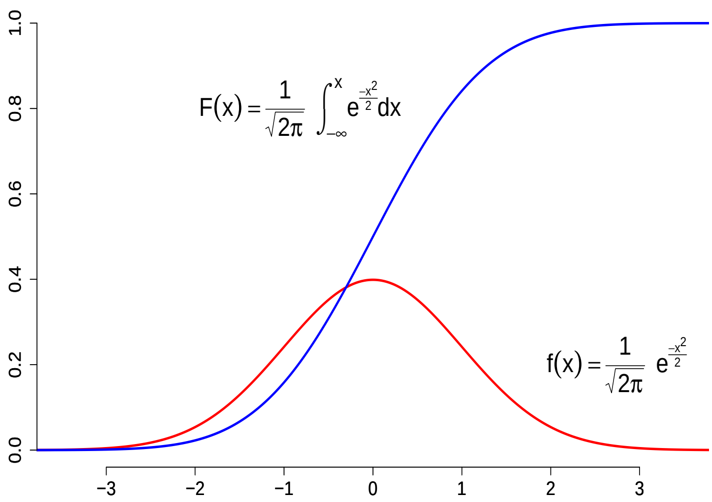
Julius-Maximilians-Universität Würzburg
Institut für Mathematik und Informatik
Lehrstuhl für Mathematik VIII (Statistik)

Comparison of Common Tests for Normality



probability and cumulative density function of N(0,1)

Diplomarbeit

vorgelegt von
Johannes Hain

Betreuer: Prof. Dr. Michael Falk

August 2010

Contents

Chapter 1

Introduction

One of the most, if not the most, used distribution in statistical analysis is the normal distribution. The topic of this diploma thesis is the problem of testing whether a given sample of random observations comes from a normal distribution. According to Thode (2002, p. 1), "normality is one of the most common assumptions made in the development and use of statistical procedures." Some of these procedures are

- the t-test,
- the analysis of variance (ANOVA),
- tests for the regression coefficients in a regression analysis and
- the F-test for homogeneity of variances.

More details of these and other procedures are described in almost all introductory statistical textbooks, as for example in Falk et al. (2002). From the authors point of view, starting a work like this without a short historical summary about the development of the normal distribution would not make this thesis complete. Hence, we will now give some historical facts. More details can be seen in Patel and Read (1996), and the references therein.

The first time the normal distribution appeared was in one of the works of Abraham De Moivre (1667–1754) in 1733, when he investigated large-sample properties of the binomial distribution. He discovered that the probability for sums of binomially distributed random variables to lie between two distinct values follows approximately a certain distribution—the distribution we today call the normal distribution. It was mainly Carl Friedrich Gauss (1777–1855) who revisited the idea of the normal distribution in his theory of errors of observations in 1809. In conjunction with the treatment of measurement errors in astronomy, the normal distribution became popular. Sir Francis Galton (1822–1911) made the following comment on the normal distribution:

> *I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion.*

After the discovery of the central limit theorem by Pierre-Simon de Laplace (1749–1827), for most of the nineteenth century, it was a common opinion in research that almost all measurable quantities were normally distributed, if only an accurate number of observations was regarded. But with the appearance of more and more phenomena, where the normality could not be sustained as a reasonable distribution assumption, the latter attitude was cast on doubt. As noted by Ord (1972, p. 1), "Towards the end of the nineteenth century [...] it became apparent that samples from many sources could show distinctly non-normal characteristics". Finally, in the end of the nineteenth century most statisticians had accepted the fact, that distributions of populations might be non-normal. Consequently, the development of tests for departures from normality became an important subject of statistical research. The following citation of Pearson (1930b, p. 239) reflects rather accurately the ideas behind the theory of testing for normality:

> *"[...] it is not enough to know that the sample* could *come from a normal population; we must be clear that it is at the same time improbable that it has come from a population differing so much from the normal as to invalidate the use of 'normal theory' tests in further handling of the material."*

Before coming to the actual topic, we first have to make some comments about basic assumptions and definitions.

Settings of this work

Basis of all considerations, the text deals with, is a random sample $y_1, \dots, y_n$ of n independent and identically distributed (iid) observations as realizations of a random variable Y. Denote the probability densitiy function (pdf) of Y by $p_Y(y)$. Since we focus our considerations on the investigation of the question, whether a sample comes from a normal population or not, the null hypothesis for a test of normality can be formulated as

$$H_0 : p_Y(y) = p_{N(\mu,\sigma^2)}(y) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad -\infty < y < \infty, \tag{1.1}$$

where both the expectation $\mu \in \mathbb{R}$ and the variance $\sigma^2 > 0$ are unknown. Note that instead of $p_{N(0,1)}(y)$ we write

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\, y^2\Big), \quad -\infty < y < \infty, \tag{1.2}$$

in this work for the pdf of a standard normally distributed random variable. A hypothesis is called **simple hypothesis**, if both parameters μ and σ^2 are completely specified that is, one is testing H_0 with $\mu = \mu_0$ and $\sigma = \sigma_0^2$, where μ_0 and σ_0^2 are both known. If at least one of the parameters is unknown, we call H_0 a **composite hypothesis**. For the remainder of this work, we suppose that μ as well as σ^2 are unknown, since this is the most realistic case of interest in practice.

For the alternative hypothesis we choose the completely general hypothesis that the distribution of the underlying population is not normal, viz.,

$$H_1 : p_Y(y) \neq p_{N(\mu,\sigma^2)}(y) \tag{1.3}$$

for all $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Historically, non-normal alternative distributions are classified into three groups based on their third and fourth standardized moments $\sqrt{\beta_1}$ and β_2, respectively. These two measures will be discussed more detailed in section 3.1. An alternative distribution with $\sqrt{\beta_1} \neq 0$ is called skewed. Alternative distributions which are symmetric are grouped into alternatives with β_2 greater or smaller 3, which is the value of the kurtosis for the normal distribution. Tests that can only detect deviations in either the skewness or the kurtosis are called **shape tests**. The tests that are able to cover both alternatives are called **omnibus tests**.

Contents

Picking up the statement of Pearson (1930b) above, the following text discusses some tests designed to test formally the appropriateness or adequacy of the normal distribution as a model of the underlying phenomen from which data were generated. The number of different tests for normality seems to be boundless when studying the corresponding literature. Therefore, we have to limit the topics presented here and exclude some testing families. The chi-square type tests (cf. Moore (1986)) are excluded as well as tests based on the empirical densitiy function (EDF) like the Kolmogorov-Smirnov test (cf. Stephens (1986)). To give reasons for this decision, we cite D'Agostino (1986a, p. 406) who state that "[...] when a complete sample is available, the chi-square test should not be used. It does not have good power [...]". Additionally they say that "For testing for normality, the Kolmogorov-Smirnov test is only a historical curiosity. It should never be used. It has poor power [...]" (see D'Agostino (1986a, p. 406)). We are also not going to consider informal graphical techniques, like boxplots or Q-Q plots to examine the question whether a sample is normally distributed. For an overview of this topic, we refer to D'Agostino (1986b).

In this diploma thesis two different approaches in testing a random sample of (iid) observations for normality are investigated. The first approach consists in using regression-type tests in order to summarize the information that is contained in a normal probability plot. The most important representantive of these type of tests is the Shapiro-Wilk test. In the first chapter of the thesis, the derivation of this test is presented as well as analytical properties of the corresponding test statistic. Further, some modifications and extensions for larger sample sizes will be introduced and discussed. For each of the mentioned tests, empiricial significance points are calculated based on Monte Carlo simulations with a number of $m = 1.000.000$ repetitions for each test and each sample size.

In the next chapter, the second approach is discussed, that consists of testing for normality using the third and fourth standardized sample moments of the observations $y_1, \ldots, y_n$, also known as sample skewness, $\sqrt{b_1}$, and sample kurtosis, b_2. After giving a motivation for employing this approach, single shape tests based only on either $\sqrt{b_1}$ or b_2 are discussed. The next step is using both tests together with the objective to get an omnibus test. The probably most popular omnibust test, the Jarque-Bera test, is introduced, as well a modification of this test. For all tests of the Jarque-Bera type, critical points are determined based on empirical sampling studies.

In the third chapter all introduced test are compared in the framework of a power study. In this study, many samples with an underlying distribution differing to the normal distribution are generated and tested for normality with all the mentioned tests. The empirical power of each test, for each alternative distribution and for each sample size is calculated by the rate of rejection for each single combination of alternative distribution and sample size. At the end of the chapter, the results will be discussed.

Chapter 2

The Shapiro-Wilk Test for Normality

An outstanding progress in the theory of testing for normality is the work of Shapiro and Wilk (1965). As noted by D'Agostino (1982, p. 200), the work "represents the first true innovation in the field since the 1930s". The main idea of the proposed test procedure consists of combining the information that is contained in the normal probability plot with the information obtained from the estimator of the standard deviation of the sample.

To understand this main idea, we will start with a closer look on this approach to get more familiar with the theoretical deviation of the test statistic. Before two attempts to interprete the test statistic are given, the properties of the test statistic for the Shapiro-Wilk test will be investigated more detailed. In the next subsection different ways to compute the coefficients of the W statistic will be presented. In the last subsection, some modifications and extensions of the Shapiro-Wilk test are discussed shortly.

2.1 Derivation of the Test Statistic

Consider a random sample $x_1, \ldots, x_n$ of n independently and identically distributed (iid) observations coming from a standard normally distributed random variable X , i.e.,

$$P\{x_i \leq t\} = \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u, \quad t \in \mathbb{R}, \quad i = 1, \ldots, n.$$

Denote by $x_{(1)} \leq \cdots \leq x_{(n)}$ the ordered values of the random sample $x_i, i = 1, \ldots, n$ and by $\mathbf{x}' = (x_{(1)}, \ldots, x_{(n)})$ the random vector of the ordered random variables. In this chapter, for the sake of easiness, we will omit the brackets around the indices in the following, so that $x_1, \ldots, x_n$ represents an already ordered sample, unless otherwise stated.

Further let $\mathbf{m}' = (m_1, \ldots, m_n)$ denote the vector of the expected values of the $x_i, i = 1, \ldots, n$

and $V = (v_{ij})$ the corresponding $n \times n$–covariance matrix, viz.,

$$\begin{aligned} E(x_i) &= m_i, \quad i = 1, \dots, n, \\ \mathrm{Cov}(x_i, x_j) &= v_{ij}, \quad i, j = 1, \dots, n. \end{aligned} \tag{2.1}$$

In the following of this work, we will retain this notation. Hence, $m_i, i = 1, \dots, n$ will always stand for the expectation of the i-th order statistic of a standard normally distributed random variable and $v_{ij}, i, j = 1, \dots, n$ will always stand for the covariance of the i-th and the j-th order statistic of the same random variable. An overview for the methods of the calculation of m_i and v_{ij} as well as techniques to find approximations of them is given in section 2.3.

Of course, the main objective is to check whether a random variable Y is normally distributed with unknown mean μ and unknown variance σ^2. Suppose that for the investigation of this question we have a random sample $y_1, \dots, y_n$ of n iid observations of Y. In order to decide whether the y_i are realizations of an underlying normally distributed random variable, the random sample has to be ordered in ascending order. Let $\mathbf{y}' = (y_{(1)}, \dots, y_{(n)})$ be the vector of the ordered random observations. Analogue to the notation of the x_i, we use for the $y_i, i = 1, \dots, n$, the same simplification and write $y_1, \dots, y_n$ in this chapter for the sample of ordered observations, unless otherwise stated. To examine the null hypothesis in (1.1), a simple regression model is constructed based on the n observations in the vector $\mathbf{y}$. Under the assumption of a normal distributed sample, we can express the y_i as

$$y_i = \mu + \sigma x_i, \quad i = 1, \dots, n. \tag{2.2}$$

If we interpret the σx_i as the error terms, we have to define new error terms $\varepsilon_i = \sigma x_i - \sigma m_i$ in order to get error terms with expectation zero. Consequently, we have to add σm_i at the right hand side of equation (2.2) and get

$$y_i = \mu + \sigma m_i + \varepsilon_i, \quad i = 1, \dots, n. \tag{2.3}$$

Because of equation (2.1), the error terms satisfy

$$\begin{aligned} E(\varepsilon_i) &= E(\sigma x_i - \sigma m_i) = \sigma m_i - \sigma m_i = 0, \quad i = 1, \dots, n, \\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) &= \mathrm{Cov}(\sigma x_i - \sigma m_i, \sigma x_j - \sigma m_j) = \sigma^2 \mathrm{Cov}(x_i, x_j) = \sigma^2 v_{ij}, \quad i, j = 1, \dots, n. \end{aligned} \tag{2.4}$$

In order to write equation (2.3) in matrix form, we introduce the following vector notation:

$$\boldsymbol{\varepsilon} := (\varepsilon_1, \dots, \varepsilon_n)' \quad \text{and} \quad \mathbf{1} := (1, \dots, 1)' \in \mathbb{R}^n.$$

Now we get

$$\mathbf{y} = \mu \mathbf{1} + \sigma \mathbf{m} + \boldsymbol{\varepsilon} =: PT + \boldsymbol{\varepsilon}, \tag{2.5}$$

where $P = (\mathbf{1}\ \mathbf{m}) \in \mathbb{R}^{n \times 2}$, $T' = (\mu\ \sigma) \in \mathbb{R}^{1 \times 2}$. The covariance matrix in (2.4) can be presented as

$$\begin{aligned} \mathrm{Cov}(\mathbf{y}) &= \mathrm{Cov}\left(E((y_i - E(y_i))(y_j - E(y_j)))\right)_{i,j} \\ &= \left(E(\varepsilon_i \varepsilon_j)\right)_{i,j} = \left(\mathrm{Cov}(\varepsilon_i, \varepsilon_j)\right)_{i,j} \\ &= \sigma^2 \left(\mathrm{Cov}(x_i, x_j)\right)_{ij} = \sigma^2 V, \end{aligned}$$

with $V := (\mathrm{Cov}(x_i, x_j))_{i,j=1,\dots,n} \in \mathbb{R}^{n\times n}$ being the covariance matrix of the standard normal order statistics.

To construct the test statistic of the Shapiro-Wilk test, we need the best estimators of the generalized regression model defined in (2.5).

2.1 Theorem. *The best linear unbiased estimators (BLUEs) of the two parameters μ and σ in the generalized regression model* (2.5) *are, respectively,*

$$\hat{\mu} = \frac{\mathbf{m}'V^{-1}(\mathbf{m}\mathbf{1}' - \mathbf{1}\mathbf{m}')V^{-1}\mathbf{y}}{\mathbf{1}'V^{-1}\mathbf{1}\mathbf{m}'V^{-1}\mathbf{m} - (\mathbf{1}'V^{-1}\mathbf{m})^2} \tag{2.6}$$

and

$$\hat{\sigma} = \frac{\mathbf{1}'V^{-1}(\mathbf{1}\mathbf{m}' - \mathbf{m}\mathbf{1}')V^{-1}\mathbf{y}}{\mathbf{1}'V^{-1}\mathbf{1}\mathbf{m}'V^{-1}\mathbf{m} - (\mathbf{1}'V^{-1}\mathbf{m})^2}. \tag{2.7}$$

Proof: Since the BLUE ist the one that minimizes the weighted sum of squares $(\mathbf{y} - PT)'V^{-1}(\mathbf{y} - PT)$ (cf. Balakrishnan and Cohen (1991, p. 80)), we get for the least square estimator

$$\hat{T} = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = (P'V^{-1}P)^{-1}P'V^{-1}\mathbf{y}. \tag{2.8}$$

For the computation of the inverse of the matrix $P'V^{-1}P \in \mathbb{R}^{2\times 2}$ we employ the well known formula for inverting a 2×2–matrix, see for example Kwak and Hong (1997, p. 25):

$$\begin{aligned}
(P'V^{-1}P)^{-1} &= \left(\begin{pmatrix} \mathbf{1}' \\ \mathbf{m}' \end{pmatrix} V^{-1}(\mathbf{1}\ \mathbf{m}) \right)^{-1} \\
&= \begin{pmatrix} \mathbf{1}'V^{-1}\mathbf{1} & \mathbf{1}'V^{-1}\mathbf{m} \\ \mathbf{1}'V^{-1}\mathbf{m} & \mathbf{m}V^{-1}\mathbf{m} \end{pmatrix}^{-1} \\
&= \frac{1}{\Delta} \begin{pmatrix} \mathbf{m}'V^{-1}\mathbf{m} & -\mathbf{1}'V^{-1}\mathbf{m} \\ -\mathbf{1}'V^{-1}\mathbf{m} & \mathbf{1}'V^{-1}\mathbf{1} \end{pmatrix},
\end{aligned} \tag{2.9}$$

where

$$\Delta = \det(P'V^{-1}P) = \det \begin{pmatrix} \mathbf{1}'V^{-1}\mathbf{1} & \mathbf{1}'V^{-1}\mathbf{m} \\ \mathbf{1}'V^{-1}\mathbf{m} & \mathbf{m}V^{-1}\mathbf{m} \end{pmatrix} = \mathbf{1}'V^{-1}\mathbf{1}\mathbf{m}V^{-1}\mathbf{m} - (\mathbf{1}'V^{-1}\mathbf{m})^2. \tag{2.10}$$

Note that we can write $\mathbf{m}'V^{-1}\mathbf{1} = \mathbf{1}'V^{-1}\mathbf{m}$ due to the fact that V^{-1} is symmetric and positive definite. Using the results in (2.9) and (2.10) we can now convert equation (2.8) to

$$\begin{aligned}
\hat{T} = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} &= \frac{1}{\Delta} \begin{pmatrix} \mathbf{m}'V^{-1}\mathbf{m} & -\mathbf{1}'V^{-1}\mathbf{m} \\ -\mathbf{1}'V^{-1}\mathbf{m} & \mathbf{1}'V^{-1}\mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{1}' \\ \mathbf{m}' \end{pmatrix} V^{-1}\mathbf{y} \\
&= \frac{1}{\Delta} \begin{pmatrix} \mathbf{m}'V^{-1}\mathbf{m}\mathbf{1}' - \mathbf{1}'V^{-1}\mathbf{m}\mathbf{m}' \\ -\mathbf{1}'V^{-1}\mathbf{m}\mathbf{1}' + \mathbf{1}'V^{-1}\mathbf{1}\mathbf{m}' \end{pmatrix} V^{-1}\mathbf{y} \\
&= \frac{1}{\Delta} \begin{pmatrix} \mathbf{m}'V^{-1}(\mathbf{m}\mathbf{1}' - \mathbf{1}\mathbf{m}') \\ \mathbf{1}'V^{-1}(\mathbf{1}\mathbf{m}' - \mathbf{m}\mathbf{1}') \end{pmatrix} V^{-1}\mathbf{y},
\end{aligned}$$

which completes the proof. $\square$

The formulas of the two estimates in (2.6) and (2.7) can be written in a more handy form. In order to do so, we first have to show a technical result.

2.2 Lemma. *For symmetric distributions—for the normal distribution in particular—we have*

$$\mathbf{1}'V^{-1}\mathbf{m} = \mathbf{m}'V^{-1}\mathbf{1} = 0.$$

Proof: Define the $n \times n$–permutation matrix

$$J = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}, \tag{2.11}$$

whose entries are 0 exept on the diagonal from bottom left to top right where the entry is 1. J is symmetric and orthogonal, i.e.

$$J = J' = J^{-1}, \quad J^2 = JJ^{-1} = I_n, \quad J'\mathbf{1} = \mathbf{1}'J = \mathbf{1}, \tag{2.12}$$

where I_n is the $n \times n$–identity matrix. Used as a multiplier from the left hand side to a vector of convenient dimension, J has the effect to reverse the order of the vector entries:

$$-J\mathbf{x} = -J \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} -x_n \\ \vdots \\ -x_1 \end{pmatrix}.$$

From the equivalence

$$x_1 \leq \cdots \leq x_n \quad \Longleftrightarrow \quad -x_n \leq \cdots \leq -x_1 \tag{2.13}$$

follows, since the x_i are symmetrically distributed, that the joint distribution of the $(x_1, \ldots, x_n)$ is the same than the joint distribution of $(-x_n, \ldots, -x_1)$ (see Balakrishnan and Cohen (1991, p. 29) and definition A.1 in the appendix). Consequently, using equation (2.13), it follows that the random vectors $\mathbf{x}$ and $-J\mathbf{x}$ have the same distribution. That means in particular $E(\mathbf{x}) = E(-J\mathbf{x}) = \mathbf{m}$ and thus,

$$\mathbf{m} = -J\mathbf{m}. \tag{2.14}$$

Moreover, it follows for the covariance matrix of $-J\mathbf{x}$ (cf. Falk et al. (2002, p. 115)):

$$V = \mathrm{Cov}(\mathbf{x}) = \mathrm{Cov}(-J\mathbf{x}) = -JV(-J)' = JVJ, \tag{2.15}$$

where equation (2.12) is used in the last equal sign. Inverting both sides of the last equation and again using (2.12) yields to

$$V^{-1} = J^{-1}V^{-1}J^{-1} = JV^{-1}J. \tag{2.16}$$

The assertion follows now, by applying the equations (2.16), (2.14) and (2.12), since we have

$$\mathbf{1}'V^{-1}\mathbf{m} = \mathbf{1}'(JV^{-1}J)(-J\mathbf{m}) = -\mathbf{1}'V^{-1}I_n\mathbf{m} = -\mathbf{1}'V^{-1}\mathbf{m}.$$

□

One can easily see, that the following result is a direct consequence of the above lemma and theorem 2.1.

2.3 Corollary. *The BLUEs in* (2.6) *and* (2.7) *can be reduced to*

$$\hat{\mu} = \frac{\mathbf{1}'V^{-1}\mathbf{y}}{\mathbf{1}'V^{-1}\mathbf{1}} \quad \textit{and} \quad \hat{\sigma} = \frac{\mathbf{m}'V^{-1}\mathbf{y}}{\mathbf{m}'V^{-1}\mathbf{m}}. \tag{2.17}$$

For the case that the underlying random variable Y is normally distributed, the estimator for the mean can be expressed even more shortly.

2.4 Remark. If the $x_1, \ldots, x_n$ are the ordered observations coming from a standard normally distributed random variable, we have $\mathbf{1}'V^{-1} = \mathbf{1}'$. Consequently, the estimate $\hat{\mu}$ in equation (2.6) for the mean reduces to the arithmetic mean $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. A proof for this can be seen in Balakrishnan and Cohen (1991, p. 59).

In the end of this section, after having determined the BLUEs for the regression model (2.5), we will finally define the test statistic for the Shapiro-Wilk test.

2.5 Definition. *Let $y_1, \ldots, y_n$ be a sample of n iid ordered random observations coming from a random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Further let*

$$S_n^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.18}$$

which is the unbiased estimate of $(n-1)\sigma^2$. The test statistic for the composite hypothesis of normality denoted by

$$W = \frac{(\mathbf{a}'\mathbf{y})^2}{S_n^2} = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.19}$$

where

$$\mathbf{a}' = (a_1, \ldots, a_n) = \frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}}$$

is called W **statistic***; for the testing procedure we will use the abbreviation* **Shapiro Wilk test** *or just W* **test***.*

The name of the test is dedicated to Shapiro and Wilk, who developed this test for normality in 1965. The abbreviation W for the test statistic has been proposed by the authors and has become widely accepted in literature.

The definition of the test statistic for the Shapiro-Wilk test seems on the first sight a little bit unexpected—especially the reasons for the use of the vector **a** are not intuitive. A justification for definition 2.5 together with an attempt to give a mathematical interpretation of the W statistic is discussed in the next section, in particular in subsection 2.2.2.

2.2 Properties and Interpretation of the W Statistic

In this section, we will try to interpret the W statistic in order to understand its functionality and to become more familiar with it. For this purpose, a few analytical properties of W will be presented first.

2.2.1 Some Analytical Features

2.6 Lemma. *Let $y_1, \dots, y_n$ be a sample of n iid ordered random observations coming from a normally distributed random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$ and let the coefficients $a_i, i = 1, \dots, n$ be defined like in definition 2.5. Then we have for $1 \leq i \leq n$:*

$$-a_i = a_{n-i+1}, \tag{2.20}$$

which means, in particular, that $\sum_{i=1}^{n} a_i = 0$.

Proof: The proof of this lemma is similar to the proof of lemma 2.2. Let $J \in \mathbb{R}^{n \times n}$ be the permutation matrix defined in (2.11). Then

$$\mathbf{a}'J = (a_1, \dots, a_n) \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} = (a_n, \dots, a_1).$$

Thus, showing the assumption (2.20) is equivalent of showing that

$$\begin{aligned} & \mathbf{a}' = -\mathbf{a}'J \\ \Longleftrightarrow & \frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}} = -\frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}} J \\ \Longleftrightarrow & \mathbf{m}'V^{-1} = -\mathbf{m}'V^{-1}J. \end{aligned}$$

By using equation (2.16) for the right hand side of the last equation, we get

$$-\mathbf{m}'V^{-1}J = -\mathbf{m}'JV^{-1}JJ = -\mathbf{m}'JV^{-1}I_n = \mathbf{m}'V^{-1},$$

where equation (2.12) is used and the fact that normal distributions are symmetric, hence we have $\mathbf{m}' = -\mathbf{m}'J$ as already explained in the proof of lemma 2.2. $\square$

2.7 Remark. (i) Note that in the case of an odd sample size, i.e., $n = 2r + 1, r \in \mathbb{N}$, the "middle" coefficient a_{r+1} is necessary zero.

(ii) The assertion of lemma 2.6 does not only hold for normal distributed samples. It can be extended to symmetric distributions.

With the result of the last lemma 2.6, we can now show a very crucial property of W.

2.8 Theorem. *W is scale and translation invariant.*

Proof: Suppose that, besides the $y_i, i = 1, \ldots, n$, we have another random sample $z_i := \alpha y_i + \beta, i = 1, \ldots, n$ with arbitrary $\alpha, \beta \in \mathbb{R}$. Then we have

$$\bar{z} := \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha y_i + \beta) = \frac{\alpha}{n}\sum_{i=1}^{n} y_i + \beta = \alpha\bar{y} + \beta.$$

Since, as shown in lemma 2.6, $\sum_{i=1}^n a_i = 0$, the W statistic for the z_i leads to

$$\begin{aligned}
W_z := \frac{\left(\sum_{i=1}^n a_i z_i\right)^2}{\sum_{i=1}^n (z_i - \bar{z})^2} &= \frac{\left(\sum_{i=1}^n a_i(\alpha y_i + \beta)\right)^2}{\sum_{i=1}^n (\alpha y_i + \beta - \alpha\bar{y} - \beta)^2} \\
&= \frac{\left(\alpha\sum_{i=1}^n a_i y_i + \beta\sum_{i=1}^n a_i\right)^2}{\alpha^2\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= \frac{\alpha^2\left(\sum_{i=1}^n a_i y_i\right)^2}{\alpha^2\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= W.
\end{aligned}$$

□

Note that the coefficients of the vector **a** are nonstochastic for an arbitrary but fixed value of the sample size n. Thus, the following corollary is an obvious consequence of theorem 2.8.

2.9 Corollary. *Let $y_1, \ldots, y_n$ be a sample of n iid ordered random observations coming from a random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Then, the distribution of the W statistic depends only on the sample size n and does not depend on the unknown location and scale parameters μ and σ. Thus, W is an appropriate statistic for testing a composite hypothesis of normality.*

This corollary is the justification for the approach in the power studies in section 4.2, where only the power behavior of standard normal distributions are analyzed. We will use the fact that W is invariant under location and scale transformations now to show that W and S_n^2 are stochastically independent from each other under the null hypothesis—a result which becomes cruical when calculating the expectation of W in lemma 2.13.

2.10 Corollary. *Let $y_1, \ldots, y_n$ denote a sample of n iid ordered random observations coming from a normally distributed random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Then, the W statistic is stochastic independent of S_n^2.*

Proof: Let

$$T := T(y_1, \ldots, y_n) = (\bar{y}, S_n^2) = \left(\frac{1}{n} \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)$$

be a statistic for the parameter $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. According to Witting (1985, p. 357), the statistic

$$\tilde{T} := \left(\frac{1}{n} \sum_{i=1}^{n} y_i, \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)$$

is a sufficient statistic for (μ, σ^2). Hence, T is a sufficient statistic for (μ, σ^2), since to get from $(n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ to $\sum_{i=1}^{n} (y_i - \bar{y})^2$ one just has to apply a one-to-one transformation (cf. Hogg and Craig (1978, p. 367)). Consequently, the assumptions of theorem A.6 in the appendix are all fulfilled and we can conclude that W and T are independent.

It remains to show that also W and S_n^2 as the second component of T are independet, as well. To this end, consider the probability

$$P(W \in A, S_n^2 \in B) = P(W \in A, T \in \mathbb{R} \times B),$$

where A is an arbitrary interval in $\left[\frac{na_1^2}{n-1}, 1 \right]$ (cf. lemma 2.11 and remark 2.12) and B is an arbitrary interval in $(0, \infty)$. Using the fact that W and T are independent, we find

$$P(W \in A, T \in \mathbb{R} \times B) = P(W \in A) P(T \in \mathbb{R} \times B) = P(W \in A) P(S_n^2 \in B),$$

which completes the proof. $\square$

2.11 Lemma. *The upper bound of the W statistic is 1.*

Proof: First, note that for all $n \in \mathbb{N}$,

$$W = \frac{(\mathbf{a}'\mathbf{y})^2}{S_n^2} \geq 0.$$

Using the translation invariance of the W statistic as shown in theorem 2.8, we can assume, without loss of generality, $\bar{y} = 0$. (Otherwise pass to $\tilde{y}_i := y_i - \bar{y}$.) Hence, the W statistic can be reduced to

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n} y_i^2}. \tag{2.21}$$

Since, by definition,

$$\sum_{i=1}^{n} a_i^2 = \mathbf{a}'\mathbf{a} = \frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}} \frac{V^{-1}\mathbf{m}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}} = \frac{\mathbf{m}'V^{-1}V^{-1}\mathbf{m}}{\mathbf{m}'V^{-1}V^{-1}\mathbf{m}} = 1, \tag{2.22}$$

we get for the nominator of (2.21)

$$\left(\sum_{i=1}^{n} a_i y_i \right)^2 \leq \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i^2,$$

where the inequality follows from the inequality of Cauchy-Schwarz (see for example Kwak and Hong (1997, p. 164)). Thus the upper bound of W is 1, while $W = 1$ if and only if $y_i = \lambda a_i, i = 1, \ldots, n$ for arbitrary $\lambda \in \mathbb{R}$. $\square$

2.12 Remark. It can be shown that there is even a larger lower bound for W than zero, namely $\frac{na_1^2}{n-1}$. For a proof of that we refer to the original work of Shapiro and Wilk (1965).

The fact, that W cannot take a value bigger than 1 is an important property of the W statistic. We will discuss later that a value of W near one is an incidence that the underlying sample is actually normally distributed, while by getting a small value of W we tend to reject the null hypothesis of a normal distribution. However, in the proof of lemma 2.11 the reader may get a first idea why choosing the vector of coefficients $\mathbf{a}$ in the W statistic. The "normalizing" property of $\mathbf{a}$ in equation (2.22) is one of the main reasons for this choice.

2.13 Lemma. *The first moment of W is given by*

$$E(W) = E\Big(\frac{b^2}{S_n^2}\Big) = \frac{\kappa(\kappa+1)}{\psi(n-1)},$$

where $b^2 = (\mathbf{a}'\mathbf{y})^2$, $\kappa = \mathbf{m}'V^{-1}\mathbf{m}$, and $\psi = \mathbf{m}'V^{-1}V^{-1}\mathbf{m}$.

Proof: In the first part of the proof we show that $E(W) = E(b^2/S_n^2) = E(b^2)/E(S_n^2)$, that is, we have to show that the expectation of the ratio $E(b^2/S_n^2)$ is the ratio of the expectations $E(b^2)/E(S_n^2)$. To this end, note that

$$\mathrm{Cov}(W, S_n^2) = \mathrm{Cov}\bigg(\frac{b^2}{S_n^2}, S_n^2\bigg) = E(b^2) - E\bigg(\frac{b^2}{S_n^2}\bigg)E(S_n^2). \tag{2.23}$$

From equation (2.23) follows that

$$E(W) = E\bigg(\frac{b^2}{S_n^2}\bigg) = \frac{E(b^2)}{E(S_n^2)} \quad \Longleftrightarrow \quad \mathrm{Cov}\bigg(\frac{b^2}{S_n^2}, S_n^2\bigg) = \mathrm{Cov}(W, S_n^2) = 0.$$

But in corolarry 2.10 we have already shown that W and S_n^2 are stochastically independet and hence uncorrelated.

In the second part of the proof we can now determine the single expectations of b^2 and S_n^2. Clearly, we have $E(S_n^2) = (n-1)\sigma^2$. Since $\mathrm{Var}(\hat{\sigma}) = E(\hat{\sigma}^2) - (E(\hat{\sigma}))^2 = \sigma^2/\kappa$ (see for example Balakrishnan and Cohen (1991, p.82)) and using equation (2.17), we can write

$$E(b^2) = \frac{\kappa^2}{\psi}(E(\hat{\sigma})^2) = \frac{\kappa^2}{\psi}(\mathrm{Var}(\hat{\sigma}) + E(\hat{\sigma})^2) = \frac{\sigma^2\kappa(\kappa+1)}{\psi},$$

which leads to the desired result. $\square$

In subsection 2.2.2 we mentioned that W can be interpreted as the squared ratio of two different estimates of the standard deviation—at least under the null hypothesis of a normal distributed sample. From the fact that under H_0, both $\hat{\sigma}^2$ and $(n-1)^{-1}S_n^2$ are estimating the variance of the underlying distribution (see Shapiro (1998, p. 480)). It follows that their ratio is approximately near 1. Hence we can write informally

$$\frac{\hat{\sigma}^2}{(n-1)^{-1}S_n^2} \approx 1 \quad \Longrightarrow \quad E\left(\frac{\hat{\sigma}^2}{(n-1)^{-1}S_n^2}\right) \approx 1. \tag{2.24}$$

Note that equation (2.24) does not follow a strictly mathematical notation but is more basis of intuitive considerations. As we will show below, W can be rewritten as that ratio of $\hat{\sigma}^2$ and $(n-1)^{-1}S_n^2$ with the constant ζ is given in equation (2.26). Consequently, this yields to

$$\begin{aligned} E(W) = \frac{\kappa(\kappa+1)}{\psi(n-1)} = \frac{\kappa^2+\kappa}{\psi(n-1)} &= E\left(\frac{\kappa^2}{\psi(n-1)}\,\frac{\hat{\sigma}^2}{(n-1)^{-1}S_n^2}\right) \\ &= \frac{\kappa^2}{\psi(n-1)}\,E\left(\frac{\hat{\sigma}^2}{(n-1)^{-1}S_n^2}\right) \\ &\approx \frac{\kappa^2}{\psi(n-1)}, \end{aligned}$$

where the last equation is based once again on intuitively considerations. Taking out common factors on both sides of the equation leads to $\kappa^2+\kappa \approx \kappa^2$. Hence, it seems that there is a little bias in estimating the ratio of $\hat{\sigma}^2$ and $(n-1)^{-1}S_n^2$ by the W statistic. However, this biasedness has, by the best knowledge of the author, never been a point of extensive mathematical investigations in literature. The main reason for that is probably that the first moment of W, $E(W)$, does not play an important role in the acutal testing procedure for normality.

The last topic we want to investigate in this subsection is the null distribution of W, i.e. the distribution of W under the null hypothesis. Unfortunately, according to Shapiro and Wilk (1965), there is no possibility for giving an explicit form of the null distribution of W for sample sizes $n \geq 4$. Shapiro and Wilk showed that there exists an implicit form for the distribution of W, for more details of this proof we refer to the original work. This fact is one of the main deficits of the Shapiro-Wilk test (and also of all the other probability plot correlation tests in section 2.4). Also after 1965, no attempts to obtains explicit forms of the distribtution of W under H_0 seem to appear in literature. Thus, to get more information about the distribution and their percentage points needed for testing for normality, empirical simulation methods have to be considered, like described in subsection 2.2.2.

Since the fact that the distribution of W is mathematically difficult to handle is very unsatisfying, at this point, we will try to get a better understandig of the distribution of W with the help of some graphics. In figure 2.1 some histograms are shown for the distribution of W for several samples sizes. For each graphic, m = 1,000,000 standard normally distributed random samples of the corresponding sample size n were generated and then plotted as a histogram. Since the number of m is quite large, the shape of these histograms comes very near to the real unknown pdf of W. By taking a more close look to the histograms, it becomes clear that the probability densitiy function of W is strongly left-skewed for every sample size (see section 3.1

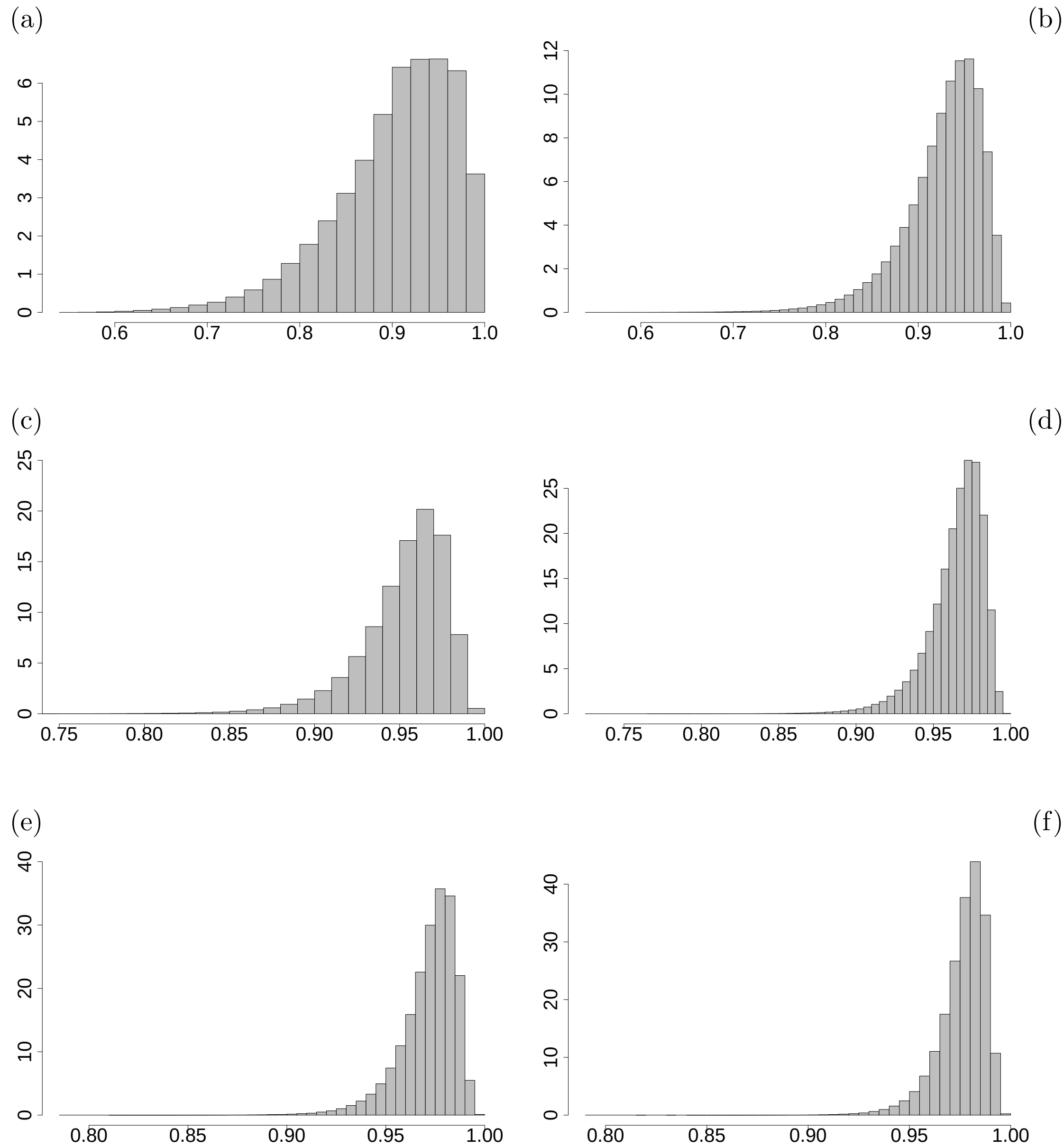


Figure 2.1: Histograms of the W statistic for several samples sizes. For each samples size, 1,000,000 standard normal samples were generated independently from each other and the W statistic was calulated each time. The samples sizes are (a) $n = 5$, (b) $n = 10$, (c) $n = 20$, (d) $n = 30$, (e) $n = 40$, (f) $n = 50$.

for a definition of skewness). For relatively small sample sizes the skewness is less pronounced. For example for $n = 5$ the empirical skewness is -1.06. However, the bigger the sample sizes get, the more left skewed becomes the pdf of W. For $n = 20$ the value of the skewness is even -1.58 and for $n = 40$ the value is -1.6. Another thing to notice is that for larger sample sizes, the distribution of W tends more and more to its maximum value 1. While for $n = 5$, the median of the m realizations of W is 0.91, it becomes 0.96 for $n = 20$ and 0.98 for $n = 50$.

2.2.2 Interpretation of the W Statistic

Due to the fact that there is no way to express the distribution of W under the null hypothesis, the interpretation of the definition of the W statistic in (2.19) is not an easy task. Nevertheless, we will try to describe the motivation of the derivation of W, in order to become more acquainted with the functionality of the Shapiro Wilk test. There are two different approaches for this problem, since W can be interpreted as a measure of linearity and as the ratio for two different estimates of the deviation of a sample. We will present both points of view in the following.

The W Statistic as a Measure of Linearity

To comprehend this approach, recall in mind that under the null hypothesis

$$E(y_i) = \mu + \sigma m_i, \quad i = 1, \ldots, n.$$

That is, a plot of the ordered observations $y_1, \ldots, y_n$ and the expectations $m_1, \ldots, m_n$ of the order statistics from a standard normal distribution should be approximately linear, if the y_i are actually coming from a normal distributed population. This is demonstrated in figure 2.2 where the probability plot of two different samples each of size $n = 50$ is plotted. In figure (a), the underlying data was generated from a standard normal distribution, while in figure (b) the data are randomly generated following an $\exp_2$ distribution (see subsection 4.2.2 for a definition of the exponential distribution). It is easy to see that the data points in figure (a) look quite linear due to the underlying normal distribution. Also it is obviously that one immediately doubts that the data from probability plot (b) are coming from a normal population since the plot is not really linear.

Another feature to notice in figure 2.2 is that the data points in (a) show a very close agreement to the regression line, while the data points in (b) are far away from that. That is, there is a strong indication that the points in (a) are originally coming from a normal population, whereas this assumption for the points in (b) would very strongly be casted on doubt. For the sake of completeness it should be mentioned that the plotting positions of the y-axis are the exact values $m_i, i = 1, \ldots, n$ taken from Parrish (1992a). The slope and the intercept of the regression line are the BLUEs given in (2.17), where the entries of the covariance matrix $v_{ij}, i, j = 1, \ldots, n$ were taken from Parrish (1992b). Note that there are many other propositions to get to suitable

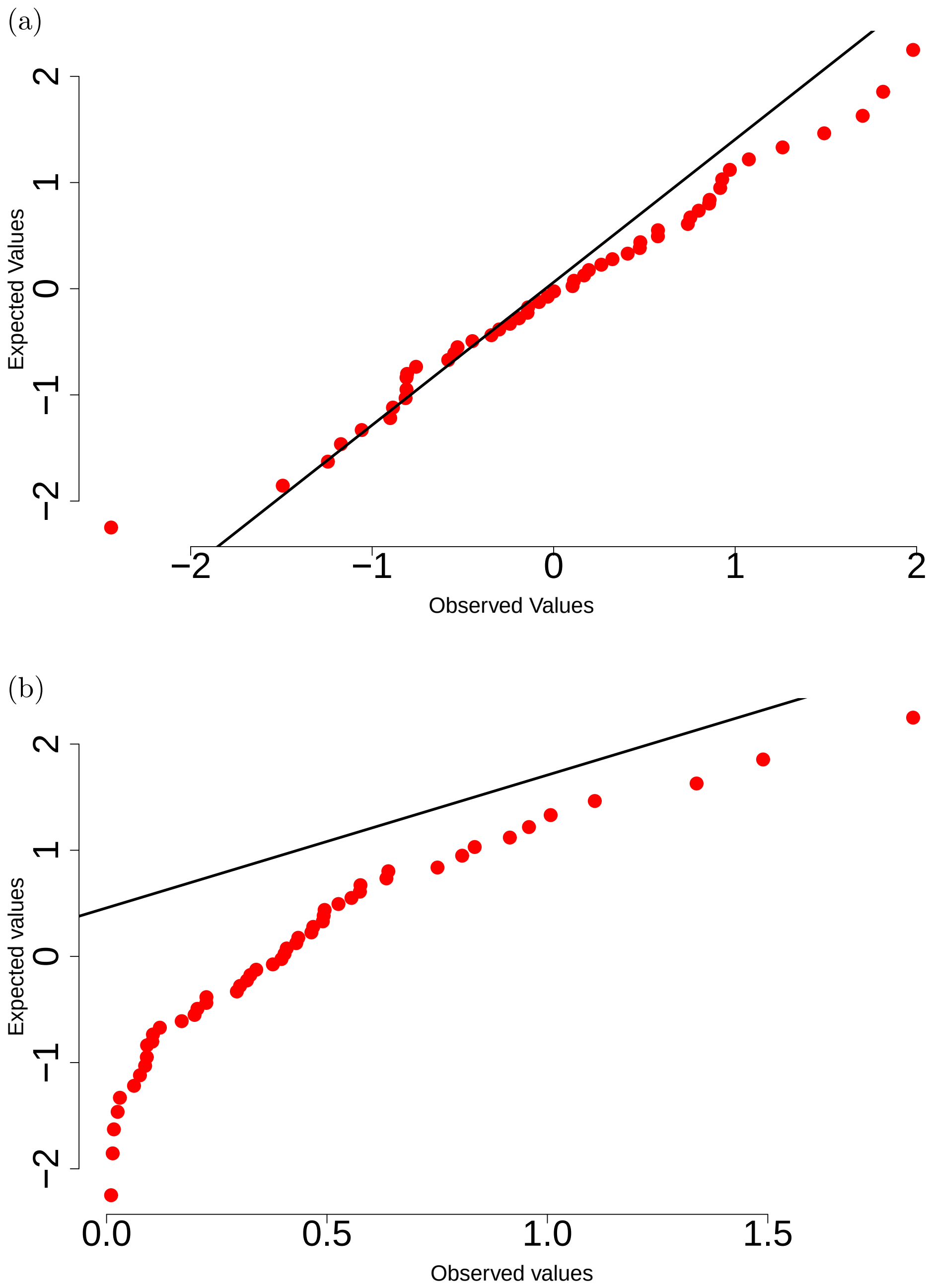


Figure 2.2: Probability plot of (a) a standard normally generated sample ($n = 50$) and of (b) a exponentially generated sample ($n = 50$) with $\lambda = 2$. The slope of the regression line was calculated as $\hat{\sigma} = (\mathbf{m}'V^{-1}\mathbf{y})/(\mathbf{m}'V^{-1}\mathbf{m})$, the intercept as $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$.

plotting points without extensive calculation of the m_i, see Sarkara and Wang (1998) or Thode (2002, table 2.1) for an overview. In section 2.3 some of these approximations are presented as well.

By looking at figure 2.2 we can summarize that the less linear the probability plot looks like, the more we doubt the null hypothesis of a normally distributed sample. The most straightforward and intuitively measure of linearity is given by the empirical correlation coefficient of the y_i and the plotting positions. Indeed, the W statistic is such a correlation coefficient as we will show in the following lemma.

2.14 Lemma. *The W statistic is the squared empirical Pearson correlation coefficient between the ordered sample values y_i and the coefficients a_i.*

Proof: We forestall the definition of the empirical Pearson correlation coefficient, that is actually given in definition 2.18, for this proof. There, the correlation coefficient is given by

$$r_{\mathbf{y},\mathbf{a}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(a_i - \bar{a})}{\left(\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (a_i - \bar{a})^2\right)^{\frac{1}{2}}},$$

where $\bar{a} = n^{-1} \sum_{i=1}^n a_i$. From lemma 2.6 we have that $\sum_{i=1}^n a_i = 0$ und thus $\bar{a} = 0$. Remembering that, by definition, $\sum_{i=1}^n a_i^2 = 1$ yields to

$$\begin{aligned} W &= \frac{\left(\sum_{i=1}^n a_i y_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\left(\sum_{i=1}^n y_i a_i - \bar{y} \sum_{i=1}^n a_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n a_i^2} \\ &= \frac{\left(\sum_{i=1}^n (y_i - \bar{y}) a_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n a_i^2} = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(a_i - \bar{a})\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (a_i - \bar{a})^2} \\ &= r_{\mathbf{y},\mathbf{a}}^2. \end{aligned}$$

□

With the result of lemma 2.14 in mind it becomes clear why Shapiro and Wilk choosed the normalized vector $\mathbf{a}$ for the vector of coefficients in the W statistic. Otherwise, the W statistic could not have been interpreted as such a correlation coefficient. The establishment of the Shaprio-Wilk test has led to many other related probability plot correlation tests using the same idea of measuring the linearity of the probabilty plot of the observed with some expected values. Some of these tests will be presented in section 2.4.1.

Like all other tests proposed in that section, the Shapiro-Wilk test is a single-tailed test, that is the null hypothesis is rejected if $W \leq c_{n,\alpha}$, where $c_{n,\alpha}$ is the critical value for the test statistic for a given sample size n and a significance level α. This is consistent with the interpretation of the correlation coeffient $r_{\mathbf{y},\mathbf{a}}$ that values bigger than 1 cannot be reached (see lemma 2.11) and that values near 1 indicate a high fit and hence are an evidence for normality.

However, according to Shapiro and Wilk the distribution of W cannot be given explicitly for $n \geq 4$ under the null distribution, hence there is no possibility in determining critical values $c_{n,\alpha}$

for W on an analytic way. To overcome this problem, Shapiro and Wilk (1965) gave empirical percentage points for the distribution of W based on Monte Carlo simulations. The principle of such a simualtion bases on the following considerations: since we do not know the explicit form of the distribution of W, we cannot calculate the quantile function to get critical values of a test. Thus, we have to compute an estimate $\hat{c}_{n,\alpha}$. The following definition describes the way to recieve to such an estimate.

2.15 Definition. *Let $\alpha \in (0,1)$ and m be the number of randomly generated independent samples of sizes n where all m samples follow a standard normal distribution. For a test of normality with test statistic T_n we can thus calculate the m test statistics and rearrange them in ascending order so that we have the ordered test statistics $T_{n,1},\ldots,T_{n,m}$. For a test of the Shapiro-Wilk type and a given sample size n, the empirical 100α % significance level $\hat{c}_{n,\alpha,m}$ of a given test for normality is defined as the value of test statistic $T_{n,\lfloor(1-\alpha)m\rfloor}$, where the floor function $\lfloor x\rfloor$ rounds x to the largest previous integer of x, i.e.,*

$$\lfloor x\rfloor = \max\{n \in \mathbb{Z} : n \leq x\}.$$

Note that in most cases m is that big that for the common values of α (e.g. $\alpha = 0.01, 0.05, 0.1$) the need of the floor function is unnecessary, since $(1-\alpha)m$ is an integer, anyway. Shapiro and Wilk choosed the following values of m in their simulation:

$$m = \begin{cases} 5{,}000 & \text{for} \quad n = 3(1)20, \\ \left\lfloor \frac{100{,}000}{n} \right\rfloor & \text{for} \quad n = 21(1)50, \end{cases}$$

where for example $n = 10(5)25$ means that the values for n go from 10 to 25 with an inkrement of 5, i.e., $n = 10, 15, 20, 25$. Since the values of m seem to be very small comparing with todays processor capabilities and the fact that after 1965, more exact values for the coefficients a_i were available, these empirical percentage points have been recalculated by Parrish (1992c), where the number of simulated W statistics was $m = 100{,}000$ for each sample size $n, n = 1,\ldots,50$. Even the extension of m to 100,000 appears to be a little too small in order to make exact statements about the percentage points of the distribution of W. That is why we recalculated the empirical percentage points of the W statistics for $m = 1{,}000{,}000$, in particular because of the need for better values for the empirical power studies in chapter 4. The results togehter with the corresponding values of Parrish are presented in table 2.1.

By taking a look at the empirical significance levels of the two sources one notices that there is only a very slight difference between the critical points of Parrish and the critical points obtained from the actual simulation study. Most of the time, the values of Parrish are a little bit higher but the difference is that marginal that might also come from rounding since Parrish only gives three decimal places in his work. To summarize the informations in table 2.1, we can state that the empirical significance points given by Parrish can be seen as very accurate.

Table 2.1: Empirical significance Points for the W statistic for several sample sizes based on $m =$ 1,000,000 repetitions compared with the critical values given by Parrish (1992c). Like in all other Monte Carlos simulation in the following of this work, the generation of the random samples was conduted in R with the function `rnorm()` (for more details R Development Core Team (see 2009)). The calculated test statistics are the same that were used to plot the histograms in figure 2.1.

	own simulation			Parrish (1992c)		
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
10	0.7856	0.8448	0.8706	0.786	0.845	0.871
15	0.8356	0.8820	0.9013	0.836	0.882	0.902
20	0.8668	0.9044	0.9200	0.867	0.905	0.929
25	0.8879	0.9195	0.9325	0.888	0.919	0.923
30	0.9032	0.9301	0.9414	0.904	0.930	0.942
35	0.9145	0.9383	0.9481	0.915	0.938	0.948
50	0.9366	0.9541	0.9613	0.937	0.954	0.961

The W Statistic as a Ratio of Two Different Estimates

As already stated in section 2.1, the sample $y_1, \dots, y_n$ of n iid ordered random observations can be expressed under the null hypothesis as

$$y_i = \mu + \sigma x_i, \quad i = 1, \dots, n. \tag{2.25}$$

Therefore, the slope of the regression line is an estimate of the scaling parameter σ, the standard deviation. According to Shapiro (1998), if the model is incorrect, i.e., if the y_i do not follow the normal distribution, the slope of this regression line is not an estimate of σ. A natural consequence in considerations for testing wether the sample of the y_i is normally distributed or not consists of comparing the estimate $\hat{\sigma}$ for the slope of the regression line with the standard deviation of the sample $y_1, \dots, y_i$ estimated by $(n-1)^{-1/2} S_n$. Note that the estimate $(n-1)^{-1/2} S_n$ does not depend on the hypothesized model (2.25), hence it is an estimate for the standard deviation no matter what underlying distribution the y_i follow. Under the null hypothesis, both quantities are estimates for the standard deviation and, thus, should be approximately equal, that is

$$\frac{\hat{\sigma}^2}{\frac{1}{n-1} S_n^2} = \frac{\left(\frac{\mathbf{m}' V^{-1} \mathbf{y}}{\mathbf{m}' V^{-1} \mathbf{m}}\right)^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} \approx 1.$$

In fact, the W statistic is up to a constant exactly the ratio of the squared values of the latter two estimates as we will see in the following lemma.

2.16 Lemma. *Let*

$$\zeta = \frac{\kappa^2}{(n-1)\psi}, \tag{2.26}$$

where $\psi = (\mathbf{m}\prime V^{-1}V^{-1}\mathbf{m})$ *and* $\kappa^2 = (\mathbf{m}'V^{-1}\mathbf{m})^2$*. Then, we have*

$$W = \frac{(\mathbf{a}'\mathbf{y})^2}{S_n^2} = \zeta\, \frac{\hat{\sigma}^2}{(n-1)^{-1}S_n^2},$$

That is, up to a constant, W is the squared ratio of the best linear unbiased estimator (BLUE) σ of the slope of a regression line of the ordered y_i on the m_i and of the estimator of the standard deviation $(n-1)^{-1/2}S_n$.

Proof: Using the definitions in (2.17) and (2.18) yields to

$$\begin{aligned}
\frac{\hat{\sigma}^2}{\frac{1}{n-1}S_n^2} &= \frac{\left(\frac{\mathbf{m}'V^{-1}\mathbf{y}}{\mathbf{m}'V^{-1}\mathbf{m}}\right)^2}{\frac{1}{n-1}\sum_{i=1}^n (y_i-\bar{y})^2} \\
&= (n-1)\frac{\mathbf{m}'V^{-1}V^{-1}\mathbf{m}}{(\mathbf{m}'V^{-1}\mathbf{m})^2}\frac{\frac{(\mathbf{m}'V^{-1}\mathbf{y})^2}{\mathbf{m}'V^{-1}V^{-1}\mathbf{m}}}{\sum_{i=1}^n (y_i-\bar{y})^2} \\
&= (n-1)\frac{\psi}{\kappa^2}\, W.
\end{aligned}$$

Since ψ, κ^2 and n are nonstochastic, they can be regarded as factors and the assertion follows. $\square$

Considering the approach of interpreting the testing procedure of Shapiro and Wilk, one is able to comprehend that the Shapiro-Wilk test is attributed to the so-called **regression tests** (cf. Thode (2002, p. 26)), because the slope of the regression line is included into the test statistic W.

2.3 The Coefficients associated with W

For the computation of the W statistic in (2.19), it is necessary to know the vector of coefficients

$$\mathbf{a}' = (a_1, \ldots, a_n) = \frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}}, \tag{2.27}$$

or equivalently, knowing the values

$$a_i = \sum_{j=1}^n m_j \frac{v_{ji}^*}{\psi}, \quad i = 1, \ldots, n,$$

where ψ is defined like in Lemma 2.13 and v_{ij}^* denotes the (i,j)-th element of the matrix V^{-1}. A serious drawback of the W statistic is that the values for $\mathbf{m}$ and V^{-1} are unknown for large sample sizes. For example, in the year 1965 when the paper of Shapiro and Wilk appeared, these values were exactly known only up to samples of size 20.

Since the knowledge how to compute the coefficients is an important issue especially for practitioners, we will give a summary of the most important developements in this field. In literature, there are two different approaches for this problem in order to find approximations for the a_i as exact as possible. The first approach chooses an indirect way and consists of trying to find approximations for the m_i and the v_{ij}. In the second, straightforward way, one attempts to solve the problem by approximating the a_i directly. In the sequel of this section, methods of resolution for these two approaches in literature will be introduced shortly.

2.3.1 The indirect approach

Approximations of the m_i

Let $x_1, \ldots, x_n$ be a sample of ordered random observations coming from a standard normally distributed random variable X. The expected value of the i-th observation is given by (see for example Balakrishnan and Cohen (1991, p. 22)):

$$\begin{aligned} E(x_i) = m_i = \frac{n!}{(n-i)!\,(i-1)!} &\int_{-\infty}^{+\infty} x \left(\frac{1}{2} - \Phi(x)\right)^{i-1} \left(\frac{1}{2} + \Phi(x)\right)^{n-i} \varphi(x)\,\mathrm{d}x, \\ &i = 1, \ldots, n, \end{aligned} \tag{2.28}$$

where $\varphi(x)$ denotes, as defined in equation 1.2, the value of the probability distribution function of the standard normal distribution evaluated at the point $x \in \mathbb{R}$. Using the latter formula and the fact that $\Phi(x) = 1 - \Phi(-x)$ it is easy to verify by insertion that for $i = 1, \ldots, n$:

$$m_i = -m_{n-i+1}, \tag{2.29}$$

(cf. Balakrishnan and Cohen (1991, p. 29)). Hence, for the calculation of the m_i one only needs to compute the values for $i = 1, \ldots, \left[\frac{1}{2}\,n\right]$, where

$$\left[\frac{1}{2}\,n\right] = \min\left\{N \in \mathbb{N} : \frac{1}{2}\,(n-1) \le N\right\} = \begin{cases} \frac{1}{2}\,n & \text{if } n \text{ is even} \\ \frac{1}{2}\,(n-1) & \text{if } n \text{ is odd.} \end{cases}$$

The most common way to determine the values of $m_i, i = 1, \ldots, n$ for different samples sizes n is to calculate the integral in (2.28) by numerical integration. An often cited paper in this framework is the work of Harter (1961) who obtained values of m_i for $n = 2(1)100(25)250(50)400$. For the mathematical details we refer to the original work and the references therein. For more accurate values of m_i with more decimal places, we refer to Parrish (1992a).

A possibility to calculate values for m_i without doing extensive numerical integration is to approximate them by an appropriate formula. The most popular approximation formula is the one given by Blom (1958):

$$E(x_i) = m_i \approx \Phi^{-1}\left(\frac{i-\eta_i}{n-2\eta_i+1}\right), \quad i = 1, \dots, n, \tag{2.30}$$

where

$$\Phi^{-1}(t) := \inf\{x \in \mathbb{R} : \Phi(x) \leq t\}, \quad t \in (0,1),$$

is the quantile function of the standard normal distribution. We will retain the definition of Φ^{-1} in the remainder of this work. For given n, the value of η_i can be obtained by

$$\eta_i \approx \frac{(n+1)\Phi(m_i) - i}{2\Phi(m_i) - 1}, \quad i = 1, \dots, \left[\frac{1}{2}\, n\right],$$

where the m_i are the exact values of the expected order statistics. Based on his results for the η_i, Blom suggested to choose $\eta = 3/8 = 0.375$ as a compromise between different values of η_i being adequate exact. The values of $\eta_i, i = 1, \dots, \left[\frac{1}{2}\, n\right]$ for $n = 2(2)10(5)20$ are given in Blom (1958, p. 70). Harter (1961) extends the table for values of η_i up to samples of size $n = 25, 50, 100, 200, 400$.

Finally it may be mentioned that Royston (1982b) presents an algorithm to compute the expected values of normal order statistics which is based on the works of Harter (1961) and Blom (1958).

Approximations of the v_{ij}

For a sample $x_1, \dots, x_n$ of n ordered random observations coming from a standard normally distributed random variable X, the covariance of the i-th and the j-th observation is defined by

$$\mathrm{Cov}(x_i, x_j) = v_{ij} = E(x_i x_j) - E(x_i)E(x_j) = E(x_i x_j) - m_i m_j, \quad i, j = 1, \dots, n.$$

Since the method to calculate the expectations m_i and $m_j, i, j = 1, \dots, n$ are already known (see above), we only need to determine the product moments $E(x_i x_j)$. The formula for the product moment of two ordered observations of a standard normal sample is given by (see for example Balakrishnan and Cohen (1991, p. 22))

$$\begin{aligned} E(x_i x_j) = {} & \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \\ & \times \int_{-\infty}^{+\infty} \int_{-\infty}^{y} xy\varphi(x)\varphi(y)\Phi(x)^{i-1}\,(1-\Phi(y))^{n-j}\,(\Phi(y)-\Phi(x))^{j-i-1}\;\mathrm{d}x\,\mathrm{d}y, \\ & i, j = 1, \dots, n. \end{aligned}$$

Of course, we have the relationship $v_{ij} = v_{ji}, i, j = 1, \dots, n$ and

$$v_{ij} = v_{n-i+1, n-j+1},$$

which can be easily verified by insertion and fundamental arithmetic operation (cf. Balakrishnan and Cohen (1991, p. 30)). That means, in particular, that not all the n^2 entries of the covariance matrix V have to be calculated, but only

$$\begin{cases} \sum_{i=1}^{n/2} 2i = \frac{n}{2}\left(\frac{n}{2}+1\right) & \text{if } n \text{ is even,} \\ \sum_{i=1}^{[n/2]} (2i-1) = \left[\frac{1}{2}n\right]^2 & \text{if } n \text{ is odd,} \end{cases}$$

which has an ease of computation effort as a consequence.

The computation of the exact values of the v_{ij} is more complicated than the computation of the expectations of the order statistics, m_i. This is mostly due to the fact that to obtain the values for $E(x_i x_j)$ the values for a large number of double integrals are needed, which is numerically difficult to handle. Like for the determination of the expectations of the normal order statistics, the calculation of the v_{ij} was done by numerical integration. For a detailed description of the procedural method, the reader my have a look in Godwin (1948). Following works on the same topic also used basically the same method.

In 1965, when the work of Shapiro and Wilk appeared, exact values for the variances and covariances of normal order statistics were known up to samples of size 20. The tabulated values can for example be seen in Sarhan and Greenberg (1956). The values in this work were used by Shapiro and Wilk (1965) for the development of their testing procedure. An extension for the variances and covariances of normal order statistics for samples up to size 50 is presented by Parrish (1992b).

Davies and Stephens (1978) presented an algorithm for the approximate calculation of the covariance matrix of normal order statistics in the programming language FORTRAN. Due to the remark of Shea and Scallan (1988), who gave an improvement for the accuracy of the algorithm, the entries for the matrix V can be computed even more exactly.

Computation of V^{-1}

By studying the various different works in literature concerning the Shapiro-Wilk test, extensions and modifications of this test or testing procedures that are based on the same regression idea, one notices that almost none of these works use the values of the covariance matrix V for the computation of the test statistic. Since it is not the matrix V that is needed for the computation of the W statistic in equation (2.19), but the inverse matrix V^{-1}, the reason for this fact is obvious. Inverting a large matrix is numerically very difficult to handle, even with today's processor capabilities, so much greater the problem was in the 1960s and 1970s when most of the testing procedures mentioned in this work were developed.

A straightforward approach for this drawback is of course the direct calculation of V^{-1}. To this end, recall in mind that if the $x_i, i = 1, \ldots, n$ are realizations of a standard normal distributed random variable X, the random variable $\Phi(X)$ is uniformly distributed on $(0,1)$ (cf. Falk et al.

(2002, corollary 1.6.4)). Let $U_i, i = 1, \dots, n$ denote the i-th order statistic of $\Phi(X)$. Then it follows (see for example Falk et al. (2002, lemma 1.6.1)) that

$$E(U_i) = \frac{i}{n+1} =: p_i, \quad i = 1, \dots, n.$$

According to Hammersley and Morton (1954) under the null hypothesis, the values of the covariances between x_i and x_j can be approximated by

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) = v_{ij} &= \frac{E(U_i)(1 - E(U_j))}{(n+2)\varphi(E(x_i))\varphi(E(x_j))} \\ &= \frac{p_i(1-p_j)}{(n+2)\varphi(m_i)\varphi(m_j)}, \quad i, j = 1, \dots, n.\end{aligned}$$

Further Hammersley and Morton stated that the matrix V^{-1} can approximately be expressed as

$$V^{-1} = (n+2)H,$$

where the entries $h_{ij}, i, j = 1, \dots, n$ of the $n \times n$ matrix H under H_0 are given by

$$h_{ii} = \varphi^2(m_i)\Big(\frac{1}{p_{i+1} - p_i} + \frac{1}{p_i - p_{i-1}}\Big) = 2(n+1)\varphi^2(m_i), \quad i = 1, \dots, n,$$

and

$$h_{i,i+1} = h_{i+1,i} = -\varphi(m_i)\varphi(m_{i+1})\frac{1}{p_{i+1} - p_i} = -(n+1)\varphi(m_i)\varphi(m_{i+1}), \quad i = 1, \dots, n-1.$$

The other entries of H are zero, i.e.,

$$v_{ij} = v_{ji} = 0, \quad j = i+2, i+3, \dots, n, \ i = 1, \dots, n.$$

Summing up the results of the last equations, we find for the inverse covariance matrix of the normal order statistics:

$$\begin{aligned}&V^{-1} = (n+1)(n+2) \\ &\times \begin{pmatrix} 2\varphi^2(m_1) & -\varphi(m_1)\varphi(m_2) & 0 & \cdots & 0 \\ -\varphi(m_1)\varphi(m_2) & 2\varphi^2(m_2) & -\varphi(m_2)\varphi(m_3) & \cdots & 0 \\ 0 & -\varphi(m_2)\varphi(m_3) & 2\varphi^2(m_3) & \cdots & 0 \\ 0 & 0 & -\varphi(m_3)\varphi(m_4) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\varphi(m_{n-1})\varphi(m_n) \\ 0 & 0 & 0 & \cdots & 2\varphi^2(m_n) \end{pmatrix}.\end{aligned} \tag{2.31}$$

However, the only work concerning tests for normality based on the idea of Shapiro and Wilk and using the approach of approximating V^{-1} is the work of Rahman and Govindarajulu (1997). More details of this test are given in subsection 2.4.2.

2.3.2 The direct approach

When avoiding the calculation of V^{-1}, methods to approximate the coefficients $a_i, i = 1, \dots, n$ of the W statistic in equation (2.27) on a direct way become an important matter. The first approach of a direct approximation of the entries for the coefficient vector $\mathbf{a}$ was given by Shapiro and Wilk (1965) when they also presented the original form of the Shapiro Wilk test. We will omit to present this way of calculation an refer to the original work instead. The reason for this is that their approximation of the a_i for $n > 20$ is of less importance, since the values for a_i given by Parrish (1992c) are based on the exact values of $\mathbf{m}$ and V^{-1} even for sample sizes larger than 20.

Another method to approximate $\mathbf{a}$ is given by Royston (1992). The main idea consists of conducting a polynomial regression analysis for the purpose of estimating the value of a_n. If the sample size is larger than 5, the value of the coefficient a_{n-1} is also estimated by another regression analysis. To become more detailed, we have a quintic regression analysis of a_n on $x = n^{-\frac{1}{2}}$ with c_n being the constant term of the regression equation, where c_i are the coefficients of the test statistic $\tilde{W}'$ of Weisberg and Bingham (1975) defined in equation (2.40) in definition 2.22. Based on the regression equation, the predicted value $\tilde{a}_n$ can be used as an approximation for a_n. Royston (1992) performed the analysis for $4 \leq n \leq 1000$ and presented the following regression equation for an approximation $\tilde{a}_n$ of a_n:

$$\tilde{a}_n = c_n + 0.221157x - 0.147981x^2 - 2.071190x^3 + 4.4424685x^4 - 2.706056x^5. \tag{2.32}$$

For a sample size larger than 5, Royston suggested as an approximation for a_{n-1} the equation

$$\tilde{a}_{n-1} = c_{n-1} + 0.042981x - 0.293762x^2 - 1.752461x^3 + 5.682633x^4 - 3.582663x^5. \tag{2.33}$$

Note that with the estimates $\tilde{a}_n$ and $\tilde{a}_{n-1}$ we always have approximations of $\tilde{a}_1 = -\tilde{a}_n$ and $\tilde{a}_2 = -\tilde{a}_{n-1}$, respectively.

For the approximation of the remaining coefficients of $\mathbf{a}$, Royston proposed

$$\tilde{a}_i = \Theta_n^{\frac{1}{2}} \tilde{m}_i \quad \text{for} \quad \begin{cases} i = 2, \dots, n-1 & \text{if} \quad n \leq 5 \\ i = 3, \dots, n-2 & \text{if} \quad n > 5, \end{cases} \tag{2.34}$$

where

$$\Theta_n = \begin{cases} \frac{1 - 2\tilde{a}_n^2}{\tilde{\mathbf{m}}'\tilde{\mathbf{m}} - 2\tilde{m}_n^2} & \text{for} \quad n \leq 5 \\ \frac{1 - 2\tilde{a}_n^2 - 2\tilde{a}_{n-1}^2}{\tilde{\mathbf{m}}'\tilde{\mathbf{m}} - 2\tilde{m}_n^2 - 2\tilde{m}_{n-1}^2} & \text{for} \quad n > 5, \end{cases} \tag{2.35}$$

and $\tilde{\mathbf{m}} = (\tilde{m}_1, \dots, \tilde{m}_n)'$ is defined like in equation (2.38) in subsection 2.4.1.

2.17 Remark. The coefficients $\tilde{a}_i, i = 1, \dots, n$ have the same normalizing property than the coefficients a_i of the Shapiro-Wilk statistic given in definition 2.5, viz.,

$$\sum_{i=1}^{n} \tilde{a}_i^2 = 1.$$

Proof: Note that for $n \leq 5$ we get from equation (2.39) that $\tilde{m}_1^2 = \tilde{m}_n^2$ and thus

$$\begin{aligned}
\sum_{i=1}^{n} \tilde{a}_i^2 &= \tilde{a}_1^2 + \sum_{i=2}^{n-1} \tilde{a}_i^2 + \tilde{a}_n^2 = 2\tilde{a}_n^2 + \sum_{i=2}^{n-1} \Theta_n \tilde{m}_i^2 \\
&= 2\tilde{a}_n^2 + \frac{1 - 2\tilde{a}_n^2}{\tilde{\mathbf{m}}'\tilde{\mathbf{m}} - 2\tilde{m}_n^2} \sum_{i=2}^{n-1} \tilde{m}_i^2 \\
&= 2\tilde{a}_n^2 + \frac{1 - 2\tilde{a}_n^2}{\sum_{i=1}^{n} \tilde{m}_i^2 - 2\tilde{m}_n^2} \sum_{i=2}^{n-1} \tilde{m}_i^2 \\
&= 2\tilde{a}_n^2 + \frac{1 - 2\tilde{a}_n^2}{\sum_{i=2}^{n-1} \tilde{m}_i^2} \sum_{i=2}^{n-1} \tilde{m}_i^2 \\
&= 2\tilde{a}_n^2 + 1 - 2\tilde{a}_n^2 = 1.
\end{aligned}$$

The proof for $n > 5$ is analogous. □

In his work from 1992, Royston also presented a comparison of his approximation $\tilde{a}_n$, the exact values a_n and of the approximation of Shapiro and Wilk (1965). The results showed that the new approximation is very accurate an can be used as an adequate estimation rule. However, no statements of the goodness-of-fit of his two regression analyses in (2.32) and (2.33) are made by the author.

2.4 Modifications of the Shapiro-Wilk Test

The innovative work of Shapiro and Wilk (1965) can be regarded as the starting point for a reinvention for the topic of testing normality in research, for example, Royston (1982a, p. 115) stated that "research into tests of non-normality was given new impetus". As a consequence, in the following years many works appeared with the objective to modify the procedure of the W test using more simple computation methods and to extend the procedure for larger sample sizes. Some of these works will shortly be presented in this section. There is a plethora of different tests with different approaches for the problem of testing for normality. Unfortunately we of course cannot present all of them here in detail. The interested reader may have a look in Thode (2002) to get an excellent overview of the different ways to tackle the problem of testing for normality and of the latest state of research.

2.4.1 Probability Plot Correlation Tests

Like the Shapiro-Wilk test, the following tests have one thing in common: they are all trying to summarize and quantify the distributional information that is contained in a normal probability plot. This is why they are also sometimes called **probability plot correlations tests** (see also subsection 2.2.2). As the Shapiro-Wilk test, these tests are omnibus tests since their test statistic is scale and translation invariant. The difference to the Shapiro-Wilk test is that their approach is more straightforward and intuitively, because they are directly measuring the linearity of the probability plot and hence the linearity of two variables. The first of the two variables is always the vector of the ordered observations $\mathbf{y}$. The second variable varies in each test and is always an exact or approximative measure of location for the ordered observations. To formalize the common ground of the probability plot correlation tests we start with the definition of the empirical Pearson correlation coefficient (cf. Falk et al. (2002, p. 92)).

2.18 Definition. *Let $y_1, \dots, y_n$ be a sample of ordered random observations coming from a random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$ and let $u_i = \mathrm{loc}(y_i), i = 1, \dots, n$ be a measure of location for the i-th order statistic y_i. Then the empirical Pearson correlation coefficient for $\mathbf{y}$ and $\mathbf{u} = (u_1, \dots, u_n)'$ is defined by*

$$r_{\mathbf{y},\mathbf{u}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(u_i - \bar{u})}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(u_i - \bar{u})^2\right)^{\frac{1}{2}}}, \tag{2.36}$$

where $\bar{u} = n^{-1} \sum_{i=1}^{n} u_i$.

The test statistics of the following testing procedures in this subsection will all base on this correlation coefficient. The idea of these testing procedures is that under the hypothesis of a normal distributed population, the plot of the y_i and the u_i will be approximately linear and thus $r_{\mathbf{y},\mathbf{u}}$ will be near 1. The more the observations y_i differ from a normal distribution, the less linear will be the plot of the y_i and the u_i since the values for the u_i are always computed under the assumption of the null hypothesis. In this case, $r_{\mathbf{y},\mathbf{u}}$ will become smaller than 1.

2.19 Remark. Note that by setting

$$\mathbf{u} = \mathbf{a} = \frac{V^{-1}\mathbf{m}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}},$$

the squared correlation coefficient $r^2_{\mathbf{y},\mathbf{u}}$ is equivalent to the W statistic, as shown in lemma 2.14. Thus, the W test can also be embedded in equation (2.36), which goes along with the interpretation of the W statistic as a correlation coefficient in subsection 2.2.2. However, since $\mathbf{a}$ is not a measure of location in the proper sense, the W statistic is not numbered among the probability plot correlation tests.

The Shapiro-Francia Test

For the computation of the W statistic in (2.19), the vector coefficients

$$\mathbf{a}' = (a_1, \dots, a_n) = \frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'V^{-1}V^{-1}\mathbf{m})^{\frac{1}{2}}}$$

has to be determined. As already explained in section 2.3, the calculation of the exact values of the $a_i, i = 1, \dots, n$ is a serious problem. Recall in mind that in 1965—when the Shapiro Wilk test appeared—the exact values for the entries of the covariance matrix of normal order statistics, V, were only known for samples of size 20 or smaller. Hence, the attempts of developing an alternative testing procedure with the same regression idea but without the described restriction can easily be comprehended. In addition, Shapiro and Wilk (1965) approximated and tabulated the values of a_i only for sample sizes $n \leq 50$. Thus, for samples of size $n > 50$ the Shapiro-Wilk testing procedure described in the latter subsections cannot be conducted.

For the purpose of overcoming these problems with the coefficients a_i and to extend the testing procedure for larger sample sizes, Shapiro and Francia (1972) chosed a somewhat different approach than Shapiro and Wilk (1965). While in the latter work, the correlation of the ordered observations plays a crucial role in the test statistic, Shapiro and Francia (1972) made the assumption that the n ordered observations y_i may be regarded as uncorrelated which is in particular for large sample sizes a justifiable assumption. Using this approach, the matrix V^{-1} can be replaced by the $n \times n$-identity matrix I_n. This is equivalent to estimate the slope of the regression line by simple least squares instead of generalized least squares. We will summarize these results in the following definition.

2.20 Definition. *Let $y_1, \dots, y_n$ be a sample of n ordered random observations coming from a random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Further let $S^2 = \sum_{i=1}^n (y_i - \bar{y})^2$. The test statistic denoted by*

$$SF = \frac{(\mathbf{b}'\mathbf{y})^2}{S^2} = \frac{\left(\sum_{i=1}^n b_i y_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \tag{2.37}$$

where

$$\mathbf{b}' = (b_1, \dots, b_n) = \frac{\mathbf{m}'}{(\mathbf{m}'\mathbf{m})^{\frac{1}{2}}}.$$

is called SF **statistic***. The related testing procedure is called the* **Shapiro-Francia test** *or briefly SF* **test***.*

The coefficients $b_i, i = 1, \dots, n$ for the SF statistic are much less complicated to calculate than the entries of the vector $\mathbf{a}$ for the W statistic. Furthermore, since the values of the expected order statistics $\mathbf{m}$ are given for even relatively large sample sizes (cf. section 2.3.1), it is easy to see that SF can be computed for samples of $n > 50$.

2.21 Remark. The SF test can be embedded in definition 2.20 by setting $u_i = m_i, i = 1, \dots, n$. Then, SF is the squared correlation coefficient of the observations y_i and the expected order statistics m_i, i.e., $SF = r^2_{\mathbf{y},\mathbf{m}}$.

Proof: Since $\bar{m} = 0$ (cf. section 2.3.1) and $\mathbf{m}'\mathbf{m} = \sum_{i=1}^n m_i^2$, we can transform the SF

statistic to

$$\begin{aligned} SF &= \frac{\left(\sum_{i=1}^{n} b_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{\left(\sum_{i=1}^{n} m_i y_i\right)^2}{\sum_{i=1}^{n} m_i^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= \frac{\left(\sum_{i=1}^{n}(m_i - \bar{m})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(m_i - \bar{m})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= r^2_{\mathbf{y},\mathbf{m}}. \end{aligned}$$

□

Based on a Monte Carlo study of 1,000 randomly generated values for each sample of size $n = 35, 50, 51(2)99$, Shapiro and Francia (1972) tabled some empirical percentage points of SF. Considering the number of only 1,000 generated values of SF, the accuracy of the percentage points for the SF test may be casted on doubt. To underline this fact, we refer to Pearson et al. (1977) who found differing percentage points for the empirical distribution of SF in their studies using at least m = 50,000 for $n = 99, 100, 125$ as the number of repetitions. In our own simulations study we recalculated the empirical significance points based on m = 1,000,000 repetitions for each sample size. In addition, as values for the expectations of the standard normal order statistics, m_i, we used those given in Parrish (1992a) for sample sizes $n < 100$ that are more exact than the ones used by Shapiro and Francia. Furthermore, we extend the results for smaller samples of size $n = 10(5)35$ and for larger samples of sizes $n = 100, 200$ using the values of m_i given in Harter and Balakrishnan (1996). The approach for this empirical simulation is analogous as the one described in subsection 2.2.2, where the empirical percentage points of the W test are calculated. The results are given in table 2.2.

The critical points of our empirical simulations are all a little bit higher (at least one unit in the second decimal place). Since we reject H_0 if the test statistic is smaller than the critical point, using the new critical points the SF does reject slightly more often the null hypothesis. As a conculsion we notice that, due to the smaller percentage points, the SF test is more conservative using the critical points of Shapiro and Francia and of Pearson et al. and might therefore lead to wrong results. For that reason we recommend the use of the new critical values which is what we did in the empirical studies in section 4.

The Weisberg-Bingham Test

Another test statistic based on the idea of the Shapiro-Francia test is presented by Weisberg and Bingham (1975). With the objective to modify the SF statistic to make it even more suitable for machine calculation, the authors replaced the expected values of the normal order statistics,

Table 2.2: Empirical significance Points for the SF statistic for several sample sizes based on $m =$ 1,000,000 repetitions compared with the critical values given by Shapiro and Francia (1972). Note that the critical points for $n = 100$ are acutally not taken form Shapiro and Francia (1972) but from the work of Pearson et al. (1977).

	own simulation			Shapiro and Francia (1972)		
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
10	0.8781	0.9177	0.9339	—	—	—
15	0.9088	0.9382	0.9502	—	—	—
20	0.9266	0.9503	0.9598	—	—	—
25	0.9389	0.9583	0.9662	—	—	—
30	0.9478	0.9641	0.9708	—	—	—
35	0.9540	0.9685	0.9742	0.919	0.943	0.952
50	0.9662	0.9766	0.9808	0.953	0.954	0.964
100	0.9819	0.9872	0.9894	0.9637	0.9741	0.9786
200	0.9904	0.9931	0.9943	—	—	—

$m_i, i = 1, \dots, n$ by the following approximation:

$$\tilde{m}_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right), \quad i = 1, \dots, n. \tag{2.38}$$

The formula for the $\tilde{m}_i$ suggested by Blom (1958) has already been mentioned in section 2.3.1. Note that from the fact that $\Phi(-x) = 1 - \Phi(x), x \in \mathbb{R}$, it follows for the quantil function $\Phi^{-1}(z)$ that $\Phi^{-1}(z) = -\Phi^{-1}(1 - z), z \in (0, 1)$. Then, for arbitrary $i = 1, \dots, n$ we get

$$\begin{aligned}
\tilde{m}_i &= \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right) \\
&= -\Phi^{-1}\left(1 - \frac{i - 0.375}{n + 0.25}\right) \\
&= -\Phi^{-1}\left(\frac{n + 0.25 - i + 0.375}{n + 0.25}\right) \\
&= -\Phi^{-1}\left(\frac{n - i + 1 - 0.375}{n + 0.25}\right) \\
&= -\tilde{m}_{n-i+1}.
\end{aligned}$$

That is, we have for the $\tilde{m}_i$ the same property than for the m_i, viz.,

$$\tilde{m}_i = -\tilde{m}_{n-i+1}, \quad i = 1, \dots, n. \tag{2.39}$$

2.22 Definition. *Let $y_1, \dots, y_n$ be a sample of ordered random observations coming from a random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Further let S^2*

be defined as like in definition 2.20 and let $\tilde{\mathbf{m}}' = (\tilde{m}_1, \dots, \tilde{m}_n)$*, whose entries are defined in equation* (2.38)*. The test statistic for testing the composite hypothesis of normality denoted by*

$$WB = \frac{(\mathbf{c}'\mathbf{y})^2}{S^2} = \frac{\left(\sum_{i=1}^n c_i y_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \tag{2.40}$$

where

$$\mathbf{c}' = (c_1, \dots, c_n) = \frac{\tilde{\mathbf{m}}'}{(\tilde{\mathbf{m}}'\tilde{\mathbf{m}})^{\frac{1}{2}}}$$

is called WB **statistic***; the testing procedure itself we will call the* **Weisberg-Bingham test** *or briefly* WB **test***.*

2.23 Remark. For $u_i = \tilde{m}_i, i = 1, \dots, n$, the WB statistic is equivalent with the squared correlation coefficient of the ordered observations y_i and the values of $\tilde{m}_i$, i.e., $WB = r^2_{\mathbf{y},\tilde{\mathbf{m}}}$.

Proof: The proof is analogous to the proof of remark 2.21. □

The advantage of WB over SF is that for the computation of the test statistic, no computer storage of any constants like the values of $\mathbf{m}$ are needed. Weisberg and Bingham (1975) conducted a Monte Carlo study with 1,000 randomly generated normal samples for $n = 5, 20, 35$ and computed the values of SF and WB for each sample. The empirical percentage points of SF and WB showed only very slight differences which leads to the consequence that the Weisberg-Bingham statistic is an adequate approximation of the Shapiro-Francia statistic and thus, of the Shapiro-Wilk statistic. Once again, we recalculated these values since the number of repetitions seemed not satisfylingliy large enough to us. Instead, we used $m =$ 1,000,000 for the number of repetitions of the Monte Carlo study. In addition, we extended our results also to samples of size $n \leq 500$ which can be seen in table 2.3.

In opposite to the results for the critical points of the Shapiro-Francia test, we observe that the given critical values are slightly higher than the values from our own simualtions. However the differences are again very small. Nevertheless we found our own simulations more accurate and used them in the power studies in chapter 4.

The Filliben Test

The main disadvantage of using the expected mean $m_i, i = 1, \dots, n$ of the order statistics is probably their difficult and time-consuming calculation. Filliben (1975) tried to overcome this problem by using not the expected mean, but the median of the i-th order statistic y_i as a measure of location. In the follow, let $M_i = \text{med}(y_i), i = 1, \dots, n$ be the median of the i-th order statistic.

2.24 Definition. *Let* $y_1, \dots, y_n$ *be a sample of ordered random observations coming from a random variable* Y *with unknown mean* $\mu \in \mathbb{R}$ *and unknown variance* $\sigma^2 > 0$*. Further let*

Table 2.3: Empirical significance Points for the WB statistic for several sample sizes based on $m =$ 1,000,000 repetitions compared with the critical values given by Weisberg and Bingham (1975)

	own simulation			Weisberg and Bingham (1975)		
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
10	0.7715	0.8427	0.8724	—	—	—
15	0.8256	0.8806	0.9031	—	—	—
20	0.8595	0.9034	0.9214	0.866	0.903	0.920
25	0.8817	0.9185	0.9336	—	—	—
30	0.8978	0.9295	0.9424	0.915	0.941	0.050
35	0.9098	0.9377	0.9491	—	—	—
50	0.9331	0.9537	0.9620	—	—	—
100	0.9639	0.9745	0.9789	—	—	—
200	0.9808	0.9863	0.9885	—	—	—
500	0.9919	0.9941	0.9950	—	—	—

$\mathbf{M} = (M_1, \ldots, M_n)'$ *be the vector of the medians of the observation vector* $\mathbf{y}$. *The test statistic denoted by*

$$FB = \frac{\mathbf{d}'\mathbf{y}}{S} = \frac{\sum_{i=1}^n d_i y_i}{(\sum_{i=1}^n (y_i - \bar{y})^2)^{\frac{1}{2}}}, \tag{2.41}$$

where

$$\mathbf{d}' = (d_1, \ldots, d_n) = \frac{\mathbf{M}'}{(\mathbf{M}'\mathbf{M})^{\frac{1}{2}}}$$

is called FB statistic. The related testing procedure is called **Filliben test** *or briefly FB* **test***.*

2.25 Remark. By setting $u_i = M_i, i = 1, \ldots, n$, the test statistic of the FI test turns out to be the empirical correlation coefficient of the ordered observations y_i and the median M_i, i.e. $FI = r_{\mathbf{y},\mathbf{M}}$.

Proof: According to Filliben (1975), $M_i = -M_{n-i+1}, i = 1, \ldots, n$, hence $\sum_{i=1}^n M_i = 0$. The rest of the proof is straightforward:

$$\begin{aligned}
\frac{\sum_{i=1}^n d_i y_i}{(\sum_{i=1}^n (y_i - \bar{y})^2)^{\frac{1}{2}}} &= \frac{\sum_{i=1}^n y_i M_i}{(\sum_{i=1}^n M_i^2)^{1/2} (\sum_{i=1}^n (y_i - \bar{y})^2)^{1/2}} \\
&= \frac{\sum_{i=1}^n y_i M_i - \bar{y} \sum_{i=1}^n M_i}{(\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n M_i^2)^{1/2}} \\
&= \frac{\sum_{i=1}^n (y_i - \bar{y}) M_i}{(\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n M_i^2)^{1/2}} \\
&= \frac{\sum_{i=1}^n (y_i - \bar{y})(M_i - \mathbf{M})}{(\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (M_i - \mathbf{M})^2)^{1/2}} = r_{\mathbf{y},\mathbf{M}}
\end{aligned}$$

□

To conduct the FB test, knowledge about the entries of the coefficient vector $\mathbf{d}$ is required. In order to do so, Filliben did not calculate the M_i directly. Instead he approximated the medians $\tilde{M}_i$ of the order statistics from a uniform distribution on $(0, 1)$. The relation between the uniform and the standard normal distribution has already been mentioned in subsection 2.3.1: if X is standard normally distributed, then $\Phi(X)$ is uniformly distributed on $(0, 1)$. Conversely, for a uniformly distributed random variable U we have that $\Phi^{-1}(U)$ is standard normally distributed Falk et al. (see 2002, p. 32). Hence, if the medians $\tilde{M}_i, i = 1, \dots, n$ of the order statistics from a uniformly distributed sample are given, one immediately obtains the corresponding medians for the $N(0, 1)$-distribution by calculating $M_i = \Phi^{-1}(\tilde{M}_i), i = 1, \dots, n$. For the approximation of the order statistics $\tilde{M}_i$, Filliben suggested the following formula:

$$\tilde{M}_i = \begin{cases} 1 - 0.5^{1/n} & \text{for} \quad i = 1 \\ (i - 0.3175)/(n + 0.365) & \text{for} \quad i = 2, \dots, n-1 \\ 0.5^{1/n} & \text{for} \quad i = n. \end{cases} \tag{2.42}$$

Note that for $i = 2, \dots, n-1$ the recommended approximation in 2.42 is the general formular in (2.30) proposed by Blom (1958) for the value $\eta = 0.3175$.

The mandatory Monte Carlo simlation study to obtain empirical significance points for the FB test was performed by Filliben for several smaple sizes where the number of replications m was 10,000 each time. Since m seems to be too small to have a satisfying accuracy concering the percentage points, we recalculated the Monte Carlo study with m = 1,000,000. However, Filliben did not give critical values for $n > 100$. We extended the simulation of larger sample sizes up to $n = 500$ for the power study in chapter 4. The results of our study are given in table 2.4.

Comparing the critical points of the two studies, there is no clear trend: sometimes the values of Filliben are slightliy higher, sometimes the values of our simulation study are higher. What the differences have in common is that there are almost negligible small. This argues for the accuracy of the simulation of Filliben.

2.4.2 Extensions of the Procedure to Large Sample Sizes

Many power studies with different testing procedures for normality and a various set of alternative distributions showed, that the Shapiro-Wilk test is probability one of the most powerful omnibus tests (cf. section 4.2). Obviously, the greatest disadvantage of the W statistic is that it is only applicable for sample sizes of $n \leq 50$. The tests presented in section 2.4 tried to simplify the calculation of the test statistic and to extend the Shapiro-Wilk test for larger sample sizes. However, there was no testing procedure that performed as good as the Shapiro-Wilk test or even better for all kind of alternative distributions in the power studies. In addition, the exten-

Table 2.4: Empirical significance Points for the FB statistic for several sample sizes based on $m =$ 1,000,000 repetitions compared with the critical values given by Filliben (1975)

	own simulation			Filliben (1975)		
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
10	0.8771	0.9171	0.9336	0.876	0.917	0.934
15	0.9075	0.9377	0.9499	0.907	0.937	0.950
20	0.9261	0.9499	0.9596	0.925	0.950	0.960
25	0.9381	0.9580	0.9660	0.937	0.958	0.966
30	0.9468	0.9637	0.9706	0.947	0.964	0.970
35	0.9533	0.9680	0.9740	0.952	0.968	0.974
50	0.9656	0.9763	0.9806	0.965	0.977	0.981
100	0.9815	0.9870	0.9893	0.981	0.987	0.989
200	0.9902	0.9930	0.9942	—	—	—
500	0.9959	0.9970	0.9975	—	—	—

sions of the SF test, the WB test or the FI test concerned only samples up to a size of 100. However, we extended the use of the testing procedure to samples of size $n \geq 100$ by presenting empirical percentage points for the tests in section 2.4. But before, these extensions have not been considered in the past by the best knowledge of the author.

The work of Royston (1992) is the probably most popular attempt to overcome this essential drawback. The underlying idea of the author consists of using a normalizing transformation for the W statistic to extend the use of this testing procedure. Another extension developed by Rahman and Govindarajulu (1997) will also be presented in this section.

The Extension of Royston

The normalizing transformation is based on a new approximation of the entries of the vector $\mathbf{a}$ in equation (2.27). This approximation has already been presented in section 2.3.2, where first the coefficients a_n and a_{n-1} were approximated by regression analysis in equations (2.32) and (2.33). The formulas for the rest of the coefficients of $\mathbf{a}$ are given in (2.34) and in (2.35). In the following, let $\tilde{\mathbf{a}} = (\tilde{a}_1, \ldots, \tilde{a}_n)'$ be the vector of approximations for $\mathbf{a}$ based on this calculation. With the new vector $\tilde{\mathbf{a}}$ we can now define the test statistic

$$W_R = \frac{(\tilde{\mathbf{a}}'\mathbf{y})^2}{S^2} = \frac{\left(\sum_{i=1}^n \tilde{a}_i y_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{2.43}$$

that is used for a normalizing transformation to get approximative significance levels of W_R. By the expression **normalizing transformation** or **normalization** of a random variable T, we mean a function $g(\cdot)$ with the property that $g(T)$ is standard normally distributed.

According to Royston, for samples of size $4 \leq n \leq 11$, the distribution of $\log(1 - W_R)$ can be approximated by a three-parameter lognormal distribution $\Lambda(\gamma, \mu, \sigma^2)$ with probability density function

$$f_{\log(1-W_R)}(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(y-\gamma)} \exp\Big(- \frac{(\mu - \log(y-\gamma))^2}{2\sigma^2}\Big) & , \quad y > \gamma, \\ 0 & , \quad \text{else.} \end{cases}$$

The parameters γ, μ and σ of the distribution were all obtained by estimation based on empirical simulation. Equations as the result of a smoothing for the parameters will be given in definition 2.26. For samples of size $12 \leq n \leq 2{,}000$, Royston found that the distribution of $1 - W_R$ can be fitted by a two-parameter lognormal distribution $\Lambda(\mu, \sigma^2)$ with the pdf

$$f_{1-W_R}(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma y} \exp\big(- \frac{(\log(1-W)-\mu)^2}{2\sigma^2}\big) & , \quad y > 0, \\ 0 & , \quad \text{else,} \end{cases}$$

where again empirical simulation was used to get the two parameters μ and σ^2. The smoothed values are presented in the next definition.

2.26 Definition. *Let $y_1, \ldots, y_n$ be a sample of ordered random observations coming from a random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. For $4 \leq n \leq 11$ define*

$$\gamma = -2.273 + 0.459n$$
$$\mu = 0.544 - 0.39978n + 0.025054n^2 - 0.0006714n^3$$
$$\sigma = \exp(1.3822 - 0.77875n + 0.062767n^2 - 0.0020322n^3)$$

and for $12 \leq n \leq 2{,}000$ let

$$\mu = -1.5861 - 0.31082\log(n) - 0.083751(\log(n))^2 + 0.0038915(\log(n))^3$$
$$\sigma = \exp(-0.4803 - 0.082676\log(n) + 0.0030302(\log(n))^2).$$

The normalization of the test statistic W_R of Royston (1992) given by

$$z_{W_R} = \begin{cases} \frac{-\log[\gamma - \log(1-W_R)] - \mu}{\sigma} & , \quad 4 \leq n \leq 11, \\ \frac{\log(1-W_R)-\mu}{\sigma} & , \quad 12 \leq n \leq 2{,}000. \end{cases} \tag{2.44}$$

is called **extension of Royston** *and the test statistic we will call z_{W_R}* **statistic**.

Following Royston, since z_{W_R} is approximately standard normal distributed the confidence interval I of the test to the significance level α is (see Falk et al. (2002))

$$I = \big(\Phi^{-1}(\alpha/2), \Phi^{-1}(1 - (\alpha/2))\big).$$

Despite of the many other trials to normalize the W statistic to extend the testing procedure and to get p-values eays to calculate, the normalizing transformation of Royston given in defintion

2.26 has become widely accepted in literature and in the applied statistics where testing for normality for large sample sizes is an essential issue. The main reason for this fact is that the testing procedure of Royston has really high power and performes very well in many different situations of the underlying alternative distribution (see section 4.2 for more details). Almost all popular statistical software packages like R (see R Development Core Team (2009)) and SAS (see SAS Institute Inc. (2003)) have today implemented the procedure of Royston as a test for normality based on the idea of Shapiro and Wilk. Thus it is an interesting fact to know that when performing the Shapiro-Wilk test for example in SAS, one is actually doing the extension of Royston, even if the sample size is smaller than 50 and the Shapiro-Wilk test actually could be performed!

The Extension of Rahman and Govindarajulu

As already mentioned in subsection 2.3.1, the approach of Rahman and Govindarajulu (1997) was based on the matrix V^{-1} given in equation (2.31). Recall in mind that the determination of the coefficients a_i of the W statistic can be done by calculating

$$a_i = \sum_{j=1}^{n} m_j \frac{v_{ji}^*}{\psi}, \quad i = 1, \dots, n, \tag{2.45}$$

where $v_{ij}^*, i, j = 1, \dots, n$ are the entries of V^{-1} and ψ is defined in lemma 2.13. Using the entries of the matrix in equation (2.31) for the v_{ij}^*, equation (2.45) can be reduced to

$$a_i = -\frac{1}{\psi}(n+1)(n+2)\varphi(m_i)\big[m_{i-1}\varphi(m_{i-1}) - 2m_i\varphi(m_i) + m_{i+1}\varphi(m_{i+1})\big], \quad i = 1, \dots, n,$$

by setting

$$m_0\varphi(m_0) = m_{n+1}\varphi(m_{n+1}) = 0.$$

As an appropriate approximation for the expected order statistics, $m_i, i = 1, \dots, n$ the authors used the approximation formula of Blom (1958) given in equation (2.38) for their test statistic.

2.27 Definition. *Let $y_1, \dots, y_n$ be a sample of ordered random observations coming from a random variable Y with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Further, let S_n^2 be defined like in definition 2.20 and*

$$g_i = -\frac{1}{\tilde{\psi}}(n+1)(n+2)\varphi(\tilde{m}_i)\big[\tilde{m}_{i-1}\varphi(\tilde{m}_{i-1}) - 2\tilde{m}_i\varphi(\tilde{m}_i) + \tilde{m}_{i+1}\varphi(\tilde{m}_{i+1})\big], \quad i = 1, \dots, n,$$

where

$$\tilde{\psi} = \tilde{\mathbf{m}}' V^{-1} V^{-1} \tilde{\mathbf{m}} = (\tilde{m}_1, \dots, \tilde{m}_n) V^{-1} V^{-1} (\tilde{m}_1, \dots, \tilde{m}_n)',$$

where the matrix V^{-1} is given in equation (2.31) and $\tilde{m}_0\varphi(\tilde{m}_0) = \tilde{m}_{n+1}\varphi(\tilde{m}_{n+1}) = 0$ and the $\tilde{m}_i, i = 1, \dots, n$ are given in equation (2.38). The test statistic denoted by

$$W_{RG} = \frac{(\mathbf{g}'\mathbf{y})^2}{S^2} = \frac{\big(\sum_{i=1}^n g_i y_i\big)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where

$$\mathbf{g}' = (g_1, \ldots, g_n)$$

is called **extension of Rahman and Govindarajulu**; *the test statistic will be briefly called* W_{RG} **statistic**.

For the purpose of comparing their approximations $g_i, i = 1, \ldots, n$ of the vector of coefficients $\mathbf{a}$ of the W statistic, the authors computed the differences of their approximations with the exact values of a_i for samples of size 10, 20 and 30. Doing so, they found that the error is very small from which they followed an appropriate accuracy for their values g_i.

In order to get percentage points of the distribution of W_{RG} under the null hypothesis, Rahman and Govindarajulu conducted a Monte Carlo study based 20,000 empirical simulations for several sample sizes up to 5,000. We recacultated these critical points in order to make them even more exact based on an empirical simulation study of m = 1,000,000 repetitions. Our results are given in table 2.5.

Table 2.5: Empirical significance Points for the W_{RG} statistic for several sample sizes based on m = 1,000,000 repetitions compared with the critical values given by Rahman and Govindarajulu (1997)

	own simulation			Rahman and Govindarajulu (1997)		
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
10	0.7812	0.8298	0.8531	0.779	0.828	0.852
15	0.8260	0.8635	0.8813	0.828	0.863	0.881
20	0.8542	0.8849	0.8993	0.853	0.884	0.899
25	0.8736	0.8998	0.9120	0.873	0.899	0.911
30	0.8878	0.9108	0.9215	0.887	0.911	0.921
35	0.8991	0.9196	0.9290	0.899	0.919	0.928
50	0.9216	0.9369	0.9440	0.921	0.937	0.944
100	0.9533	0.9619	0.9659	0.953	0.961	0.965
200	0.9731	0.9777	0.9799	0.973	0.977	0.979
500	0.9875	0.9895	0.9904	0.987	0.989	0.990

We see that the cricital points from our simluations are mostly slightly higher. However the differences to the empirical percentage points of Rahman and Govindarajulu appear only in the third decimal place which underlines the accuracy of their results. Though, for the power studies in chapter 4 we preferred our own simulation results.

Chapter 3

Moment Tests for Normality

A completely different approach for the investigation of the question whether a sample is normal distributed or not, consists in the use of moments of a distribution or, more exactly spoken, the use of sample moments. Mostly the third and the fourth standardized moment are considered to make a decision about the null hypothesis of normal distributed observations.

In this chapter, we will give a short introduction into this field of testing for normality beginning with some theoretical aspects of moments of random variables together with a justification for their use in testing for normality. After that, we present the two simplest tests for normality, namely the skewness and the kurtosis test. In the third section, approaches of combining these two tests in form of an omnibus test will be mentioned. Since we focus on the Jarque-Bera test as one of the most popular omnibus tests in applied statistics, in the fourth section we give the test statistic as well as the underlying idea of the test. At the end we present an intuitive modifications of the Jarque-Bera test. Additionally, for the Jarque-Bera test and its modification, new percentage points based on extensive empirical simulation studies are tabled in this section.

On the opposite to the mathematical notation in chapter 2, if we have a sample $y_1, \ldots, y_n$ of iid random observations coming from a random variable Y, the n observations do not necessarily have to be ordered.

3.1 Motivation of Using Moment Tests

The first one wuo realized that deviations from normality can be characterized by the standard third and fourth moment, was Karl Pearson (cf. Thode (2002, p. 41)). Hence, the theory of testing for normality can be regarded as having been initialized by Pearson who had a major impact on the developement of this theory and their mathematical methods. To become more detailed, we start with the definition of moments:

3.1 Definition. *Consider a random sample Y with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Further,* *let $p_Y(y)$ be the pdf of Y and k be an integer greater than 1. The* **k-th standardized moment** *of Y is defined as*

$$\psi_k(Y) = \frac{\mu_k(Y)}{(\mu_2(Y))^{k/2}}, \tag{3.1}$$

where

$$\mu_k(Y) = E((Y-\mu)^k) = \int_{-\infty}^{+\infty} (y-\mu)^k p_Y(y)\,\mathrm{d}y$$

is called the **k-th central moment** *of Y. In particular, $\mu_2 = E((Y-\mu)^2) = \sigma^2$.*

3.2 Remark. The k-th standardized moment is scale and translation invariant for k being an integer greater 1.

Proof: Let $k \geq 2$ be an arbitrary integer and let $\alpha, \beta \in \mathbb{R}$ be two non-stochastic constants. Further, let Y be a random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Consider the random variable $Z = \alpha Y + \beta$ that emerges from a scale and location transformation. Obviously, the mean of Z is $E(\alpha Y + \beta) = \alpha E(Y) + \beta = \alpha\mu + \beta$, due to the linearity of the expectation operator. The k-th standardized moment of Z, $\psi_k(Z)$, is then

$$\begin{aligned}
\psi_k(Z) &= \frac{E((\alpha Y + \beta - (\alpha\mu + \beta))^k)}{(E((\alpha Y + \beta - (\alpha\mu + \beta))^2))^{k/2}} \\
&= \frac{E((\alpha(Y-\mu))^k)}{(E((\alpha(Y-\mu))^2))^{k/2}} \\
&= \frac{\alpha^k E((Y-\mu)^k)}{(\alpha^2 E((Y-\mu)^2))^{k/2}} \\
&= \frac{\alpha^k E((Y-\mu)^k)}{\alpha^k (E((Y-\mu)^2))^{k/2}} = \psi_k(Y),
\end{aligned}$$

where again the linearity of the expectation operator is used to place α to the power of k and 2, respectively, outside the expectation. □

The property of invariance against scale and location transformations is a very important fact, since we want to compare the shape of an empirical distribution to that of a theoretical distribution, namely the normal distribution. Hence, when using standardized moments of the normal distribution to decide whether a random sample is normally distributed or not, we can constrain our considerations on investigations of the standard normal distribution, like in the power studies in section 4.2.

In this chapter, our main interest will focus on the third and the fourth standardized moment for which we will introduce a special notation.

3.3 Definition. *The the third standardized moment of a random variable Y is called* **skewness** *and is denoted as*

$$\sqrt{\beta_1} = \psi_3(Y) = \frac{E((Y-\mu)^3)}{(E((Y-\mu)^2))^{3/2}}.$$

The fourth standardized moment of a random variable Y is called **kurtosis** *and is denoted as*

$$\beta_2 = \psi_4(Y) = \frac{E((Y-\mu)^4)}{(E((Y-\mu))^2)^2}.$$

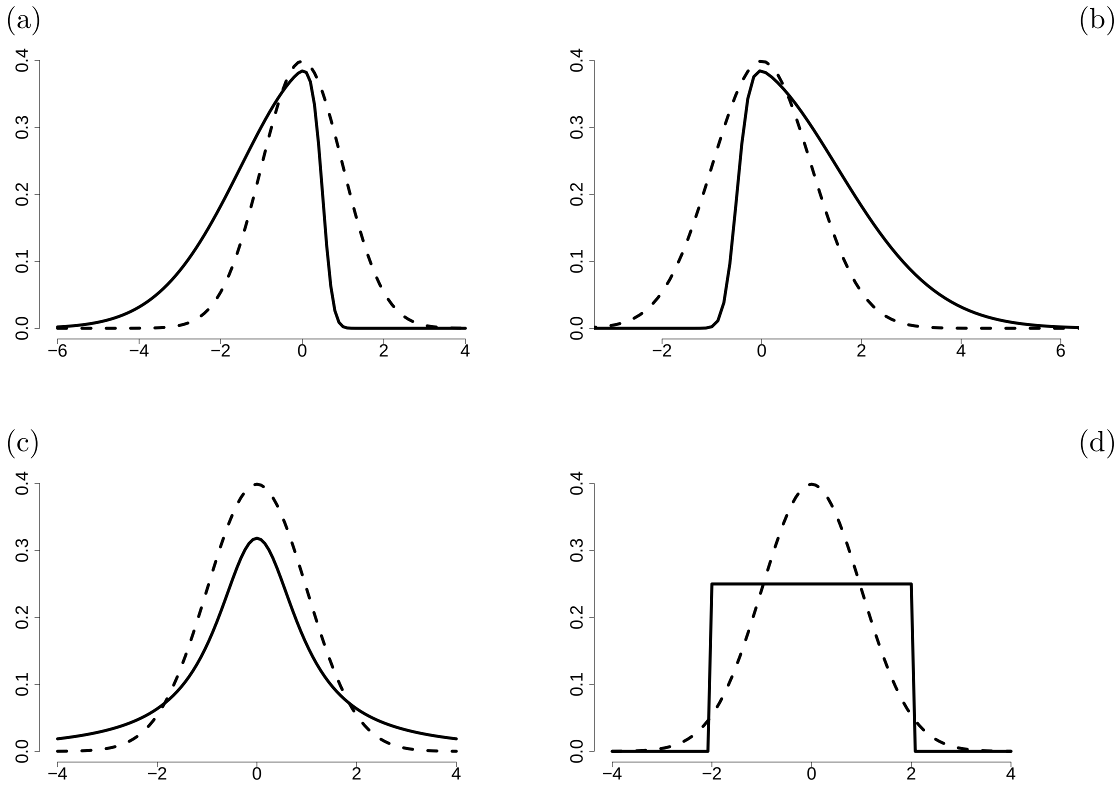


Figure 3.1: Nonnormal probability densitiy functions due to deviations in either the skewness or the kutosis value. The dashed line is the pdf of a standard normal distribution. The solid lines are (a) a left skewed distribution ($\sqrt{\beta_1} < 0$), (b) a right skewed distribution ($\sqrt{\beta_1} > 0$), (c) a heavy tailed distribution ($\beta_2 > 3$) and (d) a light tailed distribution ($\beta_2 < 3$).

The notation $\sqrt{\beta_1}$ and β_2 might seem to be a little curious, however it is the classical notation generally used in literature. Note that by writing $\sqrt{\beta_1}$ and β_2, the dependence of skewness and kurtosis of their associated random variable is omitted by the sake of simplicity. To which random variable $\sqrt{\beta_1}$ and β_2 belong will become clear from the context. A basic result of testing for normality using moments is the following:

3.4 Lemma. *Let X be a standard normally distributed random variable. Then for the skewness and kurtosis we have, respectively,*

$$\sqrt{\beta_1} = 0 \quad \textit{and} \quad \beta_2 = 3.$$

Proof: Since X is standard normally distributed, it follows that its mean μ is zero and for the variance σ^2 we have

$$\sigma^2 = E((X-\mu)^2) = E((X)^2) = \int_{-\infty}^{+\infty} x^2 \varphi(x)\,\mathrm{d}x = 1. \tag{3.2}$$

Thus,

$$\sqrt{\beta_1} = \frac{E((X-\mu)^3)}{(E((X-\mu)^2))^{3/2}} = E(X^3) = \int_{-\infty}^{+\infty} x^3 \varphi(x)\,\mathrm{d}x.$$

Since $(-x^3)\varphi(-x) = -x^3\varphi(x)$ for arbitrary x, the integrand of the integral is odd. It follows that the value of the integral is 0, which is the first part of the proof. To show the second result, we use partial integration. Since $\partial\varphi(x)/\partial x = -x\varphi(x)$, we find

$$\beta_2 = \int_{-\infty}^{+\infty} x^4\varphi(x)\,\mathrm{d}x = \underbrace{-x^3\varphi(x)\Big|_{x=-\infty}^{x=+\infty}}_{=0} + 3\underbrace{\int_{-\infty}^{+\infty} x^2\varphi(x)\,\mathrm{d}x}_{=1}.$$

The second summand is 3 because of equation 3.2. The first summand is 0 because of the fact that the exponential function converges faster than any other function and hence, as the cubic function in particular. □

In addition to the mean and the standard deviation, the skewness and the kurtosis are important quantities to characterize the shape of the distribution function of a random variable. The skewness is a measure of asymmetry. Random variables with $\sqrt{\beta_1} < 0$ are called **skewed to the left**, i.e., the long tail of their pdf goes off to the left, whereas random variables with $\sqrt{\beta_1} > 0$ are called **skewed to the right**, that is the long tail of the pdf is to the right. The kurtosis is an indicator of tail thickness. If $\beta_2 > 3$ for a random variable the pdf has "thicker" tails than normal tails which is called **heavy tailed** (or **long tailed**). Else, for $\beta_2 < 3$ the pdf of the associated random variable is called **light tailed** (or **short tailed**), since it has "thinner" tails than those of the normal distribtion. See Falk et al. (2002) for more details. Figure 3.1 tries to visualize the different types of shape for nonnormal probability density functions compared with the shape of the standard normal distributions.

In order to investigate the null hypothesis for a given random sample $y_1, \ldots, y_n$ of n observations, we have to add to definition 3.1 of skewness and kurtosis the definition of their empirical counterparts.

3.5 Definition. *Let $y_1, \ldots, y_n$ be a sample of iid random observations coming from a random variable Y with pdf $p_Y(y)$. For an integer $k > 1$, the* **k-th standardized sample moment** *of Y is defined as*

$$g_k = \frac{m_k}{m_2^{k/2}},$$

where

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k$$

is the **k-th central sample moment** *of Y.*

3.6 Remark. Like the k-th standardized moment in equation 3.1, the k-th standardized sample moment is scale and translation invariant for integers $k > 1$. The proof of this fact is simple calculus which is why we will leave the details of this proof to the reader.

As already mentioned above, the third and the fourth standardized moments are playing a crucial role in the testing theory for normality. Therefore we also introduce a special notation for their empirical counterparts.

3.7 Definition. *Let $y_1, \ldots, y_n$ be a sample of iid random observations coming from a random variable Y. The third standardized sample moment of $y_1, \ldots, y_n$ is called* **sample skewness** *and is denoted by*

$$\sqrt{b_1} = g_3 = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^3}{(\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2)^{3/2}}.$$

The fourth standardized sample moment of $y_1, \ldots, y_n$ is called **sample kurtosis** *and is denoted by*

$$b_2 = g_4 = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^4}{(\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2)^2}.$$

3.8 Remark. The sample skewness $\sqrt{b_1}$ and the sample kurtosis b_2 are consistent estimators for the theoretical skewness $\sqrt{\beta_1}$ and the kurtosis β_2, respectively, i.e., they converge in probability to their corresponding theoretical value:

$$\sqrt{b_1} \xrightarrow{P} \sqrt{\beta_1} \quad \text{and} \quad b_2 \xrightarrow{P} \beta_2.$$

For a definition of convergence in probability, see definition A.8 in the appendix. For a proof of this assertion we refer to Lehmann (1999, p. 51).

The notation $\sqrt{b_1}$ and b_2 is like for the theoretical skewness and kurtosis the classical popular notation. Under the null hypothesis that the random observations $y_1, \ldots, y_n$ are coming from a normal population, following the results of lemma 3.4, we expect the value of $\sqrt{b_1}$ to be approximately near 0 and the value of b_2 to be approximately near 3. Deviations of $\sqrt{b_1}$ and b_2 from their expected values under the hypothesis of normality is of course an indication that the observations $y_i, i = 1, \ldots, n$ are not coming form a normal distributed population. This is why, in order to judge departures from normality with moment tests, the empirical counterparts of the standard third and fourth moment have to be considered.

3.2 Shape Tests

3.2.1 The Skewness Test

Like before, consider a random variable Y with unkown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Further, let $\sqrt{\beta_1}$ denote the skewness of the probability densitiy function of Y like in definition 3.3. The skewness test investigates the following null hypothesis:

$$H_0 : \sqrt{\beta_1} = 0 \quad \text{vs.} \quad H_1 : \sqrt{\beta_1} \neq 0. \tag{3.3}$$

Note that not rejecting this null hypothesis is not equivalent in not rejecting the null hypothesis that the sample is normally distributed. The reason for that is that there are distributions that are symmetric, i.e., the skewness of their pdf is zero, but have higher or lower kurtosis than the normal distribution. For example, for the t-distribution with 5 degrees of freedom, t_5, we have $\beta_2 = 9$ (see subsection 4.2.2 for a definition of the t distribution). Hence, the skewness test can only protect against asymmetry in the data but not against heavy and light tails. Two other examples of symmetric distributions with a kurtosis β_2 different from 3 are presented in figure 3.2.

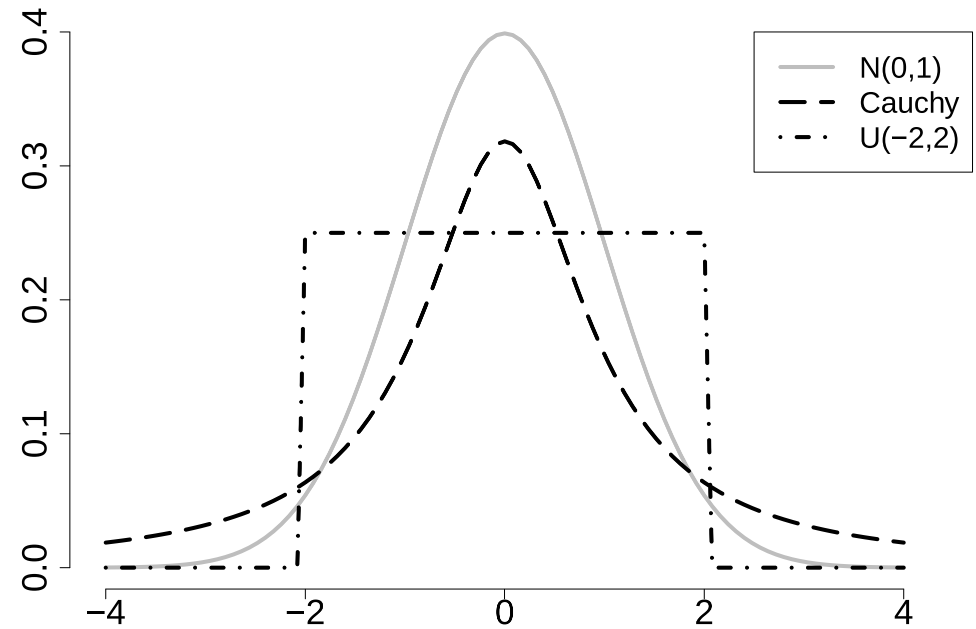


Figure 3.2: Distributions with a skewness $\sqrt{\beta_1} = 0$ that are not normally distributed and the pdf of the standard normal distribution. The kurtosis of the Cauchy distribution ($= t_1$ distribution) is ∞ and the kurtosis of the $U(-2, 2)$ distribution is 1.8

Carrying forward the arguments of the preceeding subsection, the reader can expect intuitively that by examining a random sample on normality with the sample skewness, one just have to look at the value of $\sqrt{b_1}$. Nevertheless, we start this subsection with the formal definition of the test.

3.9 Definition. *Let $y_1, \ldots, y_n$ be a sample of iid random observations coming from a random* *variable Y. The test statistic of the* **skewness test** *is denoted by*

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})2)^{3/2}}.$$

The test statistic $\sqrt{b_1}$ is mostly used as a two-sided statistic, since nonnormality can arise both in left and right skewed density functions. Because under the null hypothesis, $\sqrt{b_1} \approx 0$, we need a given critical value $c_{n,\alpha/2}$ for a given significance niveau $\alpha \in (0, 1)$ with the consequence of rejecting H_0 in (3.3) if $|\sqrt{b_1}| \geq c_{n,\alpha/2}$. As we shall see later, the distribution of $\sqrt{b_1}$ is symmetric, and therefore, taking the absolute value of $\sqrt{b_1}$ and $\alpha/2$ as a parameter for the determination of the critical value $c_{n,\alpha/2}$ are justifiable. There is also the possibility of using $\sqrt{b_1}$ as a one-sided statistic regardless of testing for left or right skewed data. Note that in this two cases the null hypothesis in (3.3) has to be modified appropriately. For testing against left skewed data on a significance level α, H_0 is rejected if $\sqrt{b_1} \leq c_{n,\alpha}$. Otherwise, for testing against right skewness in the data, H_0 is rejected if $\sqrt{b_1} \geq c_{n,\alpha}$.

Based on large-sample considerations, it is possible to determine the asymptotic distribution of $\sqrt{b_1}$:

3.10 Lemma. *Let $\sqrt{b_1}$ be the test statistic of the skewness test from definition 3.9 based on a sample $y_1, \ldots, y_n$ of iid random observations coming from a normally distributed random variable Y with unknow mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$. Then, $\sqrt{n}\sqrt{b_1}$ is asymptotically normal distributed with mean 0 and variance 6, i.e.,*

$$\sqrt{n}\sqrt{b_1} \xrightarrow{\mathcal{D}} N(0, 6) \qquad \text{as } n \to \infty. \tag{3.4}$$

See definition A.7 in the appendix for a definition of convergence in distribution.

Proof: Recall in mind that the skewness is scale and translation invariant (see remark 3.6). Therefore, it is enough to show the assertion for a sample $x_1, \ldots, x_n$ of iid random observations coming form a standard normally distributed random variable X. To this end, consider the function $g : \mathbb{R}^3 \to \mathbb{R}$, with

$$g(a, b, c) = \frac{c - 3ab + 2a^2}{(b - a^2)^{3/2}}.$$

Next, we define the following quantities:

$$\tilde{m}_2 := \frac{1}{n}\sum_{i=1}^{n} x_i^2, \qquad \tilde{m}_3 := \frac{1}{n}\sum_{i=1}^{n} x_i^3. \tag{3.5}$$

Together with the sample mean $\bar{x} := \tilde{m}_1$ we then get

$$\begin{aligned}
g(\tilde{m}_1, \tilde{m}_2, \tilde{m}_3) &= \frac{\tilde{m}_3 - 3\tilde{m}_1\tilde{m}_2 + 2\tilde{m}_1^2}{(\tilde{m}_2 - \tilde{m}_1^2)^{3/2}} \\
&= \frac{n^{-1}\sum_{i=1}^n x_i^3 - 3\tilde{m}_1 n^{-1}\sum_{i=1}^n x_i^2 + 3\tilde{m}_1^2 n^{-1}\sum_{i=1}^n x_i - \tilde{m}_1^3}{(n^{-1}\sum_{i=1}^n x_i^2 - 2\tilde{m}_1 n^{-1}\sum_{i=1}^n x_i + \tilde{m}_1^2)^{3/2}} \\
&= \frac{n^{-1}(\sum_{i=1}^n x_i^3 - 3\tilde{m}_1\sum_{i=1}^n x_i^2 + 3\tilde{m}_1^2\sum_{i=1}^n x_i - n\tilde{m}_1^3)}{(n^{-1}(\sum_{i=1}^n x_i^2 - 2\tilde{m}_1\sum_{i=1}^n x_i + n\tilde{m}_1^2))^{3/2}} \\
&= \frac{n^{-1}(\sum_{i=1}^n (x_i^3 - 3\tilde{m}_1 x_i^2 + 3\tilde{m}_1^2 x_i - \tilde{m}_1^3))}{(n^{-1}(\sum_{i=1}^n (x_i^2 - 2\tilde{m}_1 x_i + \tilde{m}_1^2)))^{3/2}} \\
&= \frac{n^{-1}\sum_{i=1}^n (x_i - \bar{x})^3}{(n^{-1}\sum_{i=1}^n (x_i - \bar{x})^2)^{3/2}} \\
&= \sqrt{b_1},
\end{aligned}$$

that is the function g provides exactly the sample skewness for the point $(\tilde{m}_1, \tilde{m}_2, \tilde{m}_3)'$. If we summarize the random variables x_i, x_i^2, x_i^3 to a three-dimensional random vector $\mathbf{x}_i = (x_i, x_i^2, x_i^3)', i = 1, \ldots, n$, we have for the mean vector $\mu = E(x_i, x_i^2, x_i^3)' = (0, 1, 0)'$, since the x_i's are iid standard normally distributed. We can now apply the multivariate central limit theorem (c.f. theorem A.9 in the appendix) and obtain that

$$\sqrt{n}(\tilde{m}_1, \tilde{m}_2 - 1, \tilde{m}_3) \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

where

$$\begin{aligned}
\Sigma &= E\left(\begin{pmatrix} x_i \\ x_i^2 - 1 \\ x_i^3 \end{pmatrix}(x_i, x_i^2 - 1, x_i^3)\right) \\
&= E\begin{pmatrix} x_i x_i & x_i(x_i^2 - 1) & x_i x_i^3 \\ x_i(x_i^2 - 1) & (x_i^2 - 1)^2 & x_i^3(x_i^2 - 1) \\ x_i x_i^3 & x_i^3(x_i^2 - 1) & x_i^3 x_i^3 \end{pmatrix} \\
&= E\begin{pmatrix} x_i^2 & x_i^3 - x_i & x_i^4 \\ x_i^3 - x_i & x_i^4 - 2x_i^2 + 1 & x_i^5 - x_i^3 \\ x_i^4 & x_i^5 - x_i^3 & x_i^6 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & 3 \\ 0 & 2 & 0 \\ 3 & 0 & 15 \end{pmatrix},
\end{aligned}$$

since $E(x_i) = E(x_i^3) = E(x_i^5) = 0$ and $E(x_i^2) = 2,\ E(x_i^4) = 3,\ E(x_i^6) = 15$ (see Stuart and Ord (1987, p. 78)). Moreover, the value $g(0, 1, 0) = 0$ is ecactly the skewness of the standard normally distributed variable X. In order to use the delta method, the gradient ∇g (vector of the first partial derivates) of g has to be determined. The partial derivates of g are:

$$
\begin{aligned}
\frac{\partial g(a,b,c)}{\partial a} &= \frac{(-3b+6a^2)(b-a^2)^{3/2} - 3/2(c-3ab+2a^3)(b-a^2)^{1/2}(-2a)}{(b-a^2)^3} \\
&= \frac{3(ac-b^2)}{(b-a^2)^{5/2}}, \\
\frac{\partial g(a,b,c)}{\partial b} &= \frac{-3a(b-a^2)^{3/2} - 3/2(c-3ab+2a)(b-a)^{1/2}}{(b-a^2)^3} \\
&= \frac{3a(a-1)+3/2(ab-c)}{(b-a^2)^{5/2}}, \\
\frac{\partial g(a,b,c)}{\partial c} &= (b-a^2)^{-3/2}.
\end{aligned}
$$

Consequently, the gradient is given by

$$
\nabla g(a,b,c) = (b-a^2)^{-3/2} \begin{pmatrix} \frac{3(ac-b^2)}{b-a^2} \\ \frac{3a(a-1)+3/2(ab-c)}{b-a^2} \\ 1 \end{pmatrix}.
$$

Thus, the gradient ∇g at the point $(0,1,0)'$ is exactly $(-3,0,1)'$. Because g is obviously differentiable at the point $(0,1,0)'$, all assumptions of theorem A.10 in the appendix are fulfilled. Hence, it follows that

$$
\sqrt{n}(g(\tilde{m}_1,\tilde{m}_2,\tilde{m}_3) - g(0,1,0)) = \sqrt{n}\sqrt{b_1} \xrightarrow{\mathcal{D}} N\left(0, (\nabla g(0,1,0))'\,\Sigma\,\nabla g(0,1,0)\right).
$$

For the variance of this normal distribution we find

$$
(\nabla g(0,1,0))'\,\Sigma\,\nabla g(0,1,0) = (-3,0,1)\begin{pmatrix} 1 & 0 & 3 \\ 0 & 2 & 0 \\ 3 & 0 & 15 \end{pmatrix}\begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix} = (-3,0,1)\begin{pmatrix} 0 \\ 0 \\ 6 \end{pmatrix} = 6.
$$

This finally yields to

$$
\sqrt{n}\sqrt{b_1} \xrightarrow{\mathcal{D}} N(0,6) \qquad \text{as } n \to \infty.
$$

□

The task of giving a critical value $c_{n,\alpha}$ for a significance level $\alpha \in (0,1)$ becomes now apparently simple. According to lemma 3.10,

$$
\sqrt{\frac{n}{6}}\,\sqrt{b_1} \xrightarrow{\mathcal{D}} N(0,1),
$$

hence, after having calculated the value of $\sqrt{n/6}\sqrt{b_1}$ the quantile function of the standard normal distribution can be employed to get critical values for testing H_0 in (3.3). Therefore, the critical value $c_{n,\alpha/2}$ for a two-sided test for the significance niveau α is $\Phi^{-1}(\alpha/2)$. For the one-sided test, the critical value if of course $\Phi^{-1}(\alpha)$ no matter if one is interested in testing for left or for right skewness. Note that these relations only hold in the asymptotic case. For finite samples sizes, these statements can only be used if the error due to the asymptotic behaviour is adequate small so that the accuracy of the results is appropriate high. We will see in the following that the error for the asymptotic statements is remarkably high for small sample sizes which is an essential disadvantage of the skewenss test.

In his works, Pearson (1930a,b) showed that for finite sample sizes n one is able to give exact expressions of the first four moments of the samples skewness $\sqrt{b_1}$. The odd moments have the value zero, i.e., we have for the expectation $E(\sqrt{b_1}) = 0$ and the skewness $\sqrt{\beta_1}(\sqrt{b_1}) = 0$. Hence, the probability density function of $\sqrt{b_1}$ is symmetric. For the variance $\mathrm{Var}(\sqrt{b_1})$, Pearson obtained

$$\mathrm{Var}(\sqrt{b_1}) = \frac{6(n-2)}{(n+1)(n+3)}. \tag{3.6}$$

The fourth moment of the distribution of $\sqrt{b_1}$ is given by

$$\beta_2(\sqrt{b_1}) = 3 + \frac{36(n+7)(n^2+2n-5)}{(n-2)(n+5)(n+7)(n+9)}.$$

A serious drawback of the asymptotic behaviour of $\sqrt{b_1}$ is the fact that the convergence is very slow which was actually one of the main motivations for Pearson to find exact expressions for the moments of the distribution of $\sqrt{b_1}$. Table 3.1 underlines the slow convergence of the variance $\mathrm{Var}(\sqrt{b_1})$ to its asymptotic value $6n^{-1}$. For small samples like $n = 25$ or $n = 50$ it is obvious to see that using the normal approximation in equation (3.4) can lead to incorrect results because the difference Δ of the finite sample and the asymptotic variance is quite big. Even for $n = 100$ the error being three units in the third decimal place does not seem to be negligible. However, for very large samples Δ becomes very small, a result which is very unsatisfying for practitioners who mostly do not have such large samples sizes.

Table 3.1: Exact and asymptotic variance for the distribution of $\sqrt{b_1}$ and their difference for several samples sizes.

n	$\mathrm{Var}(\sqrt{b_1})$	$6n^{-1}$	$\Delta := 6n^{-1} - \mathrm{Var}(\sqrt{b_1})$
25	0.18956	0.24	0.05044
50	0.10655	0.12	0.01345
100	0.05652	0.06	0.00348
150	0.03844	0.04	0.00156
250	0.02343	0.024	0.00057
500	0.01186	0.012	0.00014
1000	0.00596	0.006	0.00004
2500	0.00239	0.0024	0.00001
5000	0.001199	0.0012	0.000001

For the reason to overcome this problem D'Agostino (1970) found a transformation to approximate the distribution of $\sqrt{b_1}$ under the null hypothesis. Consider the following random variable:

$$X(\sqrt{b_1}) = \delta \log\left(\frac{Y}{\nu} + \sqrt{\frac{Y}{\nu} + 1}\right), \tag{3.7}$$

where

$$\begin{aligned} Y &= \sqrt{b_1}\left(\frac{(n+1)(n+3)}{6(n-2)}\right)^{\frac{1}{2}}, \\ \tau &= \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}, \\ \omega^2 &= \sqrt{2(\tau-1)} - 1, \\ \delta &= \frac{1}{\log(\omega)}, \\ \nu &= \sqrt{\frac{2}{\omega^2-1}}. \end{aligned}$$

According to D'Agostino, $X(\sqrt{b_1})$ is approximately standard normally distributed. Thus, critical values can be obtained by the quantile function Φ^{-1} like described above. The choice of $X(\sqrt{b_1})$ in (3.7) seems reasonable sice under the null hypothesis we can write informally

$$\sqrt{b_1} \approx 0 \iff Y \approx 0 \iff X(\sqrt{b_1}) \approx \delta \log(1) = 0.$$

The accuracy of this transformation was investigated by D'Agostino (1970) who found that the approximation (3.7) works satisfyingly for samples sizes $n \geq 8$.

For samples with $n \leq 7$ D'Agostino and Tietjen (1973) presented critical values for the distribution of $\sqrt{b_1}$ under H_0 based on Monte Carlo simlations. The number of simulated vales was at least 5,000, which is not very large considering todays computer powers. Additionally, they also presented another approximation for the distribution of $\sqrt{b_1}$. In constrast to the approximation to the standard normal distribution in equation (3.7), the transformation is approximately t-distributed. For details we refer to the original work. For larger samples sizes, like $n \geq 150$, the normal approximation

$$\sqrt{b_1}\left(\frac{(n+1)(n+3)}{6(n-2)}\right)^{1/2}$$

based on the finit sample variance in (3.6) is a valid choice for a test statistic, as claimed by D'Agostino (1986a). This suggestion is underlined by a look in table 3.1 where it can be seen that for $n = 150$ the difference between asymptotic and finite sample variance is only one unit in the third decimal place. This error is without essential impact for the most practical situations.

To give a short summary at the end of this subsection, the results of the skewness test should be regarded only with caution. If one already knows, for example from previous studies, that only asymmetry can affect the shape of the data, the skewness test is an appropriate choice of testing the data for normality. In this case, the next question arising is whether to take the one-sided or the two-sided test. If it is possible to anticipate the direction of the skewness, the one-sideed test statstic may be the better choise (see D'Agostino, 1986a). Otherwise, the two-sided test statistic is the appropriate choice.

In cases where there is no prior information one should better avoid the use of the skewness test and trust the results from an omnibus test, that considers both skewness and kurtosis deviations (cf. subsection 3.3 and section 3.4 for more details).

3.2.2 The Kurtosis Test

Anlogous to the skewness test in the preceding subsection, let Y denote a random variable with unknow mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$ and let β_2 be the kurtosis of the pdf of Y as given in definition 3.3. In the case of observing the kurtosis to decide whether the population is normally distributed or not, the null hypothesis is

$$H_0 : \beta_2 = 3 \quad \text{vs.} \quad H_1 : \beta_2 \neq 3.$$

Like for the skewness test, there are also for the kurtosis test some examples of non-normal distributions where the corresponding probability density function is having a value for the kurtosis of $\beta_2 = 3$, even so (see for example Thode (2002, p. 43)). Consequently, one has to keep in mind that not rejecting H_0 of the kurtosis test does not necessarily mean that the sample does not follow a normal distribution. However, such cases with non-normal distributions with $\beta_2 = 3$ are much more rare which does not change the fact that only looking at the kurtosis test in testing for normality can lead to erroneous conclusions about the data.

The test statistic for the kurtosis test is as simple as the one for the skewness test.

3.11 Definition. *Let $y_1, \ldots, y_n$ be a sample of iid random observations coming from a random variable Y. The test statistic of the* **kurtosis test** *is denoted by*

$$b_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^4}{(\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2)^2}$$

An examination of the asymptotic behaviour of b_2 is also possible.

3.12 Lemma. *Let b_2 be the test statistic of the kurtosis test from definition 3.11 based on a sample $y_1, \ldots, y_n$ of iid random observations coming from a standard normally distributed random variable Y. Then, $\sqrt{n}b_2$ is asmyptotically normal distributed with mean 3 and variance 24, i.e.,*

$$\sqrt{n}(b_2 - 3) \xrightarrow{\mathcal{D}} N(0, 24) \quad \text{as } n \to \infty.$$

Proof: The proof is similar to the proof of lemma 3.10. This time consider the function $\mathbb{R}^4 \to \mathbb{R}$, with

$$h(a,b,c,d) = \frac{d - 4ac + 6a^2b - 3a^4}{(b-a^2)^2}.$$

Beside the quantities $\tilde{m}_i, i = 1,2,3$ given in equation (3.5), we additionally define

$$\tilde{m}_4 = n^{-1} \sum_{i=1}^{n} x_i^4.$$

It follows, after a little algebra that

$$h(\tilde{m}_1, \tilde{m}_2, \tilde{m}_3, \tilde{m}_4) = \frac{\tilde{m}_4 - 4\tilde{m}_1\tilde{m}_3 + 6\tilde{m}_1^2\tilde{m}_2 - 3\tilde{m}_1^4}{(\tilde{m}_2 - \tilde{m}_1^2)^2} = b_2.$$

Using the same arguments than in the proof above, we have

$$\sqrt{n}(\tilde{m}_1, \tilde{m}_2 - 1, \tilde{m}_3, \tilde{m}_4 - 3) \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

where

$$\begin{aligned}
\Sigma &= E\left(\begin{pmatrix} x_i \\ x_i^2 - 1 \\ x_i^3 \\ x_i^4 - 3 \end{pmatrix} (x_i, x_i^2 - 1, x_i^3, x_i^4 - 3)\right) \\
&= E\begin{pmatrix} x_i^2 & x_i^3 - x_i & x_i^4 & x_i^5 - x_i^3 \\ x_i^3 - x_i & x_i^4 - 2x_i^2 + 1 & x_i^5 - x_i^3 & x_i^6 - x_i^4 - 3x_i^2 + 3 \\ x_i^4 & x_i^5 - x_i^3 & x_i^6 & x_i^7 - 3x_i^3 \\ x_i^5 & x_i^6 - x_i^4 - 3x_i^2 + 3 & x_i^7 - 3x_i^3 & x_i^8 - 6x_i^4 + 9 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & 3 & 0 \\ 0 & 2 & 0 & 12 \\ 3 & 0 & 15 & 0 \\ 0 & 12 & 0 & 96 \end{pmatrix},
\end{aligned}$$

since $E(x_i^7) = 0$, $E(x_i^6) = 15$ and $(x_i^8) = 105$ (see Stuart and Ord (1987, p. 78)). For the determination of the gradient ∇h we omit the calculus and give only the final results which is of the form

$$\nabla h(a,b,c,d) = (b-a^2)^{-2} \begin{pmatrix} \frac{4(3ab^2 - bc + ad - 3a^2c)}{b-a^2} \\ \frac{2a(3ab - 3a^2 + 2d - 8ac + 12a^2b - 6a^4)}{b-a^2} \\ -4a \\ 1 \end{pmatrix}$$

Hence we have $\nabla h(0,1,0,3) = (0,0,0,1)'$ and we can again use theorem A.10 in the appendix. Before, check that $h(0,1,0,3) = 3$ and

$$(\nabla h(0,1,0,3))' \Sigma \nabla h(0,1,0,3) = (0,0,0,1) \begin{pmatrix} 1 & 0 & 3 & 0 \\ 0 & 2 & 0 & 12 \\ 3 & 0 & 15 & 0 \\ 0 & 12 & 0 & 96 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 3 \end{pmatrix} = 24$$

Then the following holds:

$$\begin{aligned}\sqrt{n}(g(\tilde{m}_1, \tilde{m}_2, \tilde{m}_3, \tilde{m}_4) - g(0,1,0,3)) &= \sqrt{n}(b_2 - 3) \\ &\xrightarrow{\mathcal{D}} N\left(0, (\nabla g(0,1,0,3))' \,\Sigma\, \nabla g(0,1,0,3)\right) = N(0,24),\end{aligned}$$

which is exactly the assertion. □

Besides the first four moments of the distribution of $\sqrt{b_1}$, Pearson (1930a,b) also gave exact values of the first four moments of the distribution of b_2. For the expectation $E(b_2)$ and the variance $\mathrm{Var}(b_2)$ Pearson showed that

$$E(b_2) = \frac{3(n-1)}{n+1} \tag{3.8}$$

and

$$\mathrm{Var}(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}. \tag{3.9}$$

For the third moment he obtained

$$\sqrt{\beta_1}(b_2) = \sqrt{\frac{216}{n}} \left\{\frac{(n+3)(n+5)}{(n-3)(n-2)}\right\}^{1/2} \frac{(n^2-5n+2)}{(n+7)(n+9)}.$$

Thus, the pdf of b_2 is not symmetric like the pdf of $\sqrt{b_1}$. This complicates the use of critical values for a two-sided test together with the more important fact that the convergence of b_2 to its asymptotic distribution is very slow. The slow convergence becomes clear by looking at table 3.2 where the difference between the exact and the asymptotic mean and variance of the distribution of b_2 is presented. One can see that even for samples sizes of $n = 500$ the difference between exact and asymptotic mean is one unit in the second decimal place which is relatively high. For $n = 1{,}000$, the difference shrinks to five units in the third decimal place for the mean and to three units in the fourth decimal place for the variance of b_2. Hence, using the normal approximation

$$x = \frac{b_2 - E(b_2)}{\sqrt{\mathrm{Var}(b_2)}} = \left(b_2 - \frac{3(n-1)}{(n+1)}\right) \Big/ \left(\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}\right)^{1/2} \tag{3.10}$$

yields to an acceptable small error not until before $n \geq 1{,}000$. However, samples of that size do practically not exist. Hence, for practitioners, the normal approximation in (3.10) is not an alternative and should not be used (see D'Agostino (1986b, p. 389)).

Similar to the skewness test in the preceding subsection or for the distribution of the W statistic described in subsection 2.2.2, a convenient way to get to percentage points for the kurtosis test for smaller samples sizes than 1,000 is to perform Monte Carlo studies. On this way one is able to obtain emprircal percentage points which are adequate exact for a sufficient large value m of repetitions. D'Agostino and Tietjen (1971) presentend in their simulation study percentage points for samples sizes $7 \leq n \leq 50$ based on a number m of simulations of at

Table 3.2: Exact and asymptotic mean and variance for the distribution of b_2 and their difference for several samples sizes.

n	$E(b_2)$	$3 - E(b_2)$	$\mathrm{Var}(b_2)$	$24n^{-1}$	$24n^{-1} - \mathrm{Var}(b_2)$
25	2.76923	0.23077	0.53466	0.96	0.42534
50	2.88235	0.11765	0.35706	0.48	0.12294
100	2.94059	0.05941	0.20679	0.24	0.03321
150	2.96026	0.03974	0.14485	0.16	0.01515
250	2.9761	0.0239	0.09043	0.096	0.00557
500	2.98802	0.01198	0.04658	0.048	0.00142
1000	2.99401	0.00599	0.02364	0.024	0.00036
2500	2.9976	0.0024	0.00954	0.0096	0.00006
5000	2.9988	0.0012	0.00479	0.0048	0.00001

least 20,000. Later, D'Agostino and Pearson (1973) extended the results for samples sizes up to $n = 200$ based on a value of m of at least 10,000.

Another possibility to overcome the disadvantage of the slow convergence of the normal approximation in (3.10) consits in finding an appropriate transformation of b_2, to a known distribution function—mostly the standard normal distribution. The transformation mostly cited in literature is probably the one given in Anscombe and Glynn (1983). Based on the normal approximation x given in equation (3.10), they state that for $n \geq 20$,

$$X(b_2) = \sqrt{\frac{2}{9A}} \left(1 - \frac{2}{9A} - \left(\frac{1 - 2/A}{1 + x\sqrt{2/(A-4)}} \right) \right), \tag{3.11}$$

where

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}} \left(\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{1 + 4/\beta_1(b_2)} \right),$$

is approximately a standard normal variable with mean zero and variance unity. As explained in D'Agostino (1986b), the approximation in (3.11) can be used to test both one-sided and two-sided alternatives. For the one-sided test against heavy tailed distributions, the alternative is $H_1 : \beta_2 > 3$. In this case, one would reject H_0 on a significance level α, if $X(b_2)$ in equation (3.11) exeeds the value $\Phi^{-1}(1-\alpha)$. On the contrary, for the one-sided test against light tailed distributions, we have $H_1 : \beta_2 < 3$. Here, H_0 is rejected on a significance level α, if $X(b_2) < -\Phi^{-1}(1-\alpha)$. Note that in the case of a one-sided alternative, the null hypotheses has to be adapted corresponding to the choice of H_1.

Summing up the results concerning the kurtosis test, it is clear that—even so the distribution of b_2 is asymptotic normally distributed—the normal approximation in equation (3.10) is practically unuseable due to the very slow convergence of $\sqrt{n}b_2$ to the $N(3, 24)$ distribution. However, for $7 \leq n \leq 20$, the way out of this problem consits in using the empirical percentage points as the results of Monte Carlo studies. According to D'Agostino (1986b), for $n < 20 \leq 200$, both the empirical percentage points or the normal approximation in (3.11) given by Anscombe and Glynn are recommended. For $n > 200$, only the approximation of Anscombe and Glynn is an adequate choice.

3.3 Omnibus Tests Using Skewness and Kurtosis

The disadvantages of only performing the skewness or the kurtosis test have been mentioned several times in the last two subsections. Combining these two moment tests seems a natural thing in order to obtain an omnibust test for normality. Two attempts for such an omnibust test, namely the R test and the K^2 test will be shortly presented in this subsection.

3.3.1 The R Test

The most simple and intuitive way to perfom an omnibus test is of course to perform both the skewness and the kurtosis test simultaneously. Each of the two tests can be conducted in a suitable manner, as described in the two subsections 3.2.1 and 3.2.2 above. Thus, given a sample $y_1, \ldots, y_n$ of iid random observations coming from a random variable Y, the two test statistics $\sqrt{b_1}$ and b_2 have to be calculated. For the R test we combine the two statistics to a two-dimensional vector $(\sqrt{b_1}, b_2)$ and examine the realizations of this vector on a two-dimensional plane. The null hypothesis is

$$H_0 : \left(\sqrt{\beta_1}, \beta_2\right) = (0, 3) \quad \text{vs.} \quad H_1 : \left(\sqrt{b_1}, b_2\right) \neq (0, 3).$$

The difference now comparing to the testing procdures explained above consists in correcting the significance level due to multiple comparisons.

Considering a significance level of $100\alpha'\,\%, \alpha' \in (0, 1)$, let $\pm c_{\sqrt{b_1}}(\alpha')$ be the upper und lower percentage points of the skewness test, i.e.,

$$P\left(\sqrt{b_1} \leq -c_{\sqrt{b_1}}(\alpha')\right) = P\left(\sqrt{b_1} \geq +c_{\sqrt{b_1}}(\alpha')\right) = \frac{1}{2}\,\alpha'. \tag{3.12}$$

Further, let $\pm c_{b_2}(\alpha')$ be the upper and lower percentage points of the kurtosis test for the same significance level defined analogous to those of $\sqrt{b_1}$. Consider the rectangle R defined by the four points

$$\left(-c_{\sqrt{b_1}}(\alpha'), -c_{b_2}(\alpha')\right), \left(+c_{\sqrt{b_1}}(\alpha'), -c_{b_2}(\alpha')\right), \left(+c_{\sqrt{b_1}}(\alpha'), +c_{b_2}(\alpha')\right), \left(-c_{\sqrt{b_1}}(\alpha'), +c_{b_2}(\alpha')\right). \tag{3.13}$$

The motivation for the defintion of R becomes clear in the next lemma. It shows—under the crucial assumption of the indepencence of $\sqrt{b_1}$ and b_2—that a significance level for an omnibus test can easliy be determined.

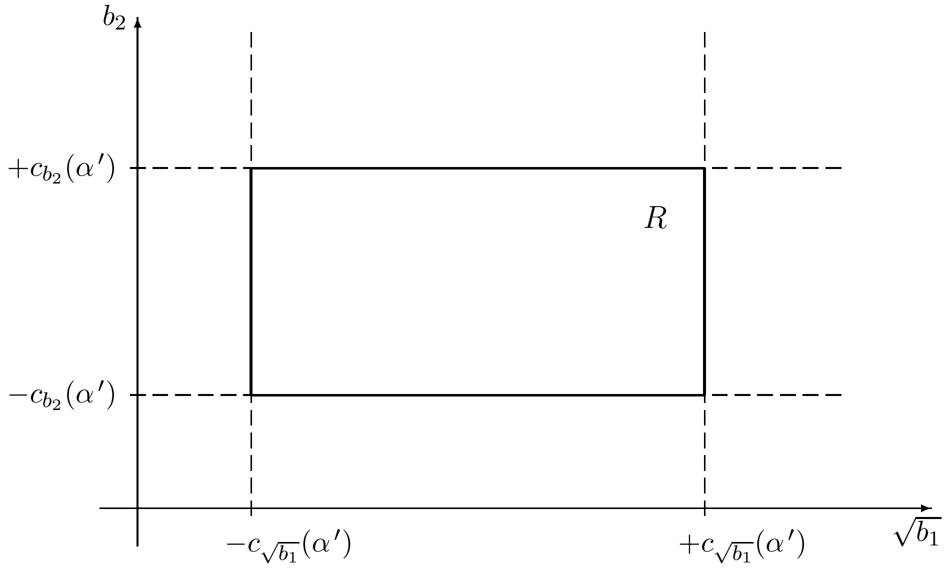


Figure 3.3: The solid lines define the rectangular R defined by (3.13). The dashed lines outline the eight areas outside of R that are summarized in the proof of lemma 3.13.

3.13 Lemma. *Aussme that $\sqrt{b_1}$ and b_2 are stochastically independent. When performing both the skewness and the kurtosis test, the probability α of the point $\left(\sqrt{b_1}, b_2\right)$ to fall outside the rectangle R given in* (3.13) *is*

$$P\left(\left(\sqrt{b_1}, b_2\right) \notin R\right) = \alpha = 2\alpha' - (\alpha')^2. \tag{3.14}$$

Proof: The probability of $\left(\sqrt{b_1}, b_2\right)$ to fall outside R is the sum of the following eight probabilities:

$$\begin{aligned}
P\left(\left(\sqrt{b_1}, b_2\right) \notin R\right) = \quad & P\left(\sqrt{b_1} \leq -c_{\sqrt{b_1}}(\alpha'), b_2 \leq -c_{b_2}(\alpha')\right) \\
&+ P\left(-c_{\sqrt{b_1}}(\alpha') < \sqrt{b_1} < +c_{\sqrt{b_1}}(\alpha'), b_2 \leq -c_{b_2}(\alpha')\right) \\
&+ P\left(\sqrt{b_1} \geq +c_{\sqrt{b_1}}(\alpha'), b_2 \leq -c_{b_2}(\alpha')\right) \\
&+ P\left(\sqrt{b_1} \geq +c_{\sqrt{b_1}}(\alpha'), -c_{b_2}(\alpha') < b_2 < +c_{b_2}(\alpha')\right) \\
&+ P\left(\sqrt{b_1} \geq +c_{\sqrt{b_1}}(\alpha'), b_2 \geq +c_{b_2}(\alpha')\right) \\
&+ P\left(-c_{\sqrt{b_1}}(\alpha') < \sqrt{b_1} < +c_{\sqrt{b_1}}(\alpha'), b_2 \geq +c_{b_2}(\alpha')\right) \\
&+ P\left(\sqrt{b_1} \leq -c_{\sqrt{b_1}}(\alpha'), b_2 \geq +c_{b_2}(\alpha')\right) \\
&+ P\left(\sqrt{b_1} \leq -c_{\sqrt{b_1}}(\alpha'), -c_{b_2}(\alpha') < b_2 < +c_{b_2}(\alpha')\right).
\end{aligned}$$

Figure 3.3 attempts to give a better understanding of the construction of the probability $P\left(\left(\sqrt{b_1}, b_2\right) \notin R\right)$. Under the assumption of independence, when using equation (3.12) the probability becomes

$$\begin{aligned}
P\left(\left(\sqrt{b_1}, b_2\right) \notin R\right) = \quad & P\left(\sqrt{b_1} \le -c_{\sqrt{b_1}}(\alpha')\right) P\left(b_2 \le -c_{b_2}(\alpha')\right) \\
& + P\left(-c_{\sqrt{b_1}}(\alpha') < \sqrt{b_1} < +c_{\sqrt{b_1}}(\alpha')\right) P\left(b_2 \le -c_{b_2}(\alpha')\right) \\
& + P\left(\sqrt{b_1} \ge +c_{\sqrt{b_1}}(\alpha')\right) P\left(b_2 \le -c_{b_2}(\alpha')\right) \\
& + P\left(\sqrt{b_1} \ge +c_{\sqrt{b_1}}(\alpha')\right) P\left(-c_{b_2}(\alpha') < b_2 < +c_{b_2}(\alpha')\right) \\
& + P\left(\sqrt{b_1} \ge +c_{\sqrt{b_1}}(\alpha')\right) P\left(b_2 \ge +c_{b_2}(\alpha')\right) \\
& + P\left(-c_{\sqrt{b_1}}(\alpha') < \sqrt{b_1} < +c_{\sqrt{b_1}}(\alpha')\right) P\left(b_2 \ge +c_{b_2}(\alpha')\right) \\
& + P\left(\sqrt{b_1} \le -c_{\sqrt{b_1}}(\alpha')\right) P\left(b_2 \ge +c_{b_2}(\alpha')\right) \\
& + P\left(\sqrt{b_1} \le -c_{\sqrt{b_1}}(\alpha')\right) P\left(-c_{b_2}(\alpha') < b_2 < +c_{b_2}(\alpha')\right) \\
= {} & 4\left(\frac{(\alpha')^2}{4} + (1-\alpha')\frac{\alpha'}{2}\right) \\
= {} & 4\left(\frac{(\alpha')}{2} - \frac{(\alpha')^2}{4}\right) \\
= {} & 2\alpha' - (\alpha')^2.
\end{aligned}$$

□

Using the results of the last lemma, we can easily find a multiple significance level for the omnibus test by writing the equation (3.14) in terms of α'. The we find

$$\alpha' = 1 - (1-\alpha)^{\frac{1}{2}}.$$

The other solution of the quadratic equation, $\alpha' = 1 + (1-\alpha)^{1/2} > 1$, can of course be excluded. We will now summarize the obtained results.

3.14 Definition. *Let $y_1, \dots, y_n$ be a sample of iid random observations coming from a random variable Y. The test statistic of the R* **test** *is given by*

$$r = \left(\sqrt{b_1}, b_2\right).$$

If r lies outside the set R given in equation (3.13)*, the null hypothesis of a normal sample has to be rejected.*

3.15 Remark. Note that the significance level $\alpha' = 1 - (1-\alpha)^{1/2}$ is actually the correction of Šidák for two tests on one sample (see for example Westfall (1999)).

In the proof of lemma 3.13, we make use of the assumption that $\sqrt{b_1}$ and b_2 are statistically independent from each other. However, for small sample sizes, this is not the case. Figure 3.4 makes an attempt to show the correlation of $\sqrt{b_1}$ and b_2 for small, moderate and relatively large samples sizes. The single plots were all drawn based on 10,000 randomly generated standard normal variables with the corresponding sample size n. Each of the random sample is independent from all the other samples. For a sample size of $n = 5$, there is definitely a very high correlation between the two quantities. The larger the sample size becomes, the less becomes the structure in the scatter plot. For $n = 500$ there is almost no structure to recognize in the plot, at the latest for $n = 5000$ the structure has gone completely. This is in accord with the fact of the slow asymptotic behaviour, especially the one of the distribution of b_2.

Consequently the assumption of the indepencence of $\sqrt{b_1}$ and b_2 is injured so that the results of lemma 3.13 can actually not be used directly. Pearson et al. (1977) who were the first to present the R test in literature, claimed that because of the strong correlation of $\sqrt{b_1}$ and b_2 for small and moderate sample sizes, the R test is a conservative test that is, the test rejects normal samples slightly less often than desired. They base this statement of the result of an empirical simulation for several samples sizes. Though, Pearson et al. (1977) present some corrected significance levels that are less conservative, the use of this test is not recommend since it is still not as powerful than for example the K^2 in the next subsection (cf. D'Agostino (1986b)).

3.3.2 The K^2 Test

Another approach for combining the skewness and the kurtosis test to an omnibus test bases upon the following idea. Since $\sqrt{b_1}$ and b_2 are both asymptotically normal, an obvious choise would be to take the sum of the two standardized test statistics

$$X^2 = \left(\frac{\sqrt{b_1} - E(\sqrt{b_1})}{\mathrm{Var}(\sqrt{b_1})}\right)^2 + \left(\frac{b_2 - E(b_2)}{\mathrm{Var}(b_2)}\right)^2.$$

Under H_0—assuming that the two summands were independent—X^2 would be chi-squard distributed with two degrees of freedom, denoted by χ^2_2 (see Falk et al. (2002, p. 54) for more details of the chi-square distribution). Recall that we already presented transformations of $\sqrt{b_1}$ and b_2 in order to get approximative standard normal random variables in the subsections 3.2.1 and 3.2.2. This leads us to the K^2 test.

3.16 Definition. *Let $y_1, \ldots, y_n$ be a sample of iid random observations coming from a random variable Y. The test statistic denoted by*

$$K^2 = X^2(\sqrt{b_1}) + X^2(b_2),$$

where $X(\sqrt{b_1})$ and $X(b_2)$ are given in the equations (3.7) *and* (3.11)*, respectively, is called the* K^2 **statistic**, *the corresponding test will be called the* K^2 **test**.

As already mentioned in this subsection before, the assumption of independence cannot be hold up for small and moderate samples sizes. Thus, the fact that K^2 is χ^2_2 distributed under the null hypothesis does not hold for the most common sample sizes. To overcome this drawback,

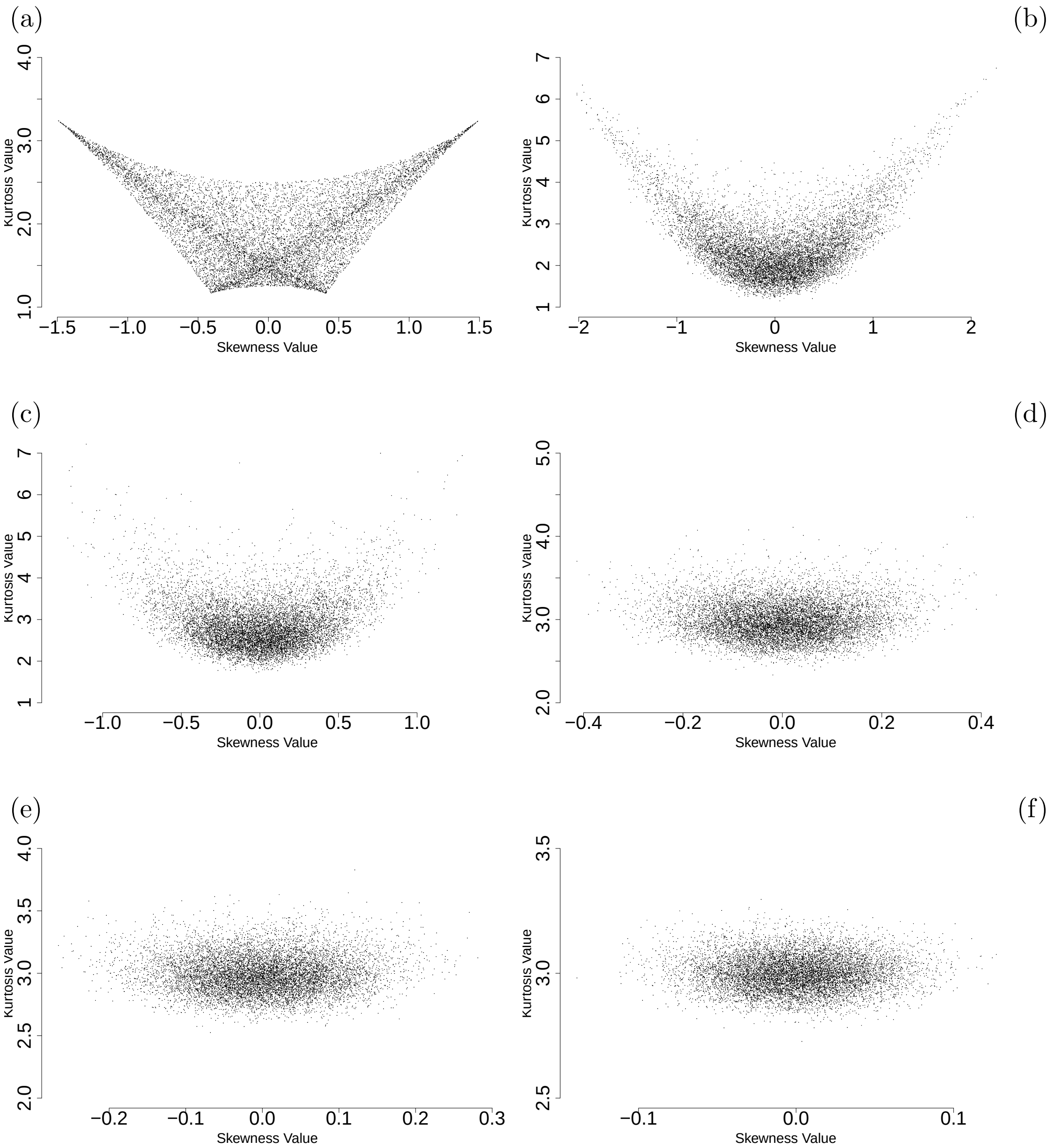


Figure 3.4: Scatterplot of $\sqrt{b_1}$ and b_2 for 10,000 randomly generated normal samples for each sample size, respectively. The sample sizes are (a) $n = 5$, (b) $n = 10$, (c) $n = 50$, (d) $n = 500$, (e) $n = 1,000$ and (f) $n = 5,000$.

Bowman and Shenton (1975) constructed contour plots for some critical values for the joint distribution of $\sqrt{b_1}$ and b_2 based on empirical simulation. They did this for sample sizes of $20 \leq n \leq 1{,}000$. For more details, we refer to the original work. The approximation of Bowman and Shenton seems to be really accurate, so that the use of the K^2 statistic is to be preferred to the use of the R test, as recommended by D'Agostino (1986b).

3.4 The Jarque-Bera Test

In this section another omnibust test based on the third and fourth sample moment is presented, the so-called Jarque-Bera test. Since this test is probabliy the most popular and mostly used moment test for normality not only in ecnonomics (cf. Gel and Gastwirth (2008), Urzua (1996), Thadewald and Büning (2007), Jahanshahi et al. (2008)), we will dedicate it an own section. Beside the original testing procedure, a modification of the test is shortly introduced. For every new introduced test, we also present emprircal significance points based on empirical sampling studies for the later power study.

3.4.1 The Original Procedure

The underlying idea of this omnibus test is once a again very simple. As the K^2 test in subsection 3.3.2, the Jarque-Bera tests picks up the same idea of taking the sum of the two test statistics of the skweness and the kurtosis test. In contrast to the K^2 test, the original values of $\sqrt{b_1}$ and b_2 run in the test statistic instead of transformations of them. To become more detailed we start with the formal definition of the test

3.17 Definition. *Let $y_1, \ldots, y_n$ be a sample of iid random observations coming from a random variable Y. Further let $\sqrt{b_1}$ and b_2 be defined like in definition 3.3. The test statistic denoted by*

$$JB = n\left(\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2-3)^2}{24}\right) = \frac{n}{6}\left((\sqrt{b_1})^2 + \frac{(b_2-3)^2}{4}\right) \tag{3.15}$$

is called **Jarque-Bera statistic** *or just JB* **statistic***, the corresponding test for normality is called* **Jarque-Bera test** *or shortly JB* **test***.*

The interpretaion of the test statistic in equation (3.15) is a simple task. As we have already shown before that $E(\sqrt{b_1}) = 0$ and $\mathrm{Var}(\sqrt{b_1}) = 6$ as well as $E(b_2) = 3$ and $\mathrm{Var}(b_2) = 24$, we can use the same argumentation like for the K^2 test in subsection 3.3.2 and say that we have the sum of the squared values of two $N(0,1)$ distributed random variables. Thus, the JB statistic is χ^2_2 distributed if $\sqrt{b_1}$ and b_2 are independent. However, the independence of $\sqrt{b_1}$ and b_2 is only valid in the asymptotic case which we have not shown yet.

In fact, Lehmann (1999, p. 345) states that under the assumption of a normally distributed population, $\left(\sqrt{n}\sqrt{b_1}, \sqrt{n}(b_2 - 3)\right)$ are asmyptotically bivariate normal distributed and uncorrelated. A direct consqeuece is that $\sqrt{b_1}$ and b_2 are asymptotically independent, since this is always the case when bivariate normal distributios have a correlation of zero. Folling this argumentation it is seen directly that under the null distribution FB has a limiting χ^2-distribution with 2 degrees of freedom.

The test is named after Jarque and Bera who developed the test and showed that FB is asympotically χ^2_2-distributed. (cf. Jarque and Bera (1980, 1987)). Jarque and Bera used another approach to prove the asymptotic behaviour of the FB statistic by uncovering that the JB test can actually derived as a Lagrange multiplier test. For more details to the Lagrange multiplier test see for example Lehmann (1999, p. 529).

It is interesting that the test statistic has already been mentioned by Bowman and Shenton (1975). In their paper, they shortly present the test statistic in equation (3.15) and also state that it is asymptotically χ^2_2-distributed. Because of the slow convergence of the JB statistic to the χ^2_2-distribution, Bowman and Shenton did not attach much importance to this test. In their work, Jarque and Bera (1987) point to this fact but proclaim the test in (3.15) as their "invention", anyway. They rely on being the first who have shown that the JB statistic is essentially a Lagrange multiplier test. However the fist ones who noted that JB is asymptotically χ^2_2-distributed were Bowman and Shenton. Thus, the test for normality in definition 3.17 should actually be named as the **Bowman-Shenton test**. But since the denotation Jarque-Bera test has become largely widespread in literature, we will also use this name of the test in the remainder of this work.

As already mentioned several times, the asymptotic convergence of the JB statistic is very slow, hence decisions for testing normality based on the quantile function of the χ^2_2-distribution can lead to serious incorrect errors. Jarque and Bera also called attention to this disadvantage of the test. The convergence of the JB statistic to its asymptotic distribution is tried to be visualized in figure 3.5. For each histogram in this figure, the Jarque-Bera statistic was calculated for $m = 1{,}000{,}000$ realizations of standard normally generated random samples of the corresponding sample size n. Additionally, the theoretical probability distribution function of the χ^2-distribution with 2 degrees of freedom is plotted in each histogram so that one is able to compare the goodness-of-fit of the empirical distribtion with the theoretical distribution. It becomes obviously that for small sample sizes the deviation of the histogram to the theoretical curve is very large, in particular for $n = 5$. The bigger the sample size, the better becomes the fit of the histogram to the curve. For $n = 500$ there is almost no difference any more between histogram and curve.

Since for finit sample sizes the joint distribution of $\sqrt{b_1}$ and b_2 is unknown once again the possibility of performing Monte Carlo simulations becomes the favourite way to obtain empirical percentage points. In their work, Jarque and Bera (1987) presented empirical significance points

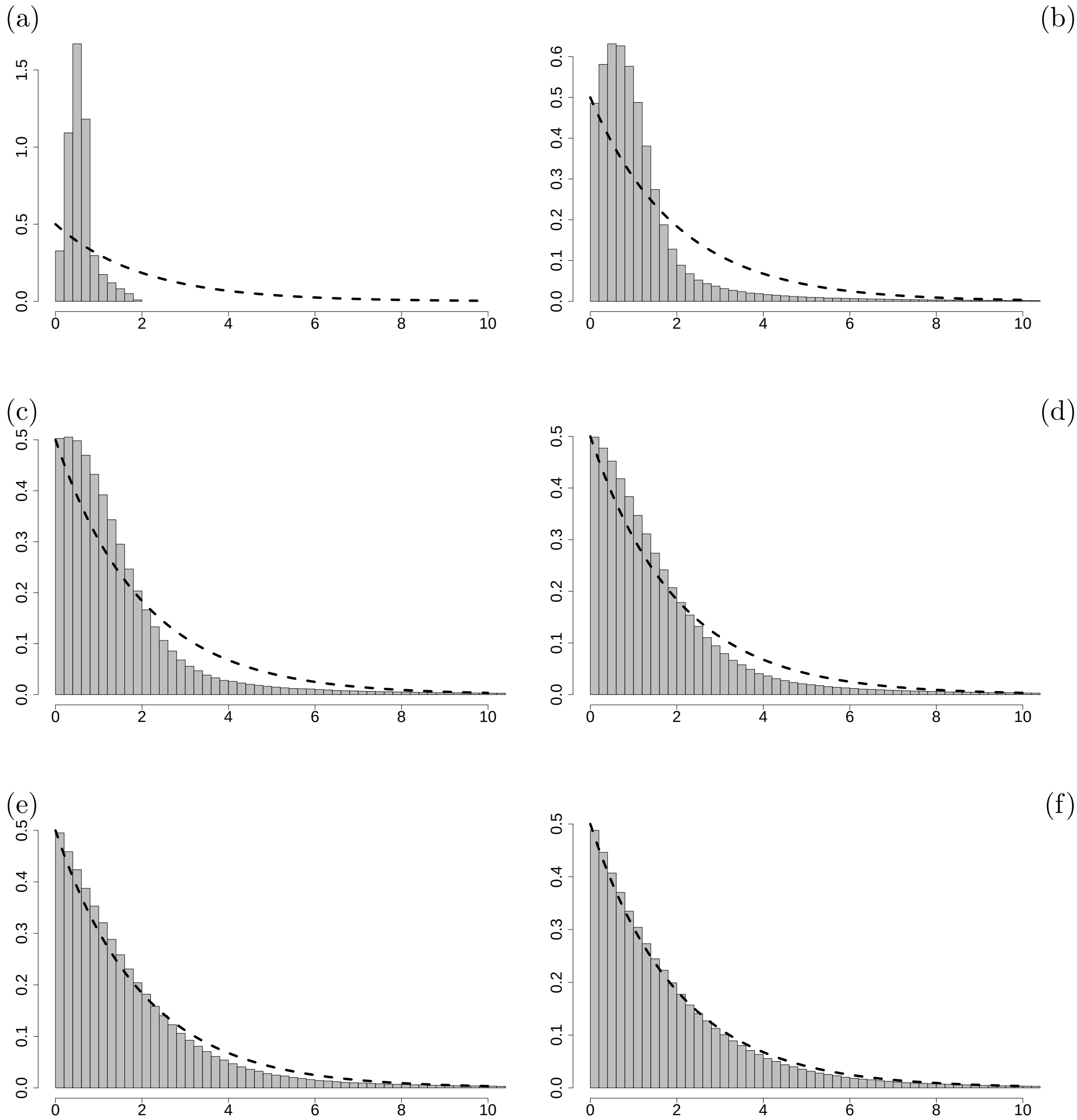


Figure 3.5: Histogram of the JB statistic for several sample sizes together with the pdf of the χ_2^2 distribution. For each sample size, 1,000,000 standard normal samples were generated independently from each other and the JB statistic was calculated each time. The sample sizes are (a) $n = 5$, (b) $n = 20$, (c) $n = 50$, (d) $n = 100$, (e) $n = 200$ and (f) $n = 500$

based on a simulation with a value of $m = 10{,}000$. The small number of only 10,000 replications seem to provide only unsatisfying accurate percentage points, which has also been pointed out by Deb and Sefton (1996). Consequently, they performed their own simulation studies with a value of $m = 600{,}000$ and found out that the empirical significance points of Jarque and Bera were too slow with the result that the test statistic tends to reject too frequently when using the empirical percentage points of Jarque and Bera. In table 3.3 new percentage points of our own simulations based on a number of $m = 1{,}000{,}000$ replication are presented with the objective to check the accuracy of the empirical significance points of Deb and Sefton and for the power studies in chapter 4.

Table 3.3: Empirical significance points for the JB statistic for several sample sizes based on $m = 1{,}000{,}000$ repetitions compared with the critical values given by Deb and Sefton (1996). For $n = \infty$ the significance point is the $1 - \alpha$ quantile of the χ^2_2 distribution as an orientation for the asymptotic convergence of the JB statistic. The calculated test statistics are the same that were used to plot the histograms in figure 3.5.

	own simulation			Deb and Sefton (1996)	
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$
10	5.6885	2.5235	1.6214	—	—
15	8.2431	3.2973	2.0581	—	—
20	9.6986	3.7947	2.3489	3.7737	2.3327
25	10.6469	4.1517	2.5723	—	—
30	11.2599	4.4055	2.7453	4.3852	2.7430
35	11.7900	4.6044	2.8866	—	—
50	12.4479	4.9671	3.1858	5.0002	3.1880
100	12.4955	5.4237	3.6649	5.4365	3.6703
200	11.9265	5.6954	4.0443	5.7113	4.0463
500	10.8050	5.8587	4.3306	5.8869	4.3522
∞	9.2103	5.9915	4.6052	5.9915	4.6052

The generation of these numbers was done on the same way than in chapter 2 with the regression type tests. The differences between the value of our own simulation comparing with the value of Deb and Sefton (1996) are often larger than one unit in the second decimal place. Considering the number of 600,000 replications in the study of Deb and Sefton, the results seem surprisingly. However, we trusted in our own significance points which is why we used them for the empirical studies in chapter 4.

3.4.2 The Adjusted Jarque-Bera Test as a Modification of the Jarque-Bera Test

The original form of the Jarque-Bera test given in definition 3.17 has become very popular and is used widespreadly in practice. The justification for this fact is the ease of computation of the JB statistic and mostly the very good power properties of the test comparing to other tests for normality, see section 4.2 for more details. A disadvantage of the JB test is the slow convergence of the third and the fourth moment to their corresponding asymptotical limit distributions which has been mentioned in the meantime multiple times in this chapter. Hence, the convergence of the JB statistic to the χ^2_2-distribution is also very slow, which we have already seen in figure 3.5. In his work, Urzua (1996) provides a natural modification to mitigate this problem.

To comprehend the approach of Urzua, recall in mind that we already presented exact values of the mean and the variance of $\sqrt{b_1}$ and b_2 for finite vales of n in the subsections 3.2.1 and 3.2.2, respectively. Urzua replaced the asymptotic values of the mean and the variance in the FB statisic with the exact vaules resulting in a new test statistic.

3.18 Definition. *Let $y_1, \ldots, y_n$ be a sample of iid random observations coming from a random variable Y. Further let $\sqrt{b_1}$ and b_2 be defined like in definition 3.3. The test statistic denoted by*

$$AJB = \frac{(\sqrt{b_1})^2}{\mathrm{Var}(\sqrt{b_1})} + \frac{(b_2 - E(b_2))^2}{\mathrm{Var}(b_2)},$$

where $\mathrm{Var}(\sqrt{b_1})$, $E(b_2)$ and $\mathrm{Var}(b_2)$ are given in the equations (3.6), (3.8) and (3.9), respectively, is called **adjusted Jarque-Bera statistic** *or just AJB* **statistic**, *the corresponding test for normality is called* **adjusted Jarque-Bera test** *or shortly AJB* **test**.

Note that since JB and AJB are asymptotically equivalent, the AJB statistic is still asympotically χ^2_2-distributed, so that for sufficiently large sample sizes the quantile function of the χ^2_2-distribution provides critical values for a significance test. Since the sample sizes where the asymptotic behaviour of AJB becomes sufficiently accurate are still too large for practical purposes, Urzua conducted a simulation study to get empirical significance points for the AJB statistic for several sample sizes. The number m of replications was with $m = 10{,}000$ relatively small so that the accuary of the critical points may be casted on doubt. For that reason, Thadewald and Büning (2007) performed their own simulation study in the framework of their power studies based on a value of $m = 100{,}000$. Since this number still seems to be too slow, we give in table 3.4 the results of our own empirical significance points. The number of repetitions here was $m = 1{,}000{,}000$ for each sample size n.

Like for the emprirical significance points in table 3.3, we note here that the differences between our simulation results and the values of Thadewald and Büning are again remarkably high—for $n = 20$ even more than one unit in the first decimal place. Considering the number of repetitions

Table 3.4: Empirical significance points for the AJB statistic for several sample sizes based on $m =$ 1,000,000 repetitions compared with the critical values given by Thadewald and Büning (2007) which are the entries in the fifth column. For $n = \infty$ the significance point is the $1 - \alpha$ quantile of the χ^2_2 distribution as an orientation for the asymptotic convergence of the AJB statistic.

	own simulation			
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$
10	18.3055	7.3840	4.1596	7.4374
15	18.8163	7.0659	4.0093	—
20	18.7801	6.9696	3.9835	6.8559
25	18.2933	6.8624	3.9691	—
30	17.8847	6.7310	3.9421	—
35	17.5237	6.6814	3.9627	—
50	16.6950	6.5436	3.9927	6.5940
100	14.8257	6.3385	4.1249	6.3381
200	13.0798	6.1868	4.2855	6.1678
500	11.2621	6.0700	4.4407	6.0378
∞	9.2103	5.9915	4.6052	5.9915

for our simulation study compared with the number of repetitions of Thadewald and Büning, their results may be casted on doubt. For that reason, we used our empirical significance points in the power studies in the next chapter.

Chapter 4

Power Studies

After having introduced a wide number of different tests for normality, the objective in this chapter is to compare all the tests directly with each others. The value in which we will measure the goodness of the tests is their power, or rather their empirical power. To determine the empirical power of a test, simulation studies based on Monte Carlo experiments have to be conducted. This is the main content of this chapter.

However, in the first part of this chapter the different tests will not be compared with each other but we will have a look at their behavoiur under the situation of a given normally distributed sample. In the second part we calculate the empirical power for the investigated tests for several sample sizes and several alternative distributinos. The results are tabled, discussed and compared with the results of other simulation studies in the past.

4.1 Type I Error Rate

4.1.1 Theoretical Background

Before comparing the different tests for normality with each other, we first will take a closer look at the properties for each single test. Maybe the most important property of a test is that it guarantees that the rate of erroneously rejecting the null hypothesis will be not too high. Otherwise, we would too often reject H_0 when in fact the sample comes from a normal population—a property that would turn the test not into a good choice. We start with the formal definition of what we just mentioned.

4.1 Definition. *The test decision of rejecting the null hypothesis when in fact H_0 is acutually true is called* **type I error**. *The probability of making a type I error is denoted by α. This probability is also called the* **significance level** *of a test.*

As already mentioned, it is obvious that for given sample sizes n the value of α_n should not exeed a certian level. The standard value is 0.05. With the objective to check for a given test of normality if α_n is higher or lower than 0.05, we can run empirical studies based on Monte Carlo simulations. The principle of these studies is easily described.

4.2 Definition. *Let m be the number of randomly generated independent samples of sizes n where all m samples follow a standard normal distribution. The empirical type I error rate $\hat{\alpha}_{n,m}$ of a given test for normality for a given sample size n is given by*

$$\hat{\alpha}_{n,m} = \frac{M}{m}, \tag{4.1}$$

where $M \leq m$ is the number of the m tests that reject the null hypothesis of a normally distributed sample on the significance level α.

The maybe most important reason for not examining the theoretical significance level but the empircial version of the type I error rate is that for all tests of the regression type, the null distribution is unknown and hence, no quantile function of the distribution can be determined. For the Jarque-Bera type tests the asymptotic distribution is well known in most cases. However, for finite sample sizes, it is not possible to express the null distribution for which reason we use empirical sampling in this case, too. Considering the form of $\hat{\alpha}_{n,m}$ in equation (4.1), one can intuitively comprehend that it is an estimator for the theoretical significance level. To express definition 4.2 in mathematical terminology, consider that we have $T_{1,n}, \dots, T_{m,n}$ as m independent realizations of the test statistic T_n. Recall that for the regression type tests we reject H_0 if T_n is smaller or equal to the critical value. Hence, we find

$$\hat{\alpha}_{n,m} = \frac{|\{i \in \{1, \dots, m\} : T_{i,n} \leq \hat{c}_{n,\alpha}\}|}{m}$$

for the empirical type I error, where $|A|$ means the cardinal numeral of the set A and $\hat{c}_{n,\alpha}$ is the empirical percentage point of the corresponding test for the sample size n and the significance level α obtained by empirical sampling studies. For tests of the Jarque-Bera type we have

$$\hat{\alpha}_{n,m} = \frac{|\{i \in \{1, \dots, m\} : T_{i,n} \geq \hat{c}_{n,\alpha}\}|}{m},$$

since for these types of test we reject H_0 when T_n is greater or equal the critical value. Finally, for m realizations $z_{1,W_R}, \dots, z_{m,W_R}$ of the extension of Royston we get

$$\hat{\alpha}_{n,m} = \frac{|\{i \in \{1, \dots, m\} : z_{i,W_R} \notin I^c\}|}{m},$$

where

$$I^c = \left(\Phi^{-1}\left(\frac{\alpha}{2}\right), \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \tag{4.2}$$

is the confidence intervall of a $N(0,1)$ distributed test statistic (cf. Falk et al. (2002, p.)). For the extension of Royston, the distribution of the test statistic z_{W_R} under H_0 is known so that an empirical analysis of the type I error rate is actually not really necessary. But since z_{W_R} is only approximate $N(0,1)$ distributed under the null hypothesis and since the parameters to calculate the statistic z_{W_R} are also only based on empirical simulation, one can argue that an empirical investigation is justified. To make the approach more comprehensible, we give a short example that should bring more light on the calculation of the empirical type I error.

4.3 Example. Assume that we have generated m = 10,000 independent $N(0,1)$ distributed random samples of size $n = 35$. After running the Shapiro-Wilk test and the Jarque-Bera test for each of the 10,000 samples on the 0.05 significance level, we obtain $M_W = 508$ rejections of H_0 for the Shapiro-Wilk test and $M_{JB} = 494$ rejections for the Jarque Bera test. Hence, the empirical type I error of the two tests are, respectively,

$$\frac{M_W}{m} = \frac{508}{10{,}000} = 5.08\% \quad \text{and} \quad \frac{M_{JB}}{m} = \frac{494}{10{,}000} = 4.94\%.$$

The Shapiro-Wilk test tends to exeed the significance level of 0.05 which can almost be regarded as negligible. On the contrary, the type I error rate of the Jarque Bera test is under 0.05 which makes the test very slightly conservative.

4.1.2 General Settings

Investigated Tests for Normality

For our empirical studies we included almost all tests for normality presented in this work. We omitted an investigation of the skewness and kurtosis test for normality since these tests are no omnibus tests and therefore not a recommended choice for practitioners. A short summary is given in table 4.1 where the notation of the test statistic is repeated. The notation is retained in the tables where the empirical results of the type I error investigations and the results of the power studies are presented.

Table 4.1: Used abbrevations for the tests of normality in the empirical power studies

Test Symbol	Test Name	Reference
W	Shapiro-Wilk test	definition 2.5 on page 12
SF	Shapiro-Francia test	definition 2.20 on page 32
WB	Weisberg-Bingham test	definition 2.22 on page 34
FB	Filliben test	definition 2.24 on page 35
RG	extension of Rahman and Govindarajulu	definition 2.27 on page 40
z_{W_R}	extension of Royston	definition 2.26 on page 39
JB	Jarque-Bera test	definition 3.17 on page 62
AJB	adjusted Jarque-Bera test	definition 3.18 on page 66

Settings for n, α and m

As it will be described more detailed in subsection 4.2.2, there is no standard setting that has been established in research for empirical simulation studies for tests of normality. In the conducted studies of this work we tried to choose our parameters as conventional as possible. The definitions of the testing parameters in the following convention should always be kept on mind when interpreting the results of the study in the next subsection.

4.4 Convention. *For the calculation of the empirical type I error $\hat{\alpha}_{n,m}$ of the tests for normality in table 4.1 we chosed the following parameters for our empirical investigations*

$$\begin{aligned} n &= 10, 15, 20, 25, 30, 35, 50, 100, 200, 500 \\ m &= 1{,}000{,}000. \end{aligned} \tag{4.3}$$

The critical values $\hat{c}_{n,\alpha}$ for the tests for normality were the points from our own simulation studies. We took the value $\alpha = 0.05$ for the empirical significance points. The exact values of $\hat{c}_{n,0.05}$ are tabled in the sections and subsections above where the corresponding test is presented.

The justification for the choice of the sample sizes n is that we wanted to investigate small sample sizes ($n = 10, 15, 20, 25, 30, 35$), moderate sample sizes ($n = 50, 100$) as well as large sample sizes ($n = 200, 500$). Since according to our own experiences, which has shown that the sample sizes in applied statistics are mostly rather too small than too large, we emphasis here on small sample sizes and calculated the empirical type I error for a broader number of different values for n than for moderate and large sample sizes. This is also consistent with the common practice in resarch where mostly small sample sizes are used in the empirical studies.

The choice for $\alpha = 0.05$ is in almost all studies the value for the significance level. For that reason, we also used this value for our simulation studies.

The number of repetitions in our study is $m = 1{,}000{,}000$. Note that this value is essentially higher that in most other studies performed in the past. Zhang (1999) used only 10,000 replications for their empirical studies while Seier (2002) and Keskin (2006) considered a number of repetitions of $m = 100{,}000$. By the best knowledge of the author, the most extensive empirical study in the past is the work of Bonett and Seier (2002) with a Monte Carlo simulation based on 300,000 repetitions. However, this work does not contain any of the tests for normality considerd in our simulation study in table 4.1.

Computational Issues

Before we present the results of the simulation studies some computational facts may be mentioned for the sake of completeness.

All simulations were done in the software R (R Development Core Team (2009)), the source code of the programms are not listed here—they can be obtained from the author by request.

The normal samples were generated in R with the function `rnorm()` and all random samples were generated independently from each other. For the calculation of the test statistic of the z_{W_R} test we used the already implemented function `shapiro.test()` in R. All other test statistics from table 4.1 were calculated by ourselves. For details of that calculation and the use of the needed parameters and coefficients we refer to the corresponding sections and subsections above where the computation is described in detail for each simulation of the empirical significance points. The decision whether to reject the null hypothesis or not was based on the critical values $\hat{c}_{n,\alpha}$ obtained from our own simulations studies also described above.

4.1.3 Results

For the tests of the regression type, the empirical type I error rates in the tables 4.2 and 4.3 show a unified view. For all sample sizes there are no critical results in the sense that the value of $\hat{\alpha}_{n,m}$ is essentially different to 0.05. The biggest discrepancy for these tests has the extension of Royston for a sample size of $n = 15$. The empirical type I error in this case is 0.0490. But this departure to 0.05 is only one unit in the third decimal place. Therefore we can ignore this fact, since it is without meaning for practial purposes, and state that for the regression tests for normality the type I error is well controlled.

Table 4.2: Empirical type I error rate for small sample sizes

Test	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$	$n = 35$
W	0.0502	0.0499	0.0503	0.0503	0.0499	0.0499
SF	0.0501	0.0499	0.0500	0.0492	0.0498	0.0505
WB	0.0503	0.0501	0.0503	0.0495	0.0496	0.0500
FB	0.0500	0.0498	0.0501	0.0499	0.0496	0.0499
RG	0.0498	0.0501	0.0499	0.0498	0.0497	0.0505
z_{W_R}	0.0501	0.0490	0.0503	0.0501	0.0505	0.0503
JB	0.0497	0.0497	0.0504	0.0498	0.0495	0.0500
AJB	0.0501	0.0505	0.0498	0.0499	0.0500	0.0500

The results for the SF test, the WB test and the FB test for $n \geq 100$ are also remarkable, since there have been conducted no simulations for such large sample sizes. We see that these tests also hold the significance level for larger sample sizes which is an indication that these tests can be recommended also for large sample sizes.

To underline the accuracy of our results due to the large number of $m = 1{,}000{,}000$ repetitions, we cite Seier (2002) who state that for $n = 50$, the W has an empirical significance level of 0.0432. According to Zhang (1999) the empirical type I error rate for the W test for the same sample size is 0.064. Considering the very exact value of 0.0506 in table 4.3, the big differences of the two results is probably mostly due the small number of repetitions comparing to our own simulation study.

The tests of the Jarque-Bera type yield to similar results. All three presented tests control the type I error very exactly. The biggest difference to 0.05 is the value 0.0494 which appears three times. Since this difference is only six units in the fourth decimal place it is even more negligible than for the regression type tests. Comparing these results with the results of Gel and Gastwirth (2008) we obtained even more accurate values for the empirical significance level. Once again this emphasizes the use of a large simulation number, since Gel and Gastwirth conducted the simulation study only with $m = 10{,}000$ repetitions.

Table 4.3: Empirical type I error rate for moderate and large sample sizes

Test	$n = 50$	$n = 100$	$n = 200$	$n = 500$
W	0.0506	—	—	—
SF	0.0503	0.0502	0.0495	—
WB	0.0501	0.0500	0.0494	0.0502
FB	0.0501	0.0499	0.0495	0.0502
RG	0.0498	0.0497	0.0500	0.0501
z_{W_R}	0.0499	0.0495	0.0491	0.0509
JB	0.0501	0.0506	0.0494	0.0501
AJB	0.0500	0.0494	0.0494	0.0499

4.2 Power Studies

4.2.1 Theoretical Background

Until now, in chapter 2 and chapter 3, we have introduced different tests for normality based on completely different approaches to the problem. One of the first natural questions that comes into mind is now how to compare the tests to state that one test is "better" than another, whatever "better" may stand for. According to Thode (2002), "the most frequent measure of the value of a test for normality is its power". To become more close, we will start with the definition of the power.

4.5 Definition. *The test decision of not rejecting the null hypothesis H_0 when in fact the alternative hypothesis H_1 is true, is called a* **type II error**. *Let the probability of making such an error be denoted with β.The* **power** $1 - \beta$ *of a statistical test is then the probability that it will not make a type II error. This is equivalent to the probability of rejecting a false null hypothesis.*

Hence, in the framework of testing for normality, the power is the ability to detect when a sample comes from a non-normal distribution. Consequently, given a sample from which we already know that it is not normally distributed, but follows an alternative distribution (e.g. a χ^2 distribution with 5 degrees of freedom, denoted by χ^2_5), tests with higher power should tend to reject H_0 more likely than tests with lower power. This is exactly the underlying idea when determining the power with an empirical simulation study. In this work, we are going to do empirical power studies by performing Monte Carlo simulations. The principle of an empirical power analysis is simple to describe. The general approach is given in the following definition.

4.6 Definition. *Let m be the number of randomly generated independent samples of sizes n where all m samples follow the same distribution that is not a normal distribution. The empirical power $1 - \hat{\beta}_{n,\alpha,m}$ of a given test for normality for a given significance level α is given by*

$$1 - \hat{\beta}_{n,\alpha,m} = \frac{M}{m},$$

where $M \leq m$ is the number of the m tests that reject the null hypothesis of a normally distributed sample on the significance level α.

The definition 4.6 is very intuitive. The ratio $\frac{M}{m}$ is an estimator for the theoretical power of the function. The higher the number of rejected tests, the higher is the ratio $\frac{M}{m}$ and, hence, the higher is the power. To come to a mathematical notation, consider that for a given sample size n we have $T_{1,n}, \ldots, T_{m,n}$ realizations of m test statistics for a test of normality. At this, the underlying samples of the m test statistics do not follow a normal distribution. Recall in mind that for a test of the regression type, H_0 is rejected if $T_{i,n} \leq c_{n,\alpha}, i = 1, \ldots, m$, where $c_{n,\alpha}$ is the critical value of the test for a given sample size n and a significance level α. Consequently, for a test of the regression type, M is given by

$$M = |\{i \in \{1, \ldots, m\} : T_{i,n} \leq c_{n,\alpha}\}|,$$

The empirical power is

$$1 - \hat{\beta}_{n,\alpha,m} = \frac{M}{m} = \frac{|\{i \in \{1, \ldots, m\} : T_{i,n} \leq c_{n,\alpha}\}|}{m}.$$

Hence, for a test form the Jarque-Bera type M is given by

$$M = |\{i \in \{1, \ldots, m\} : T_{i,n} \geq c_{n,\alpha}\}|$$

and for the empirical power we have

$$1 - \hat{\beta}_{n,\alpha,m} = \frac{M}{m} = \frac{|\{i \in \{1, \ldots, m\} : T_{i,n} \geq c_{n,\alpha}\}|}{m}.$$

For the special case of the extension of Royston we have m realizations $z_{1,W_R}, \ldots, z_{m,W_R}$ and thus we find

$$M = |\{i \in \{1, \ldots, m\} : z_{i,w_R} \notin I^c\}|,$$

with I^c given in equation (4.2). This yields to an empirical power of

$$1 - \hat{\beta}_{n,\alpha,m} = \frac{M}{m} = \frac{|\{i \in \{1, \dots, m\} : z_{i,w_R} \notin I^c\}|}{m}.$$

Again, an example attempts to demonstrate the approach of a power analysis for different test of normality and make it easier to understand

4.7 Example. Considering a χ^2_5 distribution as the alternative distribution, a sample size of $n = 40$ and a significance level of $\alpha = 0.05$, we want to compare the Shapiro-Wilk test and the Jarque Bera concerning their power for this distribution. To this end we have generated $m = 100{,}000$ independent samples of size 40 that follow a χ^2_5 distribution. When running the Shapiro-Wilk test at a significance level of 0.05 for each sample, we get the number of $M_W = 87{,}654$ rejections. While for performing the Jarque Bera test at the same significance level we get $M_{JB} = 76{,}543$ rejections. We can now compare the two empirical powers of the tests and find that

$$\frac{M_W}{m} = \frac{87{,}654}{100{,}000} = 0.87654 \approx 87.65\% > 76.54\% = 0.76543 = \frac{76{,}543}{100{,}000} = \frac{M_{JB}}{m}.$$

Thus, for this testing situation, the Shapiro-Wilk test has a higher power and therefore a "better" performance than the Jarque-Bera test.

4.2.2 General Settings of the Power Study

In this subsection we give a short overview about the settings made in the performed power studies in this work concerning alternative distributions, sample size, significance level and number of repetitions. We start with some interesting historical facts.

Development of the Design of Power Studies

The two seminal papers, that essentially set the standards for power studies and the development of tests for normality in general, were the ones of Shapiro and Wilk (1965) and Shapiro et al. (1968). The examination of the power was done like explained in the subsection before with 45 alternative distributions. The number of tests for normality in that time were relatively few, so that in the publication of 1968 the authors were able to conduct a very broad study to assess the value of the Shapiro-Wilk test.

Empirical power studies have not been performed before that time. The enforcement of this procedure was also not least due to the advantages of the calculation speed and the memory capacity of the new processors. In the following, when publishing a new test for normality or a development of an already established testing procedure, power studies following the design of the work of Shapiro and Wilk (1965) became almost obligatory.

With the increasing number of tests for normality that were developed especially in time from the middle of the 1970's to the middle of the 1980's, it became more and more extensive to compare the new tests with all other tests for normality already existing. Besides, the space limitations of the journals made complete studies impossible. Therefore, it became practice to compare a new test only with a subset of tests for normality.

A disregard in the development of power comparisons can be stated in the fact that no common standard design could be emerged. The reasons for this fact are the following:

- The value α of the significance level of the tests for normality varies in the different studies. The most popular values are 0.05 and 0.10. Thus, a comparison between two studies that used different α's is nearly infeasible.
- Different studies used different samples sizes n when generating the samples of the alternative distributions. Additionally, the number m of repetitions is changing from study to study.
- Another problem is the use of the critical points for the tests for normality. Some studies calculated their own critical points for the tests although critical point have already been available from other works.

Settings for n, α and m

For the calculation of the empricial power $1-\hat{\beta}_{n,\alpha,,m}$, the setting for the simulations parameters n, α and m were the same than for the empirical type I error investigations in convention 4.4, i.e.,

$$\begin{aligned} n &= 10, 15, 20, 25, 30, 35, 50, 100, 200, 500 \\ \alpha &= 0.05 \\ m &= 1{,}000{,}000. \end{aligned}$$

The justification for the choice of the sample sizes n is given in subsection 4.1.2, as well a the one for the choice of α. The number of repetitions in our study is with 1,000,000 again essentially higher that in most other studies performed in the past. While the first works about empirical power analysis the number of repetitions was very small (e.g. $m = 200$ in Shapiro and Wilk (1965)), even recent works used values for m that are substantially smaller that one million. For example in the power study in Deb and Sefton (1996) the value was $m = 100{,}000$ which seems to be one of the highest number of repetitions for such power studies. But even very recent works do not use values of m that are comparably high than the number of repetitions in our simulation (e.g. $m = 10{,}000$ in Thadewald and Büning (2007)).

Settings for the alternative distributions

Here the alternative distributions are shortly presented. We only give the probability densitiy function of the distribution and—as an orientation—their theoretical skewness and kurtosis value, if possible. All the information of the following distributions were taken from Johnson et al. (1994, 1970) and Johnson and Kotz (1969). Following Thode (2002) we group the alternative distributions into three subsets based on their different shape.

For the generation of the random samples following the alternative distributions, we used the corresponding functions in R, e.g. to generate a χ^2_5 distributed sample of size $n = 25$ we used the command `rchisq(25, 5)`.

Long-Tailed Symmetric Distributions

Consider a continuous random variable Y. Unless stated otherwise the skewness of the pdf of Y is $\sqrt{\beta_1} = 0$.

- Y is Laplace distributed with parameters μ and $\sigma > 0$, denoted by $Laplace(\mu, \sigma)$, if its pdf is

$$p_{Laplace(\mu,\sigma)}(y) = \frac{1}{2\sigma} \exp\left(-\frac{|y-\mu|}{\sigma}\right), \quad y \in \mathbb{R}.$$

The kurtosis is $\beta_2 = 6$.

- Y is logistic distributed with parameters α and $\beta > 0$, denoted by $l(\alpha, \beta)$ if its pdf is

$$p_{l(\alpha,\beta)}(y) = \frac{\exp\left(-\frac{y-\alpha}{\beta}\right)}{\beta\left(1+\exp\left(-\frac{y-\alpha}{\beta}\right)\right)^2}, \quad y \in \mathbb{R}.$$

The kurtosis is $\beta_2 = 4.2$.

- Y is t distributed with ν degrees of freedom, denoted by t_ν, if its pdf is

$$p_{t_\nu}(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}, \quad y \in \mathbb{R},$$

where

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u)\, \mathrm{d}u, \quad u > 0,$$

is the Gamma function. The kurtosis is $\beta_2 = \frac{3\nu-6}{\nu-4}, \nu \geq 5$. For $n \leq 4$ we have $\beta_2 = \infty$. The t_1 distribution is also known as the Cauchy distribution.

Short-Tailed Symmetric Distributions

Consider a continuous random variable Y. Unless stated otherwise the skewness of the pdf of Y is $\sqrt{\beta_1} = 0$.

- Y is beta distributed with parameters $p > 0$ and $q > 0$, denoted by $Beta(p, q)$ if its pdf is

  $$p_{Beta(p,q)}(y) = \begin{cases} \frac{1}{B(p,q)} y^{p-1}(1-y)^{q-1} & , \ 0 \le y \le 1 \\ 0 & , \ \text{else}, \end{cases}$$

  where

  $$B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 u^{p-1}(1-u)^{q-1} \, \mathrm{d}u$$

  is the beta function. The kurtosis is

  $$\beta_2 = \frac{3(p+q+1)(2(p+q)^2 + pq(p+q-6))}{pq(2+p+q)(3+p+q)}$$

  and the skewness is $\sqrt{\beta_1} = (2(q-p)\sqrt{p+q+1})/((p+q+2)\sqrt{pq})$. Hence, for $p = q$ the beta distribution is symmetric, but for $p \neq q$ the beta distribution is asymmetric.

- Y is uniformly distributed on $(a, b) \subset \mathbb{R}$, denoted by $U(a, b)$, if its pdf is

  $$p_{U(a,b)}(y) = \frac{1}{a-b} 1_{(a,b)}(y), \quad y \in \mathbb{R}, -\infty < a < b < \infty,$$

  where

  $$1_{(a,b)}(y) = \begin{cases} 1 & , \ y \in (a, b) \\ 0 & , \ \text{else} \end{cases}$$

  is the indicator function on (a, b). The kurtosis is $\beta_2 = 1.8$.

Now consider a random variable Y that is discrete distributed. Y is binomially distributed with parameters n and $p \in (0, 1)$, denoted by $B(n, p)$ if we have

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, , \ldots, n.$$

The kurtosis is $\beta_2 = 3 - (6/n) + 1/(np(1-p))$ and the skewness is $\sqrt{\beta_1} = (1-2p)/(\sqrt{np(1-p)})$. Hence, for $p = (1-p) = 0.5$ the binomial distribution is symmetric.

Asymmetric Distributions

Consider a continuous random variable Y.

- Y is χ^2 distributed with ν degrees of freedem, denoted by χ^2_ν, if its pdf is

$$p_{\chi^2_\nu}(y) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{(\nu/2)-1} \exp\left(-\frac{y}{2}\right), & y > 0 \\ 0, & y \leq 0 \end{cases}.$$

  The skewness is $\sqrt{\beta_1} = \frac{2\sqrt{2}}{\sqrt{\nu}}$, the kurtosis is $\beta_2 = 3 + \frac{12}{\nu}$.

- Y is exponentially distributed with parameter $\lambda > 0$, denoted by $\exp_\lambda$, if its pdf is

$$p_{\exp_\lambda} = \begin{cases} \lambda \exp(-\lambda y), & y \geq 0 \\ 0, & y < 0 \end{cases}.$$

  The skewness is $\sqrt{\beta_1} = 2$, the kurtosis is $\beta_2 = 9$

- Y is lognormal distributed with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, denoted by $LogN(\mu, \sigma)$, if its pdf is

$$p_{LogN(\mu,\sigma)}(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{\log(y-\mu)^2}{2\sigma^2}\right) & , x > 0 \\ 0 & , x \leq 0 \end{cases}$$

  The skewness is $\sqrt{\beta_1} = (\exp(\sigma^2) + 2)\sqrt{\exp(\sigma^2) - 1}$, the kurtosis is $\beta_2 = \exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 3$

- Y is Weibull distributed with parameters $\alpha > 0$ and $\beta > 0$, denoted by $Wei(\alpha, \beta)$, if its pdf is

$$p_{Wei(\alpha,\beta)}(y) = \begin{cases} \alpha\beta y^{\beta-1} \exp(-\alpha y^\beta), & y > 0 \\ 0, & y \leq 0 \end{cases}$$

  The skewness is

$$\sqrt{\beta_1} = \frac{\Gamma_3(\beta) - 3\Gamma_2(\beta)\Gamma_1(\beta) + 2(\Gamma_1(\beta))^3}{(\Gamma_2(\beta) - (\Gamma_1(\beta))^2)^{3/2}}$$

  and

$$\beta_2 = \frac{\Gamma_4(\beta) - 4\Gamma_3(\beta)\Gamma_1(\beta) + 6\Gamma_2(\beta)(\Gamma_1(\beta))^2 - 3(\Gamma_1(\beta))^4}{(\Gamma_2(\beta) - (\Gamma_1(\beta))^2)^2},$$

  where $\Gamma_i(\beta) = \Gamma(1 + i\beta^{-1})$.

Now consider a random variable Y that is discrete distributed. Y is Poisson distributed with parameters $\lambda > 0$, denoted by P_λ if we have

$$P(Y = y) = \frac{\lambda^y}{y!} \exp(-\lambda), \quad y = 0, 1, 2, \ldots.$$

The skewness is $\sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$, the kurtosis is $\beta_2 = 3 + \frac{1}{\lambda}$

4.2.3 Results

Now we finally present the results of the power studies grouped by the classification for the alternative distributions as described before. For each table we will give a short comment about interesting facts and compare the results with the results of power studies from other papers in the past. The tests for normality used in the empirical power studies were the same tests than for the investigation of the empirical significance level in section 4.1. For an overview take a look at table 4.1.

Long-Tailed Symmetric Distributions

The results for the long-tailed symmetric alternative distributions are given in tables 4.4, 4.5 and 4.6. The first thing we notice is that in most situations, the tests based on the sample moments, JB and AJB, have a substantially higher power than the regression type tests. For the Laplace distribution, the modifications of the Shapiro-Wilk test, SF, WB and FB have a little bit higher power for sample sizes $n \geq 25$ than the JB and the AJB test. But for the logistic distribution and for the t distribution, in particular for the t distribution with high degrees of freedom, both Jarque-Bera type tests have the highest power of the considered tests. It is also notable that the AJB test always performes better than the JB test. The difference is, however, slightly which agrees with the statement of Thadewald and Büning (2007) that for this type of alternative distributions it makes almost no difference which of the two tests is taken.

The power of the W test is for this type of alternative distribution in all situations less than the power of the Jarque-Bera type tests which has also been stated by Thadewald and Büning (2007) and Deb and Sefton (1996). It is remarkable that in accordance with the results of Filliben (1975), the FB test performes in almost all cases better than the W test, in particular for the Laplace distribution the difference is quite large. But not only the FB test shows better results than the Shapiro-Wilk test. Its modifications SF and WB also have in all situations essentially higher power than the W test. Very similar results to the W test shows the z_{W_R} test, there are only slightly differences between the original procedure, W, and its extension to large sample sizes. These results do not completely agree with the results of Seier (2002) who state that for $n = 50$ the z_{W_R} test has higher power than the W test. The RG test shows the worst performance of all considered tests in this study. This is in accordance with Farrell and Rogers-Stewart (2006) who state that the W test peformes better than the RG test for symmetric long-tailed alternatives.

Short-Tailed Symmetric Distributions

The results of the short-tailed symmetric alternativ distributions given in tables 4.7, 4.8 and 4.9 do not give a unified picture. Due to the different power behaviours of the considered tests for

Table 4.4: Empirical power for symmetric long-tailed alternative distributions for $n = 10, 15, 20, 25$.

Alternative	$\sqrt{\beta_1}$	β_2	$n=10$								$n=15$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Laplace(0,0.5)$	0	6	15.3	17.9	17.8	18.1	11.5	15.3	17.9	18.4	21.0	25.1	25.0	25.4	13.7	20.8	24.5	25.6
$Laplace(0,1)$	0	6	15.3	17.9	17.9	18.1	11.5	15.3	18.0	18.5	21.0	25.1	25.1	25.4	13.7	20.9	24.6	25.6
$l(0,0.5)$	0	4.2	8.2	9.1	9.1	9.2	6.7	8.1	9.7	9.9	10.1	11.7	11.7	11.8	7.2	10.0	12.5	12.8
$l(0,1)$	0	4.2	8.1	9.1	9.1	9.1	6.7	8.1	9.7	9.9	10.0	11.6	11.6	11.8	7.2	9.9	12.5	12.8
t_1	0	∞	59.1	62.7	62.6	62.9	52.5	59.1	58.7	59.6	76.5	79.7	79.7	79.9	68.0	76.4	76.1	77.4
t_2	0	∞	29.7	32.8	32.7	33.0	24.6	29.7	32.6	33.2	42.6	46.9	46.8	47.2	33.5	42.4	46.2	47.3
t_4	0	∞	13.8	15.6	15.6	15.7	10.9	13.8	16.4	16.8	19.3	22.2	22.2	22.4	13.8	19.2	23.2	23.8
t_6	0	6	9.7	10.9	10.9	11.0	7.8	9.6	11.6	11.9	12.7	14.6	14.6	14.8	9.0	12.6	15.6	16.0
t_8	0	4.5	8.1	8.9	8.9	9.0	6.7	8.0	9.5	9.7	10.0	11.5	11.5	11.6	7.4	9.9	12.3	12.7
t_{10}	0	4	7.2	7.9	7.9	8.0	6.2	7.2	8.4	8.7	8.6	9.8	9.8	9.8	6.6	8.5	10.5	10.8
t_{15}	0	3.55	6.3	6.7	6.7	6.8	5.6	6.2	7.1	7.2	7.0	7.8	7.8	7.9	5.8	7.0	8.3	8.5
t_{20}	0	3.38	5.9	6.3	6.3	6.3	5.4	5.9	6.5	6.5	6.4	7.0	7.0	7.0	5.5	6.4	7.4	7.5

Alternative	$\sqrt{\beta_1}$	β_2	$n=20$								$n=25$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Laplace(0,0.5)$	0	6	26.2	31.4	31.4	31.8	15.5	26.2	30.3	31.4	31.1	36.9	37.0	37.7	17.2	31.2	35.4	36.7
$Laplace(0,1)$	0	6	26.2	31.4	31.4	31.9	15.4	26.2	30.3	31.4	31.1	36.9	36.9	37.7	17.2	31.1	35.3	36.7
$l(0,0.5)$	0	4.2	11.7	13.9	13.9	14.1	7.5	11.7	14.9	15.2	13.2	15.7	15.8	16.1	7.7	13.2	17.0	17.4
$l(0,1)$	0	4.2	11.7	13.9	13.9	14.1	7.5	11.7	14.9	15.3	13.2	15.7	15.8	16.1	7.7	13.2	16.9	17.3
t_1	0	∞	86.7	89.1	89.1	89.3	78.6	86.6	86.2	87.1	92.5	94.1	94.1	94.3	86.0	92.5	92.0	92.7
t_2	0	∞	53.0	57.6	57.6	58.0	40.8	52.9	56.8	57.8	61.5	66.2	66.2	66.7	47.4	61.5	65.2	66.3
t_4	0	∞	24.1	27.8	27.8	28.1	16.0	24.1	28.9	29.5	28.5	32.8	32.8	33.3	18.1	28.5	34.0	34.9
t_6	0	6	15.4	18.0	18.0	18.2	10.0	15.4	19.2	19.6	14.2	17.8	20.9	21.3	10.8	17.8	22.3	22.9
t_8	0	4.5	11.7	13.7	13.7	13.9	7.7	11.7	14.8	15.0	9.1	13.3	15.6	16.0	8.1	13.3	16.9	17.3
t_{10}	0	4	9.8	11.3	11.3	11.5	6.8	9.8	12.3	12.6	10.9	12.8	12.8	13.1	6.9	11.0	13.9	14.1
t_{15}	0	3.55	7.7	8.7	8.7	8.8	5.8	7.7	9.4	9.6	8.3	9.5	9.5	9.7	5.8	8.3	10.3	10.5
t_{20}	0	3.38	6.9	7.7	7.7	7.7	5.5	6.9	8.2	8.2	7.2	8.1	8.1	8.2	5.4	7.3	8.8	8.8

Table 4.5: Empirical power for symmetric long-tailed alternative distributions for $n = 30, 35, 50, 100$.

Alternative	$\sqrt{\beta_1}$	β_2	$n=30$								$n=35$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Laplace(0, 0.5)$	0	6	35.6	42.2	42.1	42.9	18.9	35.8	40.1	41.7	39.9	46.9	46.8	47.5	20.8	40.1	44.3	46.0
$Laplace(0, 1)$	0	6	35.6	42.2	42.1	42.8	18.9	35.7	40.0	41.6	40.0	47.0	46.9	47.6	20.9	40.1	44.4	46.0
$l(0, 0.5)$	0	4.2	14.6	17.7	17.7	18.0	7.9	14.6	19.0	19.6	15.9	19.4	19.3	19.7	8.1	15.9	20.8	21.4
$l(0, 1)$	0	4.2	14.5	17.7	17.7	18.0	7.8	14.6	19.0	19.5	15.8	19.4	19.3	19.7	8.0	15.9	20.8	21.4
t_1	0	∞	95.9	96.9	96.9	97.0	90.9	95.9	95.5	95.9	97.7	98.3	98.3	98.4	94.2	97.7	97.4	97.7
t_2	0	∞	68.4	72.9	72.9	73.4	53.2	68.4	71.8	73.1	74.3	78.5	78.5	78.9	58.7	74.4	77.4	78.4
t_4	0	∞	32.6	37.5	37.5	38.0	20.0	32.7	38.7	39.7	36.5	41.8	41.7	42.3	22.1	36.5	43.0	44.0
t_6	0	4.5	20.0	23.9	23.8	24.2	11.4	20.2	25.4	26.0	22.2	26.4	26.3	26.7	12.2	22.2	28.0	28.8
t_8	0	4	14.6	17.5	17.4	17.8	8.3	14.7	18.9	19.4	16.1	19.4	19.3	19.6	8.7	16.1	20.8	21.3
t_{10}	0	3.55	11.9	14.3	14.2	14.4	7.0	12.0	15.4	15.7	12.9	15.5	15.4	15.7	7.2	12.9	16.8	17.2
t_{15}	0	3.38	8.9	10.4	10.3	10.5	5.8	8.9	11.2	11.5	9.3	11.1	11.0	11.1	5.8	9.4	11.9	12.3
t_{20}	0	0.38	7.5	8.7	8.7	8.8	5.3	7.6	9.4	9.5	7.9	9.2	9.1	9.3	5.4	7.9	9.9	10.0

Alternative	$\sqrt{\beta_1}$	β_2	$n=50$								$n=100$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Laplace(0, 0.5)$	0	6	52.1	59.0	59.1	59.9	26.7	52.1	55.5	57.2	—	83.9	84.1	84.6	48.9	79.6	79.9	80.9
$Laplace(0, 1)$	0	6	52.0	58.9	59.0	59.9	26.7	52.0	55.5	57.2	—	83.9	84.1	84.6	49.0	79.6	80.0	81.0
$l(0, 0.5)$	0	4.2	19.6	24.0	24.0	24.6	8.5	19.5	25.8	26.6	—	36.5	36.7	37.5	10.6	30.4	39.4	40.2
$l(0, 1)$	0	4.2	19.7	24.1	24.1	24.6	8.5	19.6	25.8	26.6	—	36.5	36.7	37.4	10.6	30.4	39.4	40.2
t_1	0	∞	99.6	99.7	99.7	99.8	98.6	99.6	99.5	99.6	—	100	100	100	100	100	100	100
t_2	0	∞	86.3	89.1	89.1	89.4	72.1	86.3	88.2	89.0	—	99.9	99.0	99.0	93.4	98.6	98.8	98.9
t_4	0	∞	46.9	52.6	52.6	53.3	27.5	46.8	54.0	55.1	—	76.0	76.2	76.7	44.7	71.0	77.3	78.1
t_6	0	6	28.4	33.3	33.4	34.0	14.1	28.3	35.4	36.2	—	51.2	51.5	52.2	20.7	45.0	54.1	54.9
t_8	0	4.5	19.9	23.9	24.0	24.4	9.4	19.8	25.9	26.5	—	36.5	36.7	37.4	12.1	30.8	39.6	40.4
t_{10}	0	4	15.6	18.9	18.9	19.3	7.4	15.5	20.6	21.1	—	27.8	28.0	28.6	8.4	22.9	30.7	31.3
t_{15}	0	3.55	10.7	12.8	12.8	13.1	5.7	10.6	12.1	14.5	—	17.6	17.7	18.1	5.5	14.2	19.8	20.2
t_{20}	0	3.38	8.8	10.3	10.3	10.5	5.2	8.7	11.3	11.6	—	13.4	13.4	13.4	4.8	10.9	15.1	15.3

Table 4.6: Empirical power for symmetric long-tailed alternative distributions for $n = 200, 500$.

Alternative	$\sqrt{\beta_1}$	β_2	$n = 200$								$n = 500$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Laplace(0, 0.5)$	0	6	—	98.1	98.2	98.3	83.9	97.5	96.6	96.9	—	—	100	100	100	100	100	100
$Laplace(0, 1)$	0	6	—	98.2	98.2	98.3	83.8	97.5	96.6	96.9	—	—	100	100	100	100	100	100
$l(0, 0.5)$	0	4.2	—	55.4	56.0	56.9	17.2	49.1	59.1	60.0	—	—	87.1	87.6	45.5	83.7	89.0	89.4
$l(0, 1)$	0	4.2	—	55.6	56.1	57.1	17.2	49.2	59.2	60.2	—	—	87.1	87.6	45.5	83.7	89.0	89.4
t_1	0	∞	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
t_2	0	∞	—	100	100	100	100	100	100	100	—	—	100	100	100	99.8	100	100
t_4	0	∞	—	94.3	94.4	94.6	72.9	92.5	94.8	95.1	—	—	100	100	98.7	99.9	100	100
t_6	0	6	—	74.0	74.4	75.1	35.6	68.9	76.8	77.6	—	—	96.9	97.1	75.9	95.8	97.6	97.7
t_8	0	4.5	—	55.0	55.5	56.3	18.7	48.8	59.0	59.8	—	—	85.7	86.3	44.4	82.0	88.2	88.6
t_{10}	0	4	—	41.8	42.3	43.1	11.5	35.8	46.0	46.9	—	—	71.1	71.9	25.4	65.4	75.1	75.8
t_{15}	0	3.55	—	25.0	25.3	26.0	5.8	20.5	28.3	29.0	—	—	43.4	44.3	9.0	37.1	48.1	48.8
t_{20}	0	3.38	—	17.9	18.1	18.6	4.5	14.5	20.4	20.9	—	—	29.4	30.2	5.3	24.2	33.2	33.9

the uniform distribution, we have to consider these results separately.

First, it is obviously that the power of the Jarque-Bera type tests is surprisingly poor for the beta and the binomial distribution. In particular, the results for the beta distribution for sample sizes $n < 100$ are very poor, for the AJB test the power is sometimes even zero. Jahanshahi et al. (2008) obtained comparably results for this type of distribution. However, in the power study of Yazici and Yolacan (2007), the authors report about very good results for the beta distribtion with a power near 1 using the Jarque-Bera test—their results my be casted on doubt. For the binomial distribution the power for the JB and the AJB test are also very poor, even for large samples sizes the power does not increase. These results are in contradiction to the ones found by Jahanshahi et al. (2008), who report about higher values for the binomial distribution. Better results for the beta and the binomial distribution are achieved by the regression type tests, for all tests the power is in every situation higher than for the Jarque-Bera type tests. For the binomial and the beta distribution W and z_{W_R} show similar results. The RG test has mostly a lower power than the W test for the binomial distribution which has also been found out by Rahman and Govindarajulu (1997). For the beta distribution, however, the power of the RG test is the best of all considered tests, which confirms the findings of Farrell and Rogers-Stewart (2006). In contradiction to the work of Yazici and Yolacan (2007) the power values for the W test in our power study do not decrease for larger sample sizes. We also note, that the original W test performes most of the times better than the modifications SF, WB and FB.

On the contrary, for the uniform distribution, we find a completely opposite power behaviour. The JB and the AJB test have permanently a higher power than the regression type tests. Jahanshahi et al. (2008) found power values of the JB and the AJB test that decrease for larger samples sizes, so we cannot agree with their results. However, it may be mentioned that they only considered the $U(2, 4)$ distribution which is not a popular chioce for an alternative distribution compared to other power studies. Hence, our results are not exactly comparable to their findings. The three modifications of the W test show for some sample sizes ($n = 30, 35, 50$) all substantially higher power than the original W test and the RG test has the worst performance of all tests. This is in contradiction to the study of Rahman and Govindarajulu (1997) who found that the RG test has higher power than the Shapiro-Wilk test. We also have to disagree to the statement of Rahman and Govindarajulu that the SF shows lower power results than the W test.

Asymmetric Distributions

The power results for asymmetric alternative distributions are presented in tables 4.10, 4.11 and 4.12. The power behaviour in these cases are much more in accord with the results of other power studies than the results for the symmetric alternative distributions.

The highest power is achieved by the W test and its two extensions, the RG test and the z_{W_R} test. Like for the symmetric beta distributions, the RG has highest power for the asymmetric

Table 4.7: Empirical power for symmetric short-tailed alternative distributions for $n = 10, 15, 20, 25$.

Alternative	$\sqrt{\beta_1}$	β_2	$n=10$								$n=15$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(0.5, 0.5)$	0	1.5	29.8	26.6	17.2	15.5	42.2	29.5	2.7	1.8	52.0	29.6	30.4	27.0	70.5	51.6	1.1	0.6
$Beta(1.1, 1.1)$	0	1.85	7.2	3.9	4.0	3.7	11.8	7.1	1.6	1.4	10.7	4.3	4.5	3.9	20.5	10.5	0.6	0.5
$Beta(1.5, 1.5)$	0	2	5.1	3.1	3.1	2.9	8.0	5.0	1.7	1.5	6.1	2.7	2.8	2.4	12.2	6.0	0.7	0.6
$Beta(2, 2)$	0	2.14	4.2	2.9	2.9	2.8	6.3	4.2	1.9	1.8	4.6	2.3	2.4	2.1	8.6	4.5	0.9	0.8
$U(0, 1)$	0	1.8	8.1	9.1	9.1	9.2	6.6	8.1	9.7	9.9	1.9	11.6	11.6	11.7	7.1	9.8	12.4	12.8
$U(-1, 1)$	0	1.8	8.2	9.2	9.1	9.2	6.7	8.2	9.7	9.9	10.0	11.7	11.6	11.8	7.2	9.9	12.4	12.8
$U(-2, 2)$	0	1.8	8.2	9.1	9.1	9.2	6.7	8.1	9.7	9.9	10.0	11.7	11.6	11.8	7.2	9.9	12.5	12.9
$B(4, 0.5)$	0	2.5	47.3	37.1	40.4	34.2	45.3	47.3	3.2	4.1	62.1	51.0	52.6	49.9	57.8	59.8	2.2	1.9
$B(8, 0.5)$	0	2.75	21.0	17.0	17.8	16.1	21.3	21.0	3.8	3.9	25.9	20.7	21.1	20.3	27.7	25.4	3.5	3.3
$B(10, 0.5)$	0	2.8	16.8	13.6	14.1	13.1	17.3	16.8	4.0	4.1	20.0	16.2	16.4	15.9	22.0	19.7	3.8	3.7
$B(20, 0.5)$	0	2.9	9.9	8.5	8.7	8.3	10.5	9.9	4.5	4.6	10.7	9.3	9.3	9.1	11.8	10.6	4.4	4.4
Alternative	$\sqrt{\beta_1}$	β_2	$n=20$								$n=25$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(0.5, 0.5)$	0	1.5	72.3	47.2	47.9	43.3	88.1	72.4	0.6	0.2	86.5	65.0	65.4	61.1	96.1	86.6	0.5	0.1
$Beta(1.1, 1.1)$	0	1.85	15.9	5.8	6.0	5.0	31.4	15.9	0.3	0.2	22.6	8.3	8.4	7.0	43.6	22.6	0.1	0.1
$Beta(1.5, 1.5)$	0	2	8.1	3.0	3.0	2.6	17.7	8.2	0.3	0.3	10.9	3.7	3.7	3.1	24.2	10.9	0.1	0.1
$Beta(2, 2)$	0	2.14	5.3	2.2	2.2	1.9	11.3	5.3	0.4	0.3	6.4	2.2	2.3	2.0	14.6	6.4	0.2	0.2
$U(0, 1)$	0	1.8	11.7	13.9	13.9	14.1	7.5	11.7	14.9	15.2	13.2	15.8	15.8	16.2	7.7	13.2	17.0	17.4
$U(-1, 1)$	0	1.8	11.6	13.8	13.8	14.0	7.5	11.6	14.8	15.1	13.3	15.9	15.9	16.2	7.7	13.3	17.1	17.5
$U(-2, 2)$	0	1.8	11.7	13.9	13.9	14.1	7.5	11.7	14.9	15.2	13.2	15.8	15.8	16.2	7.7	13.2	17.0	17.5
$B(4, 0.5)$	0	2.5	70.9	66.8	66.8	66.1	66.2	71.6	1.1	0.9	95.0	84.5	84.9	83.0	86.2	95.0	0.7	0.6
$B(8, 0.5)$	0	2.75	33.1	27.0	27.1	26.0	33.9	33.3	2.9	2.9	44.4	35.0	35.3	33.9	42.0	44.4	2.6	2.4
$B(10, 0.5)$	0	2.8	24.9	20.2	20.3	19.5	26.1	25.0	3.3	3.2	31.8	25.0	25.2	24.3	31.3	31.9	3.1	2.9
$B(20, 0.5)$	0	2.9	12.1	10.2	10.2	9.9	13.2	12.1	4.2	4.1	13.7	11.2	11.2	11.0	14.6	13.7	4.0	3.9

Table 4.8: Empirical power for symmetric short-tailed alternative distributions for $n = 30, 35, 50, 100$.

Alternative	$\sqrt{\beta_1}$	β_2	$n = 30$								$n = 35$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(0.5, 0.5)$	0	1.5	94.1	79.8	79.7	75.9	98.9	94.2	0.5	0.1	97.8	89.6	89.3	86.6	99.7	97.8	1.1	0.1
$Beta(1.1, 1.1)$	0	1.85	30.2	12.1	12.0	9.9	55.5	30.5	0.1	0	38.9	16.9	16.5	13.6	66.7	39.1	0.1	0
$Beta(1.5, 1.5)$	0	2	14.0	4.7	4.7	3.8	31.2	14.1	0.1	0.1	17.8	6.2	6.0	4.8	39.1	17.9	0.1	0
$Beta(2, 2)$	0	2.14	7.6	2.6	2.6	2.1	18.2	7.7	0.1	0.1	9.2	3.1	3.0	2.4	22.4	9.3	0.1	0.1
$U(0, 1)$	0	1.8	14.5	17.7	17.6	18.0	7.8	14.5	19.0	19.6	15.9	19.5	19.4	19.8	8.1	16.0	20.9	21.5
$U(-1, 1)$	0	1.8	14.5	17.7	17.6	18.0	7.8	14.6	18.9	19.5	15.9	19.5	19.4	19.7	8.1	16.0	20.8	21.4
$U(-2, 2)$	0	1.8	14.5	17.7	17.6	18.0	7.8	14.6	19.0	19.6	15.8	19.4	19.3	19.7	8.1	15.9	20.8	21.4
$B(4, 0.5)$	0	2.5	100	98.0	98.0	96.5	100	100	0.4	0.3	100	100	100	100	100	100	0.3	0.2
$B(8, 0.5)$	0	2.75	53.9	46.0	45.8	43.8	49.5	54.3	2.3	2.2	64.8	56.7	56.2	54.0	57.6	65.0	2.0	1.9
$B(10, 0.5)$	0	2.8	38.7	31.8	31.7	30.2	36.8	39.0	2.8	2.7	46.9	39.2	38.8	37.0	43.0	74.1	2.6	2.5
$B(20, 0.5)$	0	2.9	15.5	12.7	12.7	12.2	16.4	15.7	3.9	3.9	17.7	14.4	14.2	13.7	18.6	17.8	3.7	3.7

Alternative	$\sqrt{\beta_1}$	β_2	$n = 50$								$n = 100$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(0.5, 0.5)$	0	1.5	99.9	99.2	99.1	98.7	100	99.9	39.6	8.0	—	100	100	100	100	100	100	99.9
$Beta(1.1, 1.1)$	0	1.85	65.0	36.1	35.3	30.4	88.8	64.5	0.4	0	—	92.3	91.5	88.6	100	98.7	60.9	39.2
$Beta(1.5, 1.5)$	0	2	32.5	12.8	12.4	10.0	61.5	32.0	0.1	0	—	55.3	53.3	47.3	97.1	80.4	19.9	8.1
$Beta(2, 2)$	0	2.14	15.5	5.2	5.0	4.0	35.8	15.2	0.1	0	—	22.3	20.9	17.2	78.1	45.3	0.5	0.1
$U(0, 1)$	0	1.8	19.7	24.1	24.1	24.6	8.5	19.6	25.9	26.6	—	36.5	36.8	37.5	10.7	30.4	39.4	40.3
$U(-1, 1)$	0	1.8	19.7	24.0	24.1	24.6	8.5	19.6	25.8	26.6	—	36.5	36.7	37.5	10.7	30.4	39.4	40.2
$U(-2, 2)$	0	1.8	19.6	24.0	24.0	24.6	8.5	19.5	25.8	26.5	—	36.5	36.8	37.5	10.6	30.4	39.4	40.3
$B(4, 0.5)$	0	2.5	100	100	100	100	100	100	0.1	0.1	—	100	100	100	100	100	0.1	0.1
$B(8, 0.5)$	0	2.75	93.2	86.6	86.3	84.4	81.2	93.0	1.5	1.4	—	100	100	100	100	100	0.8	0.7
$B(10, 0.5)$	0	2.8	76.5	66.7	66.3	63.7	63.9	76.1	2.1	2.0	—	100	100	100	99.0	100	1.3	1.2
$B(20, 0.5)$	0	2.9	26.3	20.9	20.7	19.8	25.2	26.0	3.5	3.4	—	63.1	62.3	59.5	54.8	71.1	3.0	2.9

Table 4.9: Empirical power for symmetric short-tailed alternative distributions for $n = 200, 500$.

Alternative	$\sqrt{\beta_1}$	β_2	$n = 200$								$n = 500$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(0.5, 0.5)$	0	1.5	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$Beta(1.1, 1.1)$	0	1.85	—	100	100	100	100	100	100	99.9	—	—	100	100	100	100	100	100
$Beta(1.5, 1.5)$	0	2	—	98.7	98.4	97.6	100	99.9	96.1	93.0	—	—	100	100	100	100	100	100
$Beta(2, 2)$	0	2.14	—	76.8	74.5	69.4	99.5	92.3	68.1	57.3	—	—	100	100	100	100	100	100
$U(0, 1)$	0	1.8	—	55.5	56.0	57.0	17.2	49.1	59.2	60.2	—	—	87.1	87.6	45.6	83.7	89.0	89.5
$U(-1, 1)$	0	1.8	—	55.6	56.1	57.0	17.2	49.1	59.2	60.2	—	—	87.1	87.6	45.5	83.6	89.0	89.4
$U(-2, 2)$	0	1.8	—	55.5	56.0	57.0	17.2	49.1	59.1	60.1	—	—	87.1	87.6	45.6	83.6	89.0	89.4
$B(4, 0.5)$	0	2.5	—	100	100	100	100	100	1.4	0.7	—	—	100	100	100	100	46.4	40.3
$B(8, 0.5)$	0	2.75	—	100	100	100	100	100	0.9	0.6	—	—	100	100	100	100	4.5	3.6
$B(10, 0.5)$	0	2.8	—	100	100	100	100	100	1.1	0.9	—	—	100	100	100	100	3.1	2.5
$B(20, 0.5)$	0	2.9	—	100	100	100	97.3	100	2.5	2.4	—	—	100	100	100	100	2.7	2.4

beta distributions and a slightly higher power than W and z_{W_R} for some χ^2 distributions with small degrees of freedom. With these results we can agree with Rahman and Govindarajulu (1997) and Farrell and Rogers-Stewart (2006) who found similar results. The power of W and z_{W_R} is almost identical for all distributions and samples sizes which has also been stated by Seier (2002). The modifications of the W test show slightly less power behaviour but have still substantially better power than the Jarque-Bera typed tests. The worst performance for asymmetric distributions show the JB and the AJB test. Their power is in all situations distinctly smaller than for the regression type tests. Thus, our findings confirm the results given by Jarque and Bera (1987) and Deb and Sefton (1996). Also note that for some cases the power for the AJB test is substantially smaller than for the JB test whereas for the symmetric alternatives there were only slight differences between the two tests.

4.3 Recommendations

In order to summarize the results of our power study and to give some recommendations for using a test for normality we have to reflect the results in subsetion 4.2.3 above. First we want to mention that one objective of the study was to extend the modifications of the Shapiro-Wilk test, the SF test, the WB test and the FB test for larger sample sizes. Until now there were only used for relatively small sample sizes ($n \leq 100$) and by the best of the authors knowledge, they have not been extended to large values of n. Such an extension including the calculation of empricial significance points like it was done in section 2.4 seems justifiable since the power behaviour does not change esentially the already known power properties of the test discussed in preceding power studies. In some cases, the modifications performed even better than the original W test, e.g. for long-tailed symmetric distributions. However, for symmetric short-tailed and for asymmetric alternative distributions, the power of the original procedure was higher most of the time.

Another objective was to assess the power of the two extensions RG and z_{W_R} simultaneously in one power study since it seems that this has not been done before. Seier (2002) only compared the W test and the z_{W_R} test, whereas Rahman and Govindarajulu (1997) and Farrell and Rogers-Stewart (2006) both only compared the W test and the RG test with each other. In this study, a comparison for all three tests was done with the result that these three tests have the highest power in most situations. The agreement of power of the W test and the z_{W_R} test for almost all alternative distributions is impressive which shows that there is no loss of power in extending the W test using the procedure of Royston. The RG has often simular results, it is even much better for the beta distributed alternatives. Nevertheless, the z_{W_R} seems to be the better chioce because of the fact that its power is much better for symmetric long-tailed distributions than the power of the RG test and because for the symmetric short-tailed alternatives where the RG test acieves the best results, the power of the z_{W_R} test was only slightly smaller.

Table 4.10: Empirical power for asymmetric alternative distributions for $n = 10, 15, 20, 25$.

Alternative	$\sqrt{\beta_1}$	β_2	$n=10$								$n=15$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(2,1)$	-0.57	2.4	13.2	10.2	10.4	10.0	15.9	13.1	6.7	5.6	20.9	14.6	14.9	14.0	26.9	20.7	7.1	5.5
$Beta(3,2)$	-0.29	2.36	5.2	4.1	4.2	4.0	6.7	5.2	3.2	3.0	3.3	4.0	4.1	3.8	9.0	5.9	2.5	2.2
χ^2_1	2.83	15	73.5	69.1	69.4	68.6	74.5	73.4	52.1	31.5	92.6	89.6	89.7	89.1	93.4	92.5	71.5	56.5
χ^2_2	1.41	9	44.5	41.6	41.9	41.3	44.0	44.4	33.7	17.2	68.0	63.2	63.5	62.6	68.2	67.7	49.5	33.8
χ^2_4	0.71	6	24.0	23.0	23.2	22.9	22.8	23.9	20.8	18.7	39.0	36.5	36.6	36.1	37.2	38.7	31.2	27.9
χ^2_{10}	0.28	4.2	12.0	12.0	12.0	12.0	11.1	12.0	11.7	11.1	18.1	17.6	17.7	17.4	16.2	17.9	16.8	15.4
χ^2_{20}	0.14	3.6	8.4	8.5	8.5	8.4	7.8	8.4	8.5	8.2	11.3	11.2	11.3	11.2	10.2	11.2	11.1	10.5
$\exp_1$	2	9	44.7	41.8	42.0	41.4	44.1	44.5	33.8	29.4	67.9	63.2	63.4	62.5	68.2	67.7	49.5	43.8
$\exp_{0.5}$	2	9	44.6	41.7	41.9	41.3	44.1	44.5	33.6	29.3	68.0	63.3	63.5	62.6	68.2	67.7	49.6	43.8
$LogN(0,1)$	6.18	113.9	60.8	58.6	58.8	58.3	59.7	60.7	50.2	44.8	82.6	79.9	80.0	79.5	82.1	82.4	69.6	64.3
$LogN(0,0.5)$	1.75	8.9	24.7	24.4	24.5	24.3	22.7	24.7	22.9	21.0	39.1	37.9	38.0	37.6	36.1	38.9	34.5	31.5
$Wei(0.5,1)$	6.62	87.72	44.6	41.7	41.9	41.3	44.1	44.4	33.7	29.3	68.0	63.3	63.6	62.7	68.3	67.8	49.6	43.8
$Wei(2,1)$	0.63	3.25	44.6	41.7	41.9	41.3	44.1	44.5	33.7	29.4	68.1	63.3	63.5	62.6	68.2	67.8	49.6	43.9
$Wei(10,1)$	-0.64	3.57	44.5	41.6	41.8	41.2	44.0	44.4	33.6	29.3	68.0	63.3	63.6	62.7	68.2	67.8	49.6	43.9
P_4	0.5	3.25	14.4	12.4	12.7	12.1	14.7	14.4	6.7	6.5	18.4	16.0	16.2	15.8	19.2	18.2	8.3	7.9
P_{10}	0.32	3.1	8.1	7.5	7.5	7.4	8.4	8.1	5.6	5.6	9.2	8.5	8.5	8.4	9.4	9.0	6.3	6.1

Alternative	$\sqrt{\beta_1}$	β_2	$n=20$								$n=25$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(2,1)$	-0.57	2.4	30.3	20.5	20.7	19.3	39.5	30.4	7.6	5.4	41.0	27.7	27.9	26.1	52.4	41.1	8.3	5.6
$Beta(3,2)$	-0.29	2.36	7.2	4.2	4.3	4.0	11.8	7.3	2.1	1.7	8.7	4.7	4.8	4.4	14.8	8.8	1.9	1.5
χ^2_1	2.83	15	98.4	97.2	97.2	97.0	98.7	98.4	84.4	74.1	99.7	99.4	99.4	99.3	99.8	99.7	92.2	86.0
χ^2_2	1.41	9	83.6	79.1	79.2	78.3	84.1	83.6	62.9	49.2	92.4	89.0	89.0	88.4	92.8	92.4	73.7	62.7
χ^2_4	0.71	6	53.0	49.1	49.2	48.4	51.2	53.0	40.8	36.4	65.2	60.3	60.5	59.7	63.7	65.2	49.6	44.5
χ^2_{10}	0.28	4.2	24.3	23.1	23.2	22.9	21.6	24.3	21.5	19.5	30.6	28.7	28.7	28.4	27.2	30.6	26.1	23.6
χ^2_{20}	0.14	3.6	14.4	14.0	14.1	13.9	12.5	14.4	13.7	12.7	17.6	16.8	16.9	16.8	15.1	17.6	16.3	15.0
$\exp_1$	2	9	83.5	79.0	79.2	78.3	84.0	83.5	62.9	56.3	92.4	89.0	89.0	88.4	92.8	92.4	73.7	67.2
$\exp_{0.5}$	2	9	83.5	79.0	79.1	78.2	84.1	83.5	62.9	56.2	92.3	88.9	89.0	88.4	92.8	92.3	73.6	67.1
$LogN(0,1)$	6.18	113.9	93.2	91.3	91.4	91.0	93.0	93.2	82.5	77.7	97.5	96.5	96.5	96.3	97.5	97.6	90.4	86.8
$LogN(0,0.5)$	1.75	8.9	52.1	49.9	50.0	49.5	48.6	52.1	44.7	40.8	63.1	60.3	60.4	59.9	59.5	63.2	53.7	49.5
$Wei(0.5,1)$	6.62	87.72	83.5	79.0	79.2	78.2	84.1	83.5	62.8	56.2	92.3	88.9	89.0	88.3	92.8	92.4	73.6	67.1
$Wei(2,1)$	0.63	3.25	83.6	79.0	79.2	78.3	84.1	83.6	62.9	56.3	92.3	88.9	89.0	88.4	92.8	92.4	73.5	67.0
$Wei(10,1)$	-0.64	3.57	83.5	78.9	79.1	78.2	84.0	83.5	62.9	56.2	92.4	88.9	89.0	88.4	92.8	92.4	73.6	67.1
P_4	0.5	3.25	23.6	20.2	20.4	19.7	23.5	23.6	9.8	9.1	29.7	25.0	25.2	24.5	28.9	29.8	11.3	10.5
P_{10}	0.32	3.1	10.6	9.6	9.7	9.5	10.6	10.6	6.9	6.6	12.1	10.8	10.9	10.8	11.9	12.1	7.5	7.2

Table 4.11: Empirical power for asymmetric alternative distributions for $n = 30, 35, 50, 100$.

Alternative	$\sqrt{\beta_1}$	β_2	$n = 30$								$n = 35$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(2,1)$	-0.57	2.4	51.4	36.1	36.0	33.6	64.0	51.6	9.5	6.2	61.4	45.0	44.6	41.7	74.2	61.6	11.0	6.9
$Beta(3,2)$	-0.29	2.36	10.5	5.6	5.5	5.0	18.4	10.6	1.8	1.3	12.5	6.5	6.4	5.7	22.4	12.6	1.7	1.2
χ^2_1	2.83	15	99.9	99.9	99.9	99.9	100	100	96.5	93.1	100	100	100	100	100	100	98.6	96.7
χ^2_2	1.41	9	96.7	94.7	94.7	94.2	97.0	96.8	82.1	73.9	98.7	97.6	97.6	97.3	98.9	98.7	88.3	83.5
χ^2_4	0.71	6	74.8	70.1	70.1	69.1	73.8	74.9	57.8	52.5	82.4	78.2	78.0	77.1	81.9	82.5	65.2	59.7
χ^2_{10}	0.28	4.2	36.7	34.4	34.3	33.8	32.8	36.8	30.6	28.0	42.5	39.8	39.7	39.0	38.6	42.7	35.0	32.1
χ^2_{20}	0.14	3.6	20.6	19.8	19.8	19.6	17.5	20.8	18.8	17.5	23.8	22.7	22.6	22.2	20.4	23.9	21.2	19.7
$\exp_1$	2	9	96.7	94.7	94.7	94.2	97.1	96.8	82.0	76.3	98.7	97.7	97.6	97.3	98.9	98.7	88.3	83.4
$\exp_{0.5}$	2	9	96.7	94.7	94.7	94.2	97.0	96.7	82.1	76.3	98.7	97.6	97.6	97.3	98.9	98.7	88.3	83.4
$LogN(0,1)$	6.18	113.9	99.2	98.7	98.7	98.6	99.2	99.2	95.0	92.7	99.7	99.6	99.6	99.5	99.8	99.8	97.6	96.1
$LogN(0,0.5)$	1.75	8.9	72.1	69.3	69.3	68.7	68.9	72.3	61.9	57.7	79.4	76.7	76.5	75.9	76.7	79.5	68.9	64.8
$Wei(0.5,1)$	6.62	87.72	96.7	94.7	94.7	94.2	97.0	96.7	82.0	76.3	98.7	97.6	97.6	97.3	98.9	98.7	88.3	83.4
$Wei(2,1)$	0.63	3.25	96.7	94.7	94.7	94.2	97.0	96.7	82.0	76.3	98.7	97.6	97.6	97.3	98.9	98.7	88.2	83.4
$Wei(10,1)$	-0.64	3.57	96.7	94.7	94.7	94.2	97.0	96.7	82.0	76.3	98.7	97.6	97.6	97.3	98.9	98.7	88.3	83.4
P_4	0.5	3.25	36.2	30.8	30.7	29.7	34.6	36.4	12.8	11.9	43.5	37.1	36.8	35.5	41.2	43.7	14.3	13.2
P_{10}	0.32	3.1	13.7	12.4	12.3	12.1	13.2	13.8	8.2	7.9	15.4	13.9	13.8	13.5	14.8	15.5	8.8	8.4

Alternative	$\sqrt{\beta_1}$	β_2	$n = 50$								$n = 100$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(2,1)$	-0.57	2.4	84.3	70.0	69.5	66.3	92.3	84.1	20.0	11.4	—	99.2	99.1	98.8	100	99.9	84.8	71.6
$Beta(3,2)$	-0.29	2.36	20.3	10.3	10.0	8.8	35.1	20.0	1.8	1.1	—	33.8	32.6	29.0	75.1	53.0	9.7	4.4
χ^2_1	2.83	15	100	100	100	100	100	100	100	99.8	—	100	100	100	100	100	100	100
χ^2_2	1.41	9	99.9	99.9	99.8	99.8	100	99.9	97.7	95.8	—	100	100	100	100	100	100	100
χ^2_4	0.71	6	95.0	92.5	92.4	91.8	94.8	94.9	82.5	77.8	—	99.9	99.9	99.9	100	100	99.5	99.1
χ^2_{10}	0.28	4.2	59.1	55.1	54.9	54.1	54.5	58.9	48.0	44.2	—	87.5	87.2	86.5	88.0	90.1	81.6	78.1
χ^2_{20}	0.14	3.6	33.3	31.2	31.1	30.6	28.2	33.1	28.6	26.6	—	57.5	57.2	56.3	54.3	60.8	52.9	49.6
$\exp_1$	2	9	99.9	99.8	99.8	99.8	100	99.9	97.6	95.7	—	100	100	100	100	100	100	100
$\exp_{0.5}$	2	9	99.9	99.9	99.8	99.8	100	99.9	97.7	95.7	—	100	100	100	100	100	100	100
$LogN(0,1)$	6.18	113.9	100	100	100	100	100	100	99.8	99.6	—	100	100	100	100	100	100	100
$LogN(0,0.5)$	1.75	8.9	92.5	90.5	90.5	90.0	90.9	92.4	84.5	81.1	—	99.8	99.8	99.7	99.8	99.9	99.3	99.0
$Wei(0.5,1)$	6.62	87.72	99.9	99.9	99.8	99.8	100	99.9	97.6	95.7	—	100	100	100	100	100	100	100
$Wei(2,1)$	0.63	3.25	99.9	99.8	99.8	99.8	100	99.9	97.6	95.7	—	100	100	100	100	100	100	100
$Wei(10,1)$	-0.64	3.57	99.9	99.8	99.8	99.8	100	99.9	97.6	95.7	—	100	100	100	100	100	100	100
P_4	0.5	3.25	66.4	57.8	57.5	55.7	59.7	66.1	18.8	17.4	—	99.5	99.5	99.2	99.0	100	35.2	32.4
P_{10}	0.32	3.1	21.4	18.7	18.6	18.2	19.6	21.2	10.6	10.0	—	40.5	40.1	38.7	39.3	45.7	17.1	16.0

Table 4.12: Empirical power for asymmetric alternative distributions for $n = 200, 500$.

Alternative	$\sqrt{\beta_1}$	β_2	$n=200$								$n=500$							
			W	SF	WB	FB	RG	z_{W_R}	JB	AJB	W	SF	WB	FB	RG	z_{W_R}	JB	AJB
$Beta(2,1)$	-0.57	2.4	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$Beta(3,2)$	-0.29	2.36	—	85.5	84.1	81.0	99.1	94.7	70.5	60.7	—	—	100	100	100	100	100	100
χ^2_1	2.83	15	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
χ^2_2	1.41	9	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
χ^2_3	0.94	7	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
χ^2_4	0.71	6	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
χ^2_5	0.57	5.4	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
χ^2_8	0.35	4.5	—	100	100	99.4	100	100	99.9	99.8	—	—	100	100	100	100	100	100
χ^2_{10}	0.28	4.2	—	99.7	99.6	99.6	99.7	99.8	99.3	99.0	—	—	100	100	100	100	100	100
χ^2_{20}	0.14	3.6	—	88.6	88.3	87.7	87.0	90.7	86.2	84.4	—	—	100	100	99.9	100	100	99.9
$\exp_1$	2	9	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$\exp_{0.5}$	2	9	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$LogN(0,1)$	6.18	113.9	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$LogN(0,0.5)$	1.75	8.9	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$Wei(0.5,1)$	6.62	87.72	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$Wei(2,1)$	0.63	3.25	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$Wei(3,1)$	0.17	2.73	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
$Wei(10,1)$	-0.64	3.57	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
P_1	1	4	—	100	100	100	100	100	100	100	—	—	100	100	100	100	100	100
P_4	0.5	3.25	—	100	100	100	100	100	68.4	65.3	—	—	100	100	100	100	99.4	99.2
P_{10}	0.32	3.1	—	88.0	87.4	86.0	79.6	91.9	31.2	30.0	—	—	100	100	100	100	71.4	70.1

Even though the JB test is very often used in practice, the power results are sometimes surprisingly bad which makes this test acutally not a good chioce in the everyday work of a statistician. For symmetric long-tailed distributions and for the uniform distribution the Jarque-Bera type tests JB and AJB achieved the best power. But for all other types of alternative distributions their power behaviour was really inferior compared to the regression type tests. An often cited argument for the JB test is its easy computably test statistic. However, there is still the need for critical significane points due to the slow asymptotic convergence as we have seen in section 3.4 so that this argument seems to be incomprehensible. Additionally, all these procedures can today easily implemented in a statistic software so that the argument of easy computation cannot hold any more.

For these reasons we recommend the use of the z_{W_R} test as a test for normality since its overall power behaviour is better than for any other considered test in our study and because it can be used even for large sample sizes. Besides, the z_{W_R} test is also implemented in almost all statistical software packages like R, SPSS, SAS or STATISTICA. However, the results of our empirical power studies underline again what has been stated many times in preceding studies with the same subject: it seems that no omnibus test for detecting departures from the normal distribution appears to exist.

Chapter 5

Conclusion

Besides the use of formal tests for normality—no matter which test is taken—there is one thing that we have not mentioned yet. Testing for normality also always includes a detailed graphical analysis involving for example a normal probability plot or a boxplot. Making a decision wether a sample is normally distributed or not without looking at a graphic makes the investigation not complete. A famous and often cited quote of John W. Tukey (taken from Thode (2002, p. 15)) brings it in one nutshell:

> *There is no excuse for failing to plot and look.*

As a pracitioneer one should always keep these words in mind because not unfrequently the results of the formal tests contradict with the impressions of the graphical analysis. In particular for large sample sizes the blind trust on the formal tests for normality can lead to erroneous results.

In our work, in particular in the power study, it was of course impossible to consider all known tests for normality, we had to limit our investigations to some very famous ones and we could also only examine regression type test and test based on the third and fourth sample moment. Anyway, by studing the broad literature concerning the theory of testing for normality one finds that most of the tests were develped in the 1960s, 1970s and 1980s. After the development of the Jarque-Bera test in the late 1980s, only few new tests seemed to appear after that time (cf. Thode (2002) for an overview). Moreover, none of these new tests could establish itself as real alternative for testing for normality with an essential improvement of empirical power behaviour. This leads us to the conclusion that the theory of testing for normality, in particular the search of tests for normality with better power behaviour than the already existing tests seems to be mostly closed. Thus, the validity of our power study should be rated as relatively acceptable since all of the most important todays known tests for normality were included.

The normal distribution is definitely one of the most important distributions in applied statistics but its importance is, according to the authors experiences, overestimated, at least in some areas of applied statistics. The reason for this opinion is that for almost every parametric statistical procedure that demands a fulfilled normality assumption, there is a nonparametric alternative that can also be conducted without the need of an underlying normal distribution. Additionally in another research area (robust statistics) it was found that the validity for many parametric tests is not or almost not affected when the sample is not normally distributed. The analysis of variance is such an example when the design is balanced, i.e., when for every factor level the sample sizes are equal. After the authors experiences, another fact that can often be regarded in applied science is some kind of unreflected use of the normal distribution as a common practice which is nicely summarized in the following citation of Daw (1966, p. 2):

> *[...] everybody believes in the Law of Errors (i.e. the Normal distribution), the experimenters because they think it is a mathematical theorem and the mathematicians because they think it is an experimental fact.*

Many researchers forget that they should not check wether the data *comes* from a normal distribution but should rather realize that testing for normality means checking wether the data *could be assessed* by a model that has as its underlying distribution the normal distribution.

To give a final statement, in conjunction with the literature research for this work, the author found a very interesting citation of Geary (1947, p. 241) which became to a favourite statement during the completion of this diploma thesis. It puts all the efforts in developing new tests and improving established tests for normality in a nutshell:

> *Normality is a myth; there never was, and never will be a normal distribution. This is an overstatement for the practical point of view, but it represents a safer initial mental attitude than any in fashion in the past two decades.*

Appendix A

General Definitions

A.1 Definition. (Joint probability density function) *Let* $X_1, \ldots, X_n$ *be* n *random variables in the probabilty space* $(\Omega, \mathcal{A}, P)$ *and* f *a random variable such that* $\int_\Omega f \, \mathrm{d}P = 1$. *If for* $i = 1, \ldots, n$,

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{A_n} \cdots \int_{A_1} f(x_1, \ldots, x_n) \, \mathrm{d}x_1 \cdots \, \mathrm{d}x_n, \quad A_i \in \mathcal{A},$$

then $f(x_1, \ldots, x_n)$ *is called* **joint probability density function (pdf)** *of* $X_1, \ldots, X_n$. $\diamondsuit$

A.2 Definition. (Joint cumulative densitiy function) *Let* $X_1, \ldots, X_n$ *be* n *random variables in the probabilty space* $(\Omega, \mathcal{A}, P)$ *and* $F : \mathbb{R}^n \to [0, 1]$. *If*

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

then $F(x_1, \ldots, x_n)$ *is called* **joint cumulative density function (cdf)** *of* $X_1, \ldots, X_n$. $\diamondsuit$

Sufficency and Completenes

A.3 Definition. (Sufficient statistic) *Let* $y_1, \ldots, y_n$ *denote a sample of iid random observations coming from a random variable* Y *with pdf* $f(Y; \vartheta); \vartheta \in \theta$. *Let* $Z = u(y_1, \ldots, y_n)$ *be a statistic whose pdf is* $g(Z; \vartheta)$. *Then* Z *is called a* **sufficient statistic for** $\vartheta \in \theta$ *if and only if*

$$\frac{f(y_1; \vartheta) \cdots f(y_n; \vartheta)}{g\left(u(y_1, \ldots, y_n); \vartheta\right)} = H(y_1, \ldots, y_n),$$

where $H(y_1, \ldots, y_n)$ *is a function that does not depend upon* $\vartheta \in \theta$ *for every fixed value of* Z. $\diamondsuit$

A.4 Definition. (Complete statistic) *Let the random variable Y have a pdf that is a member of the family $\{f(y;\vartheta) : \vartheta \in \theta\}$ and let $u(Y)$ be a statistic. If the condition $E(u(Y)) = 0$ for every $\vartheta \in \theta$ requires that $u(Y)$ be zero execept on a set of points that has probability zero for each pdf $f(y;\vartheta), \vartheta \in \theta$, then the family $\{f(Y;\vartheta) : \vartheta \in \theta\}$ is called a* **complete family** *of pdfs.* ◊

A.5 Definition. (Complete sufficient statistic) *Let T be a sufficient statistic for a parameter $\vartheta \in \theta$ and let the family $\{f(T;\vartheta) : \vartheta \in \theta\}$ of pdfs of T be complete. Then T is also called a* **complete sufficient statistic**. ◊

A.6 Theorem. *Let $X_1, \ldots, X_n$ denote a random sample from a distribution having a pdf $f(x;\vartheta), \vartheta \in \theta$, where θ is an interval set. Let $T = t(X_1, \ldots, X_n)$ be a complete sufficient statistic for $\vartheta \in \theta$. If the distribution of any other statistic $U = u(X_1, \ldots, X_n)$ does not depend upon ϑ, then U is stochastically independet of the sufficient statistic T.*

Proof: See Hogg and Craig (1978, p. 389-390). □

Large-Sample Theory

A.7 Definition. (Convergence in Distribution) *A sequence of random vectors $\{X_n\}$ with cdf $\{F_{X_n}(\cdot)\}$ is said to converge in distribution to a random vector X with cdf $F_X(\cdot)$ if*

$$F_{X_n}(x_1, \ldots, x_n) \to F_X(x_1, \ldots, x_n) \qquad \text{as } n \to \infty,$$

for every $x = (x_1, \ldots, x_n)$ at which $F_X(x_1, \ldots, x_n) = P\{X_1 \le x_1, \ldots, X_n \le x_n\}$ is continous. Convergence in distribution is denoted by $X_n \xrightarrow{\mathcal{D}} X$. Alternative names for this convergence are weak convergence or convergence in law. ◊

A.8 Definition. (Convergence in Probability) *A sequence of random vectors $\{X_n\}$ converges in probability to a random vector X, written $X_n \xrightarrow{P} X$, if, for all $\varepsilon > 0$,*

$$P\{|X_n - X| > \varepsilon\} \to 0 \qquad \text{as } n \to \infty.$$

◊

A.9 Theorem. (Multivariate Central Limit Theorem) *Let* $X_n' = (X_{n,1}, \ldots, X_{n,k})$ *be a sequence of iid random vectors with mean vector* $\mu' = (\mu_1, \ldots, \mu_k)$ *and covariance matrix* $\Sigma = E((X_1 - \mu)(X_1 - \mu)')$. *Let* $\bar{X}_{n,j} = n^{-1} \sum_{i=1}^{n} X_{i,j}, j = 1, \ldots, k$. *Then*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) = \sqrt{n}((\bar{X}_{n,1} - \mu_1), \ldots, (\bar{X}_{n,k} - \mu_k))' \xrightarrow{\mathcal{D}} N(0, \Sigma) \qquad \text{as } n \to \infty.$$

Proof: See Lehmann (1999, p. 313). □

A.10 Theorem. (Delta Method) *Suppose* $X_1, X_2, \ldots$ *and* X *are random vectors in* $\mathbb{R}^k$. *Assume* $\tau_n(X_n - \varphi) \xrightarrow{\mathcal{D}} X$, *where* φ *is a constant vector and* $\{\tau_n\}$ *is a sequence of constants* $\tau_n \to \infty$. *Suppose* g *is a function from* $\mathbb{R}^k$ *to* $\mathbb{R}$ *which is differentiable at* φ *with gradient (vector of first partial derivates) of dimension* k *at* φ *equal to* $\nabla g(\varphi)$. *Then,*

$$\tau_n(g(X_n) - g(\varphi)) \xrightarrow{\mathcal{D}} (\nabla g(\varphi))' X \qquad \text{as } n \to \infty.$$

In particular, if $\sqrt{n}(X_n - \varphi)$ *converges in law to a multivariate normal distribution* $N(0, \Sigma)$ *the conclusion of the theorem is that the sequence* $\sqrt{n}(g(X_n) - g(\varphi))$ *converges in law to the* $N(0, (\nabla g(\varphi))' \, \Sigma \, \nabla g(\varphi))$ *distribution.*

Proof: See Lehmann (1999, p. 86). □

Bibliography

F. J. Anscombe and W. J. Glynn. Distribution of the kurtosis ktatistic b_2 for normal statistics. *Biometrika*, 70:227–234, 1983.

N. Balakrishnan and A. C. Cohen. *Order Statistics and Inference*. Academic Press, Boston, 1991.

G. Blom. *Statistical Estimates and Transformed Beta-Variables*. Johan Wiley & Sons, New York, 1958.

D. G. Bonett and E. Seier. A test of normality with high uniform power. *Computational Statistics & Data Analysis*, 40:435–445, 2002.

K. O. Bowman and L. R. Shenton. Omnibus Test Contours for Departures from Normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, 62:243–250, 1975.

R. B. D'Agostino. Transformation to Normality to the Null Distribution of g_1. *Biometrika*, 57: 679–681, 1970.

R. B. D'Agostino. Departures from Normality, Testing for. In North Holland, editor, *Encyclopedia of Statistical Sciences*, volume 2. S. Kotz and N. L. Johnson and C. B. Read, Amsterdam, 1982.

R. B. D'Agostino. Tests for the Normal Distribution. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness-of-fit Techniques*, chapter 9. Marcel Dekker, Inc., New York, 1986a.

R. B. D'Agostino. Graphical Analysis. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness-of-fit Techniques*, chapter 2. Marcel Dekker, Inc., New York, 1986b.

R. B. D'Agostino and E. S. Pearson. Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika*, 60:613–622, 1973.

R. B. D'Agostino and G. L. Tietjen. Simulation probability points of b_2 for small samples. *Biometrika*, 58:669–672, 1971.

R. B. D'Agostino and G. L. Tietjen. Approaches to the Null Distribution of $\sqrt{b_1}$. *Biometrika*, 60:169–173, 1973.

C. S. Davies and M. A. Stephens. Approximating the Covariance Matrix of Normal Order Statistics. *Applied Statistics*, 27:206–212, 1978.

R. H. Daw. Why the normal distribution. *Journal of Statistical Simulation*, 18:2–15, 1966.

P. Deb and M. Sefton. The distribution of a Lagrange multiplier test of normality. *Econometrics Letters*, 51:123–130, 1996.

M. Falk, F. Marohn, and B. Tewes. *Foundations of Statistical Analysis and Applications with SAS*. Birkhäuser, Basel, 2002.

P. J. Farrell and K. Rogers-Stewart. Comprehensive Study of Tests for Normality and Symmetry: Extending the Spiegelhalter Test. *Journal of Statistical Computation and Simulation*, 76:803–816, 2006.

J. J. Filliben. The Probability Plot Correlation Coefficient Test for Normality. *Technometrics*, 17:111–117, 1975.

R. C. Geary. Testing for Normality. *Biometrika*, 34:209–242, 1947.

Y. R. Gel and J. L. Gastwirth. A robust modification of the Jarque-Bera test of normality. *Economics Letters*, 99:30–32, 2008.

H. J. Godwin. Some Low Moments of Order Statistics. *The Annals of Mathematical Statistics*, 20:279–285, 1948.

J. M. Hammersley and K. W. Morton. The estimation of location and scale parameters from group data. *Biometrika*, 41:296–301, 1954.

H. L. Harter. Expected Values of Normal Order Statistics. *Biometrika*, 48:151–165, 1961.

H. L. Harter and N. Balakrishnan. *Tables for the Use of Order Statistics in Estimation*. CRC Press, Boca Raton, 1996.

R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. Macmillan, New York, 1978.

S. M. A. Jahanshahi, H. Naderi, and R. Moayeripour. Jarque Bera Normality Test and Its Modification. In *Iranian Statistical Conference*, pages 38–55, 2008.

C. M. Jarque and A. K. Bera. Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals. *Economics Letters*, 6:255–259, 1980.

C. M. Jarque and A. K. Bera. A Test for Normality of Observations and Regression Residuals. *International Statistical Review*, 55:163–172, 1987.

N. L. Johnson and S. Kotz. *Discrete distributions*. John Wiley & Sons, New York, 1969.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions 2*. Mifflin, Boston, 1970.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions 1*. Mifflin, Boston, 1994.

S. Keskin. Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation based Small Samples. *Journal of Applied Science Research*, 2:296–300, 2006.

J. H. Kwak and S. Hong. *Linear Algebra*. Birkhäuser, Boston, 1997.

E. L. Lehmann. *Elements of Large Sample Theory*. Springer, New York, 1999.

D. S. Moore. Tests of Chi-Squared Type. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness-of-fit Techniques*, chapter 3. Marcel Dekker, Inc., New York, 1986.

K. J. Ord. *Families of Frequency Distributions*. Number 30 in Griffin's statistical monographs & courses. Griffin, London, 1972.

R. S. Parrish. Computing Expected Values of Normal Order Statistics. *Communications in Statistics – Simulation and Computation*, 21:57–70, 1992a.

R. S. Parrish. Computing Variances and Covariances of Normal Order Statistics. *Communications in Statistics – Simulation and Computation*, 21:71–101, 1992b.

R. S. Parrish. New tables of coefficients and percentage points for the w test for normality. *Journal of Statistical Computation and Simulation*, 41:169–185, 1992c.

J. K. Patel and C. B. Read. *Handbook of the Normal Distribution*. Dekker, New York, 1996.

E. S. Pearson. A Further Development of Tests for Normality. *Biometrika*, 22:239–249, 1930a.

E. S. Pearson. Note on Tests for Normality. *Biometrika*, 22:423–424, 1930b.

E. S. Pearson, R. B. D'Agostino, and K. O. Bowman. Tests for Departure from Normality: Comparison of Powers. *Biometrika*, 64:231–246, 1977.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

M. M. Rahman and Z. Govindarajulu. A Modification of the Test of Shapiro and Wilk for Normality. *Journal of Applied Statistics*, 24:219–235, 1997.

J. P. Royston. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Applied Statistics*, 31:115–124, 1982a.

J. P. Royston. Algorithm AS 177. Expected Normal Order Statistics (exact and approximate). *Applied Statistics*, 31:161–165, 1982b.

J. P. Royston. Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, 2:117–119, 1992.

A. E. Sarhan and B. G. Greenberg. Estimation of Location and Scale Parameters by Order Statistics from Singly and Doubly Censored Samples Part I. *Annals of Mathematical Statistics*, 27:427–451, 1956.

S. K. Sarkara and W. Wang. Estimation of Scale Parameters Based on a Fixed Set of Order Statistics. In N. Balakrishnan and C. R. Rao, editors, *Handbook of Statistics*, volume 17, chapter 6. Elsevier, Amsterdam, 1998.

SAS Institute Inc. *SAS Statistical Procedures, Version 9.* Cary, NC, 2003.

E. Seier. Comparison of Tests for Univariate Normality. *Interstat*, 1:1–17, 2002.

S. S. Shapiro. Distribution Assesement. In N. Balakrishnan and C. R. Rao, editors, *Handbook of Statistics*, volume 17, chapter 17. Elsevier, Amsterdam, 1998.

S. S. Shapiro and R. S. Francia. An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, 67:215–216, 1972.

S. S. Shapiro and M.B. Wilk. An analysis of variance Test for Normality (complete samples). *Biometrika*, 52:591–611, 1965.

S. S. Shapiro, M. B. Wilk, and H. J. Chen. A Comparative Study of Various Tests for Normality. *Journal of the American Statistical Association*, 63:1343–1372, 1968.

B. L. Shea and A. J. Scallan. AS R72. A Remark on Algorithm AS128: Approximating the Covariance Matrix of Normal Order Statistics. *Applied Statistics*, 37:151–155, 1988.

M. A. Stephens. Tests Based on EDF Statistics. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness-of-fit Techniques*, chapter 4. Marcel Dekker, Inc., New York, 1986.

A. Stuart and J. K. Ord. *Kendall's Advanced Theory of Statistics*, volume 1. Oxford University Press, London, 1987.

T. Thadewald and H. Büning. Jarque-Bera Test and its Competitors for Testing Normality – A Power Comparison. *Journal of Applied Statistics*, 37:87–105, 2007.

H. C. Thode. *Testing for Normality.* Marcel Dekker, New York, 2002.

C. M. Urzua. On the correct use of omnibus tests for normality. *Econometrics Letters*, 53: 247–251, 1996.

S. Weisberg and C. Bingham. An Approximate Analysis of Variance Test for Non-Normality Suitable for Machine Calculation. *Technometrics*, 17:133–134, 1975.

P. H. Westfall. *Multiple Comparisons and Multiple Tests*. SAS Institute Inc., Cary, NC, 1999.

H. Witting. *Mathematische Statistik I*. Teubner, Stuttgart, 1985.

B. Yazici and S Yolacan. A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77:175–183, 2007.

P. Zhang. Omnuibus Test for Normality Using the Q Statistic. *Journal of Applied Statistics*, 26:519–528, 1999.

Erklärung

Ich versichere, dass ich, Johannes Hain, die vorliegende Diplomarbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur angefertigt habe.

Würzburg, den 04. August 2010