A STUDY OF NORMALITY TESTS AND AN EXTENSION TO THE NORMALITY PLOT

By

Felipe H. Acosta Archila

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

MATHEMATICS STATISTICS

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS

May, 2010

Approved by:

Julio C. Quintana Díaz, Ph.D. Member, Graduate Committee	Date
Dámaris Santana Morant, Ph.D. Member, Graduate Committee	Date
Wolfgang Rolke, Ph.D. President, Graduate Committee	Date
Isidoro Couvertier, Ph.D. Representative of Graduate Studies	Date
Silvestre Colón, M.S. Chairperson of the Department	Date

Abstract of Dissertation Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

A STUDY OF NORMALITY TESTS AND AN EXTENSION TO THE NORMALITY PLOT

By

Felipe H. Acosta Archila

May 2010

Chair: Wolfgang Rolke
Major Department: Mathematics

We are considering a number of existing normality tests and study their power against different alternative hypotheses. We also develop an extension to the normality plot by adding a fixed confidence band. This procedure is named the envelope test. We use a Monte Carlo simulation to do a power study among all normal tests considered and in the development of the envelope test. The results show that although there was no best normality test, we can provide guidelines to improve their efficiency. With the envelope test we construct a method that eliminates the subjectivity that the normality plot carries within.

Resumen de Disertación Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

UN ESTUDIO DE PRUEBAS DE NORMALIDAD Y UNA EXTENSIÓN AL GRÁFICO DE NORMALIDAD

Por

Felipe H. Acosta Archila

Mayo 2010

Consejero: Wolfgang Rolke
Departamento: Matemática

Consideramos un número de pruebas de normalidad existentes y estudiamos su potencia utilizando diversas distribuciones como hipótesis alternativas. También desarrollamos una extensión a la gráfica de normalidad añadiendo una banda de confianza de cubrimiento establecido. Este procedimiento es llamado la prueba del sobre. Se usa una simulación de Monte Carlo para hacer un estudio de la potencia entre las pruebas consideradas y también para el desarrollo de la prueba del sobre. Los resultados muestran que aunque no existe una mejor prueba de normalidad, podemos tener guías para mejorar su eficiencia. Con la prueba del sobre construimos un método que elimina la subjetividad que conlleva el uso de gráficos de normalidad.

To my parents, Felipe and Eva, thank you for the support through the years.

ACKNOWLEDGMENTS

To my thesis advisor, Professor Wolfwang Rolke, who has guided me through my studies.

To Professor Paul Castillo who has encouraged me to keep going.

To the Mathematics Department's faculty and staff for granting me the opportunity to do my Master's studies.

To my family, my new friends that I have met during my stay in Puerto Rico, and my old friends from Honduras for their constant support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

i.i.d.	Independent and identically distributed
E.D.F.	Empirical Distribution Function
p.d.f.	Probability Density Function
C.D.F.	Cumulative Distribution Function
H_o	Null hypothesis
H_a	Alternative hypothesis
QQ	Quantile–Quantile

LIST OF SYMBOLS

μ	Mean of a random variable (location parameter).
$\hat{\mu}$	Sample mean. $\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i/n$
σ	Standard deviation of a random variable (scale parameter).
σ^2	Variance of a random variable.
$\hat{\sigma}$	Square root of the sample second moment. $\hat{\sigma} = \sqrt{m_2}$
s	Sample standard deviation. $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
μ_i	$i-$th centered moment. $\mu_i = E(x - \bar{x})^i$
m_i	Sample $i-$th centered moment. $m_i = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^i$
$x_{(i)}$	$i-$th order statistic.
$\hat{z}$	Standardized observation using estimated parameters.
$\hat{z}_{(i)}$	$i-$th ordered standardized observation using estimated parameters.
$F_n(x)$	Empirical Distribution Function.
$\Phi(x)$	Standard normal cumulative distribution function.

CHAPTER 1
INTRODUCTION: NORMALITY TESTS

Statistical models play a fundamental part in many areas of research, they help in issues such as quantification of uncertainty in data and calculations, characterization of numerical results of experiments and mathematical models for the better understanding of a system's nature, estimation and prediction of the behavior of a system, among other uses. Statistical models, like other mathematical models, carry conditions for their validity. The condition of normality is not an unusual one, commonly in regression models and in inference about a model's parameters. This issue of testing for normality has been studied for a long time, beginning in 1900 when Pearson [1] outlined his χ^2 test.

A normality test is a special type of a hypothesis test. Checking for normality is usually done in one of two ways, with graphics or with a formal hypothesis test. Graphical methods include histograms and normality plots; formal hypothesis tests consist of the calculation of a statistic from the data. Classical examples of hypothesis tests include the Pearson test and the Kolmogorov-Smirnov test.

A standard normality test has the following form. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a random variable X with an unspecified distribution. The null (H_o) and the alternative (H_a) hypothesis are defined by

$$H_o : X \sim N(\mu, \sigma)$$
$$H_a : X \nsim N(\mu, \sigma)$$

As with all hypothesis tests, we should be careful when dealing with normality tests, if the test rejects the null hypothesis all we can say is that it is unlikely that the sample was drawn from a normal population. On the other hand, if the test fails to reject the null hypothesis it tells us that there is not enough information to claim that it did not come

from a normal population. The test's conclusion refers to the population from were the sample was taken, it does not say if a sample is normal or not.

One question that may arise when dealing with normality tests is the meaning of the alternative hypothesis. Stating that a sample does not come from a normal distribution does not tell us much. If a specific alternative hypothesis is needed, is up to the researcher—using histograms, boxplots and other methods—to come up with an alternative distribution.

Since in practice we do not usually know the parameters μ and σ of a distribution, we will only discuss tests for composite normality. The tests to be considered also have the property of being invariant to location–scale transformations, which is important because it lets us focus on the shape of the distribution rather than also worry about these location–scale parameters.

Our interest is to compare various tests for normality, hence we need a measure for their efficiency. For this we use the *power* of the test, this is, the probability of rejecting H_o when it is not true. Since in the case of normality testing it is not possible to estimate a general power against every non–normal distribution, we will measure the power using different distributions and estimating the power for each alternative.

CHAPTER 2
GRAPHICAL NORMALITY TESTS

There are many graphical procedures that help in the inspection of data to detect deviations from normality—for example histograms and normality plots—the last one being the most widely used. A difficulty that these types of tests present is that they can be very subjective, especially if the investigator is not very experienced.

2.1 Histograms

A very practical, but often unreliable, test is to plot a histogram of the sample with its area scaled to 1 along with the p.d.f. of a normal distribution. Notice that if we don't know the parameters of the normal distribution we will need to estimate them.

Example. *Figure* ?? *is a histogram constructed from a sample of size 200 taken from a standard normal distribution with its p.d.f. We can see that even for a reasonable big sample the data and the p.d.f. does not fit perfectly.*

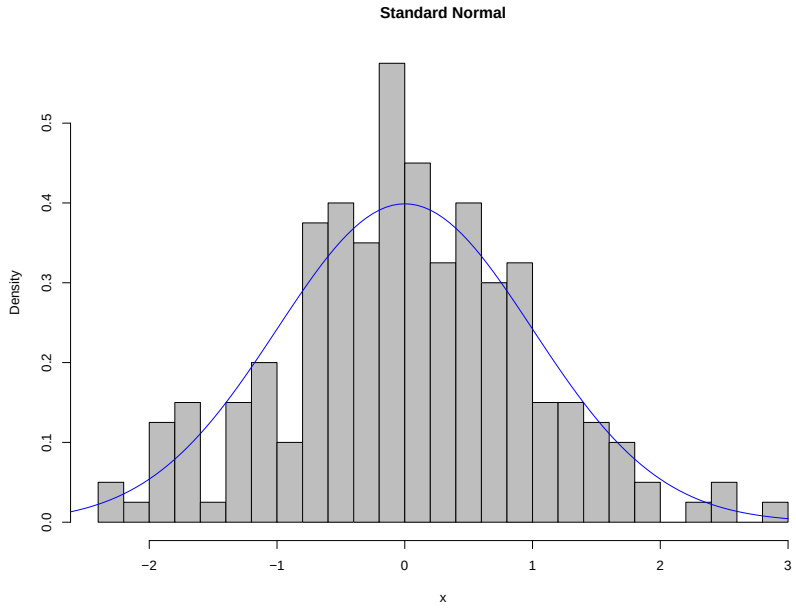

Figure 2–1: Histogram with p.d.f.

It is easy to see that for smaller sample sizes it will be harder to reach a decision. Because of this, histograms are rarely used as a conclusive test. Nevertheless they are useful to check for characteristics such as outliers, symmetry or peakedness of the sample.

A positive aspect of histograms is that it is easy to jump from testing for normality to testing for other distributions as long as the form of the p.d.f. is know.

2.2 Normality plots

A normality plot is a quantile–quantile (QQ) plot of the sample order statistics against the population quantiles $\Phi^{-1}(p_i)$ of a standard normal distribution. The values $\Phi^{-1}(p_i)$ can be calculated in different ways, for example two common positions are $p_i = (i - .5)/n$ and $p_i = i/(n + 1)$ [2].

Theoretically, under the null hypothesis we have that $X \sim N(\mu, \sigma)$ and we can define $Z = \dfrac{X - \mu}{\sigma}$ to be a standard normal variable. Therefore we can write

$$F_Z\left(\frac{x - \mu}{\sigma}\right) = F_X(x)$$

We are interested in the ordered pairs of quantiles $\left(F_Z^{-1}(p), F_X^{-1}(p)\right)$ for $0 < p < 1$. Let $p_0 \in (0, 1)$ fixed such that

$$F_Z\left(\frac{x - \mu}{\sigma}\right) = F_X(x) = p_0 \tag{2.1}$$

Since both F_Z and F_X are strictly increasing continuous functions their inverses are well defined. From 2.1 we find that

$$F_Z^{-1}(p_0) = \frac{x - \mu}{\sigma} \tag{2.2}$$

$$F_X^{-1}(p_0) = x \tag{2.3}$$

Substituting 2.3 in 2.2 and rearranging terms we can write

$$F_X^{-1}(p_0) = \mu + \sigma F_Z^{-1}(p_0) \tag{2.4}$$

Therefore from 2.4 we can see that the quantiles of X and Z have a linear relationship. This means that in QQ plots deviance from a straight line is evidence to reject the hypothesis of normality.

Example. *Figure ?? shows a normality plot for a sample of size 50 taken from a standard normal distribution.*

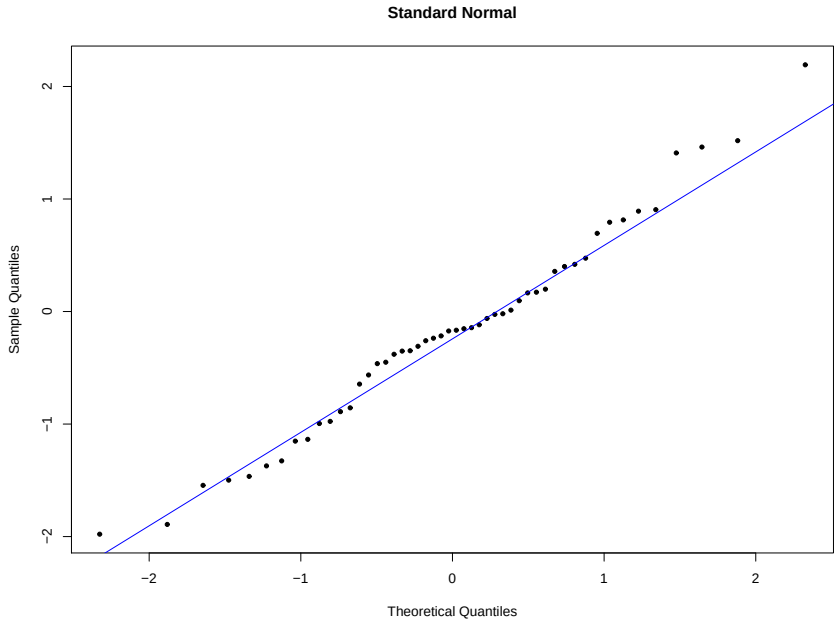

Figure 2–2: Normality plot

If the real distribution is not a normal distribution, there are some common deviations of the points from a straight line. Some well known deviations and their patterns are: a heavy–tailed distribution will bend the lower tail downward and the upper tail upward, a short-tailed distribution will bend the lower tail to the left and the upper tail to the right, creating a "S"-shape, an asymmetric distribution skewed to the left will be concave upward, and skewed to the right concave downward. Other deviations such as outliers will appear farther of the straight line, and bimodality as separated lines. [2]

2.2.1 Confidence bands

There are methods to add confidence bands to the normality plot to do a formal hypothesis test on top of a graphical procedure. Two of these methods are presented in the library *car* available in *R* and in the probability plot available in *Minitab*.

The first method relies on the calculation of confidence intervals for each of the theoretical quantiles in the QQ plot instead of calculating a confidence band for overall coverage. For the method available in *Minitab* we do not have the procedure of how the

confidence band is calculated. However in Chapter 4 we show that it does not have the total coverage confidence level.

Also in Chapter 4, we develop a method of calculation a confidence band based on a Monte Carlo simulation. We call this method the envelope test. With the envelope test we want to create a confidence band that has a total coverage confidence level. We compare the envelope test to the methods available in *R* and in *Minitab*.

CHAPTER 3
NUMERICAL NORMALITY TESTS

A numerical normality test consists in the calculation of a statistic from the sample, then this statistic is compared to a fixed critical value or a significance level is calculated to decide the outcome of the test. We will classify numerical tests in four categories: tests based on the empirical distribution function (E.D.F.); moments; correlation and entropy.

3.1 E.D.F. based tests

The E.D.F. of a sample, $F_n(x)$, is the proportion of observations less than or equal to x. It is defined as

$$F_n(X) = \begin{cases} 0 & x < x_{(1)} \\ i/n & x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, \ldots, n-1 \\ 1 & x \geq x_{(n)} \end{cases}$$

A typical example of a E.D.F. ploted with $\Phi(x)$ is shown in figure 3–1.

Tests based on the E.D.F of a sample were the first normality tests to be developed, starting with Pearson in 1900 [1] and Kolmogorov in 1933 [3]. These types of tests compare the E.D.F. of the sample to the C.D.F. of the null distribution, the problem with classic E.D.F. tests was that prior knowledge of the null distribution's parameters is required to perform the test. It was not until later that, with asymptotic theory and Monte Carlo studies, critical values for the composite version of the tests could be calculated. [4]

3.1.1 Pearson's χ^2 test

This test groups the observed data in k classes and calculates the difference between the observed frequency and the expected frequency of each cell.

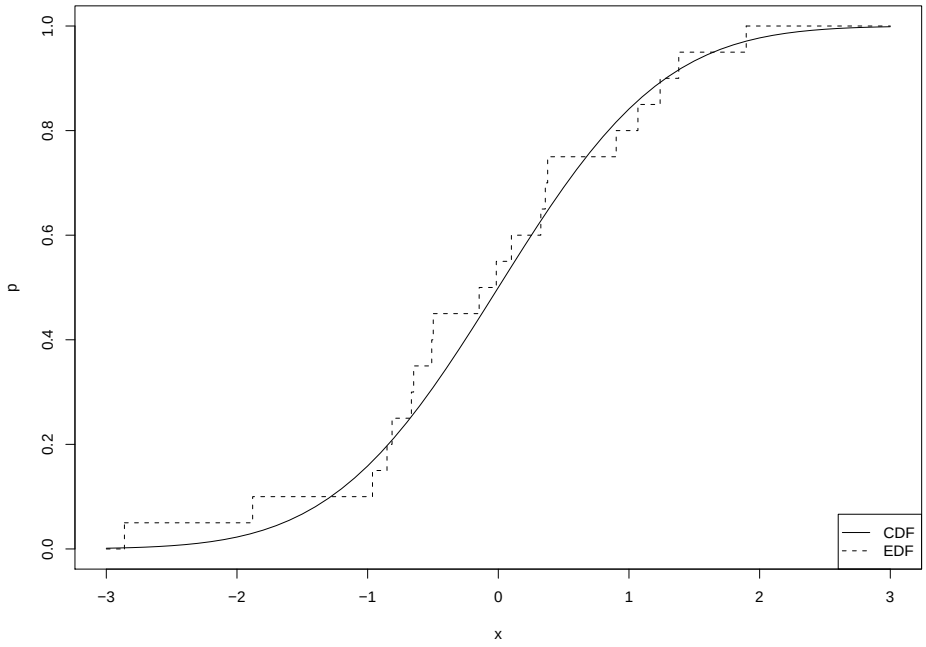

Figure 3–1: $\Phi(x)$ with the E.D.F. of a sample

The statistic for this test is calculated as:

$$P^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.1}$$

Where O_i is the observed frequency and E_i is the expected frequency of each class.

The asymptotic distribution of 3.1 is a χ^2_{k-1} if the parameters of the null distribution are know, in the case that the parameters are estimated it is a χ^2_{k-m-1}, where m is the number of estimated parameters, but if all sampled data was used to estimate the parameters using an efficient estimator— like the maximum likelihood estimators— there is a recovery of some of the m degrees of freedom resulting in an asymptotic distribution between χ^2_{k-1} and χ^2_{k-m-1} [2]

In the case of composite normality, the null distribution that is used is a standard normal distribution, and the sample is standardized using the estimated values of μ and σ, respectively $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = s$.

Because of this standardization and Proposition 1 in Appendix A, the statistic 3.1 is invariant to location–scale transformations

Choosing an adequate number of classes is a difficulty that this test presents, Moore [4] suggests

$$k \leq 4\left(\frac{2n^2}{c(\alpha)^2}\right)^{1/5} \tag{3.2}$$

where $c(\alpha)$ is the upper $\alpha-$point of a standard normal distribution, this value can be halved with little effect over the power.

For the confidence level of $\alpha = 0.05$ we find that $k \leq 3.7654n^{2/5}$, half the value is $1.8827n^{2/5}$ so we will use the recommended $k = 2n^{2/5}$. This value is reasonable for $.01 \leq \alpha \leq .10$ so it will be the default value used. The choice of the number of cells is sometimes a critical issue, the outcome of the test may be different for a data set using a different value of k. An example of this situation is given:

Example. *Performing the Pearson's χ^2 test assuming the asymptotic distribution for 3.1 and choosing $\alpha = 0.05$, with the data -6.7, -2.4, -0.2, 0.1, 0.6, 0.8, 0.8 2.2, 2.8, 3.9 using 5 classes we get $P = 7$, which has $p-$value of 0.0302 hence rejecting the null hypothesis. Using 6 classes we get $P = 2$, which has a $p-$value of 0.5724 and the test fails to reject the null hypothesis.*

3.1.2 Lillifors' test

The Lillifors's test for normality is an extension of the the Kolmogorov-Smirnov test. The test statistic is the same as the Kolmogorov-Smirnov statistic [5]:

$$KS = \max\{D^+, D^-\} \tag{3.3}$$

where

$$D^+ = \max\{i/n - \Phi(\hat{z}_{(i)})\} \qquad \text{and} \qquad D^- = \max\{\Phi(\hat{z}_{(i)}) - (i-1)/n\}$$

D^+ measures the upper diference between the E.D.F. and $\Phi(x)$ while D^- measures the lower difference between the E.D.F. and $\Phi(x)$.

The difference between the Komogorov-Smirnov and the Lillifors test is the set of critical values for each test. Lillifors calculated some of the corrected values for testing composite normality [5].

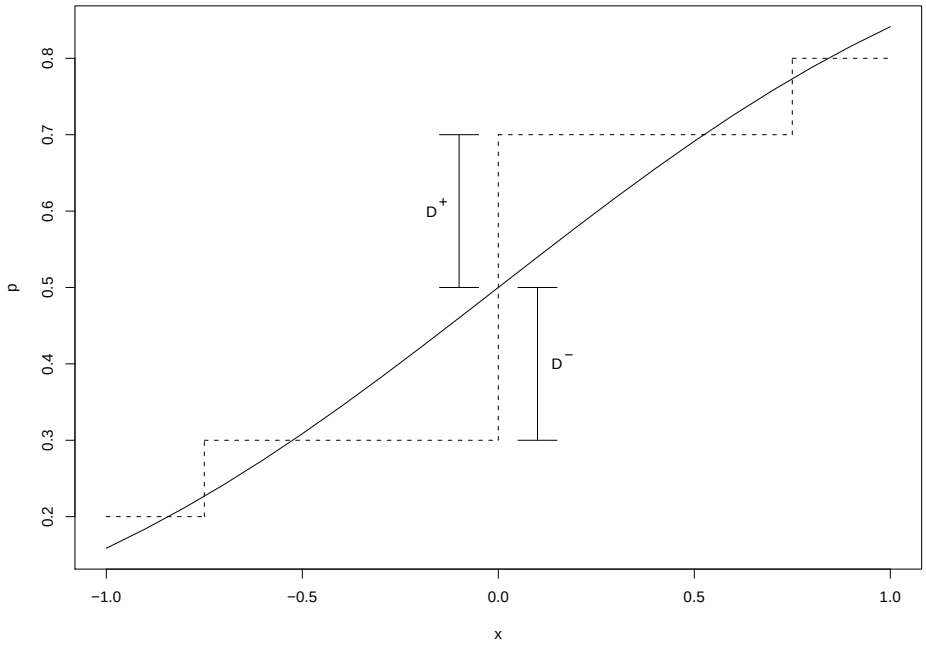

Figure 3–2: D^+ and D^-

Since the test statistic is calculated by using the standardized order statistics, by Proposition 1 in Appendix A the test is invariant to location–scale transformations.

3.1.3 Cramér-von Misses' criterion

The criterion is named after Harold Cramér and Richard Edler von Mises, who first proposed it in 1928-1930 [6] and [7]. This normality test comes from a family of tests that compares the squares of the differences between the E.D.F. of a sample and the C.D.F. by using the statistic:

$$\omega^2 = n \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 \, \psi\left(F(x)\right) dF(x) \tag{3.4}$$

where $\psi\left(F(x)\right)$ is a weighting function. In particular if $\psi\left(F(x)\right) = 1$ and $F(x) = \Phi(x)$, ω^2 is the Cramér-von Misses statistic for testing normality [2]:

$$CVM = \sum_{i=1}^{n} \left\{ \Phi(\hat{z}_{(i)}) - \frac{(2i-1)}{2n} \right\}^2 + \frac{1}{12n} \tag{3.5}$$

This test is invariant to location–scale because it is based on the standardization of the observations and by Proposition 1 in Appendix A this standardization is invariant to location–scale transformations.

3.1.4 Anderson-Darling test

This test is part of the same family defined in 3.4. Anderson and Darling proposed the weighting function $\psi(x) = (x(1-x))^{-1}$ [8], and this function yields the statistic [2]:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left\{\ln\Phi(\hat{z}_{(i)}) + \ln(1-\Phi(\hat{z}_{(n+1-i)}))\right\} \tag{3.6}$$

The resulting test gives more weight to the tails of the distribution than the Cramér-von Misses test.

One problem that this test presents is the calculation of the value of $\ln\Phi(\hat{z}_{(i)})$ and $\ln(1-\Phi(\hat{z}_{(n+1-i)}))$, because the value of $\hat{z}_{(i)}$ can be too close to 0 or 1, which would cause the logarithm to tend to infinity and then the test statistic can not be calculated. In the experiments performed (see Chapter 5) this happens more often with heavy–tailed distributions and large sample sizes.

As with the other E.D.F. based tests discussed before, this test is invariant because it uses the standardized observations to compare the empirical distribution to $\Phi(x)$.

3.2 Moments based tests

The $k-$th central moment of a random variable X is defined as $\mu_k = E((X-\mu)^k)$ for $k \geq 2$. Note that $\mu_2 = \sigma^2$. The tests to be discussed in this section are based on the third and fourth moments and more specifically the standardized moments, named *skewness* and *kurtosis* respectively. The $k-$th standardized moments are found by dividing the $k-$th moment by $\mu_2^{k/2}$.

The skewness of a random variable is a measure of asymmetry, it is defined as $\sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$, a positive skewness means that the right tail of the density is longer while a negative skewness means that the left tail is longer. The sample skewness is denoted by $\sqrt{b_1}$ and is calculated as:

$$\sqrt{b_1} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}} \tag{3.7}$$

For a standard normal distribution, the value of the skewness is 0. It is important to notice that since skewness is only a measure of asymmetry, the value of 0 is not unique for

the standard normal distribution, to have a sample with an estimated skewness near 0 is not evidence enough to assume normality but a significative value different from 0 would be good enough to reject normality. [2]

The kurtosis of a random variable is a measure of the "peakedness" and the "heaviness" of the tails of the p.d.f. of a random variable X, it is defined as $\beta_2 = \mu_4/\mu_2^2$. For a standard normal distribution, it has a value of 3, for this reason it is common to use the *kurtosis excess* which is the calculated as $\beta_2 - 3$, assigning this way to the normal distribution an excess kurtosis of 0.

The sample kurtosis is denoted as b_2 and is calculated as:

$$b_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \tag{3.8}$$

A positive kurtosis excess means that the distribution can have more peakedness or heavier tails or both, while a negative kurtosis excess means that the distribution can be more flat or have shorter tails or both. It is important to be careful and not think that a higher kurtosis implies a higher variance. In the next example we show the excess kurtosis of a normal distribution, a double exponential distribution and an uniform distribution, also the p.d.f is plotted in figure 3–3 all distributions are scaled to $\sigma^2 = 1$.

Example. *Let* $X \sim N(0,1)$, $Y \sim dexp(1/\sqrt{2})$ *and* $Z \sim unif(-\sqrt{3},\sqrt{3})$. *For these variables we have that* $\sigma_x^2 = 1$, $\sigma_y^2 = 2/(\sqrt{2})^2 = 1$ *and* $\sigma_z^2 = (2\sqrt{3})^2/12 = 1$. *And the kurtosis excess of each variable is* $\beta_{2_x} - 3 = 0$, $\beta_{2_y} - 3 = 3$ *and* $\beta_{2_z} - 3 = -6/5$.

3.2.1 Jarque-Bera test

The test was proposed by Jarque and Bera in 1987 [9]. The test statistic is calculated as:

$$JB = n\left\{\frac{1}{6}(\sqrt{b_1})^2 + \frac{1}{24}(b_2 - 3)^2\right\} \tag{3.9}$$

Where $\sqrt{b_1}$ and b_2 are the sample standardized moments defined in 3.7 and 3.8.

The statistic defined in 3.9 is asymptotically distributed as χ_2^2, although its has a very slow convergency, having problems even with moderately big samples, e.g. $n = 100$ [10].

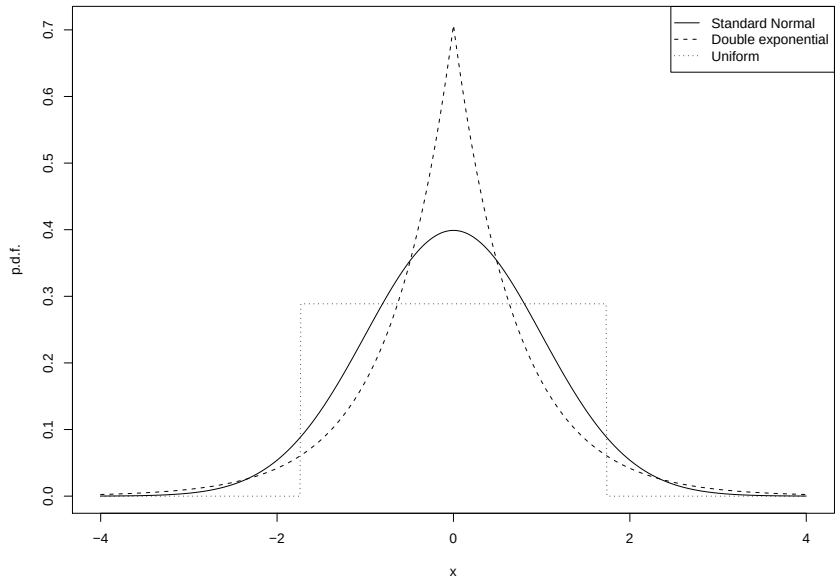

Figure 3–3: p.d.f. of X, Y and Z

One way to overcome this issue is to approximate critical values of JB using Monte Carlo simulation instead of using its asymptotic distribution.

By Proposition 2 of Appendix A the test is invariant to location–scale transformations because it is based on standardized moments and those are invariant to location–scale.

3.2.2 D'Agostino-Pearson test

D'Agostino and Pearson developed a normality test based on the skewness and kurtosis of a sample. They proposed a transformation $Z(x)$ for $\sqrt{b_1}$ and b_2 that asymptotically leads $Z(\sqrt{b_1}) \sim N(0,1)$ and $Z(b_2) \sim N(0,1)$. This yields to the test statistic [11]:

$$DP = Z^2(\sqrt{b_1}) + Z^2(b_2) \tag{3.10}$$

DP is asymptotically distributed as χ^2_2, and this statistic has a faster convergence to its asymptotic distribution than JB.

The transformation is outlined in [12]. It is based on a standardization using the mean and the variance of $\sqrt{b_1}$ and b_2.

Again, as in the Jarque-Bera test, by Proposition 2 of Appendix A the test is invariant to location–scale transformations because it is based on standardized moments and those are invariant to location–scale.

3.3 Correlation based tests

Correlation tests are based on the QQ plots described in chapter 2. The relation of the order statistics $x_{(i)}$ and the quantiles of the expected values of $x_{(i)}$, $\Phi^{-1}(p_i)$, under the null hypothesis tends to be linear, so it is possible to use a measure of the linear correlation between them.

3.3.1 Shapiro-Wilk test

The test was proposed by Shapiro and Wilk in 1965 [13], the test statistic is defined as:

$$SW = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.11}$$

where,

$$\begin{aligned} \boldsymbol{a'} &= \frac{\boldsymbol{w'}\,\boldsymbol{V^{-1}}}{(\boldsymbol{w'}\,\boldsymbol{V^{-1}}\,\boldsymbol{V^{-1}}\,\boldsymbol{w})^{1/2}} \\ w_i &= E(Z_{(i)}) \\ v_{ij} &= \mathrm{cor}(Z_{(i)}, Z_{(j)}) \end{aligned}$$

The reasoning of the test comes from the fact that in a normality plot we can write each observation as $x_i = \mu + \sigma z_i$, then an estimate of σ is found by using generalized least squares, and up to a constant to standardize the linear coefficients, this estimate is $\boldsymbol{a'x}$ [13].

To show why the statistic is invariant to location–scale transformations, for the case of a normal distribution we have that $-a_i = a_{n-i+1}$, since these are based on the order statistics which are symmetric as well [13], and if we have an odd number of observations, say $n = 2k + 1$ for the median $Z_{(k+1)}$ we have $E(Z_{(k+1)}) = 0$ hence $a_{(k+1)} = 0$. Using this

we can rewrite 3.11 as

$$SW = \frac{\left(\sum_{i=1}^{\lfloor n/2 \rfloor} a_{n-i+1}(x_{(n-i+1)} - x_{(i)})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.12}$$

Now given a transformation $y = \alpha x + \beta$ for each x_i, and calculating 3.12 we get:

$$\begin{aligned}
SW_y =& \frac{\left(\sum_{i=1}^{\lfloor n/2 \rfloor} a_{n-i+1}(y_{(n-i+1)} - y_{(i)})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
=& \frac{\left(\sum_{i=1}^{\lfloor n/2 \rfloor} a_{n-i+1}(\alpha x_{(n-i+1)} + \beta - \alpha x_{(i)} - \beta)\right)^2}{\sum_{i=1}^{n}(\alpha x_i + \beta - \alpha\bar{x} - \beta)^2} \\
=& \frac{\left(\sum_{i=1}^{\lfloor n/2 \rfloor} a_{n-i+1}(\alpha x_{(n-i+1)} - \alpha x_{(i)})\right)^2}{\sum_{i=1}^{n}(\alpha x_i - \alpha\bar{x})^2} \\
=& \frac{\alpha^2\left(\sum_{i=1}^{\lfloor n/2 \rfloor} a_{n-i+1}(x_{(n-i+1)} - x_{(i)})\right)^2}{\alpha^2\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
=& SW
\end{aligned}$$

Then SW is invariant to location–scale transformations.

For a significance level of α the test has a rejection region of the form $SW < sw_{crit}$ where the sw_{crit} is the corresponding critical value. Since under the null hypothesis the numerator and denominator of 3.11 estimate the same quantity, σ^2, and in general SW will have a maximum value of 1 [13].

3.3.2 D'Agostino test

Introduced in 1971, the test is based on a statistic D which is up to a constant the ratio of a linear unbiased estimator of the population standard deviation to the sample standard deviation proposed by Downton in 1966 [14]. It is defined as:

$$D = \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) x_{(i)}}{n^{3/2} \sqrt{\sum_{j=1}^{n} (x_i - \bar{x})^2}} \tag{3.13}$$

D'Agostino gives the asymptotic mean and standard deviation and proposed the asymptotic distribution $Y = \frac{D - (2\sqrt{\pi})^{-1}}{asd(D)} \sim N(0,1)$. A problem with this statistic is that it is not really appropriate except for very large values of n, so the use of critical values is recommended with a two-sided rejection region of the form $D \leq d_{\alpha/2}$ and $D \geq d_{(1-\alpha)/2}$ [14].

The statistic in 3.13 is invariant to location–scale transformation, let $y = \alpha x + \beta$ for each x_i, calculation D for the transformed samples we have:

$$\begin{aligned}
D_y &= \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) y_{(i)}}{n^{3/2} \sqrt{\sum_{j=1}^{n} (y_i - \bar{y})^2}} \\
&= \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) (\alpha x_{(i)} + \beta)}{n^{3/2} \sqrt{\sum_{j=1}^{n} (\alpha x_i + \beta - \alpha\bar{x} - \beta)^2}} \\
&= \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) \alpha x_{(i)} + \left(i - \frac{n+1}{2} \right) \beta}{n^{3/2} \sqrt{\sum_{j=1}^{n} (\alpha x_i - \alpha\bar{x})^2}} \\
&= \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) \alpha x_{(i)} + \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) \beta}{\alpha\, n^{3/2} \sqrt{\sum_{j=1}^{n} (x_i - \bar{x})^2}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\alpha \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) x_{(i)} + \sum_{i=1}^{n} i\beta - \frac{n(n+1)}{2}\beta}{\alpha\, n^{3/2} \sqrt{\sum_{j=1}^{n} (x_i - \bar{x})^2}} \\
&= \frac{\alpha \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) x_{(i)} + \frac{n(n+1)}{2}\beta - \frac{n(n+1)}{2}\beta}{\alpha\, n^{3/2} \sqrt{\sum_{j=1}^{n} (x_i - \bar{x})^2}} \\
&= \frac{\alpha \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) x_{(i)}}{\alpha\, n^{3/2} \sqrt{\sum_{j=1}^{n} (x_i - \bar{x})^2}} \\
&= \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) x_{(i)}}{n^{3/2} \sqrt{\sum_{j=1}^{n} (x_i - \bar{x})^2}} \\
&= D
\end{aligned}
$$

Then 3.13 is invariant to location–scale transformations.

3.4 Entropy based tests

The entropy of a continuous random variable is defined as

$$H = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

Shannon [15], showed that the maximum entropy for continuous distributions is $H = \log \sqrt{2\pi e}\, \sigma$ and it is attained by a Normal distribution with standard deviation σ. Since we are interested in testing composite normality we choose $\sigma = 1$ and standardize the observations from the sample. Using this it is possible to construct a test statistic of the form:

$$T = \frac{\exp \hat{H}}{\hat{\sigma}} \tag{3.14}$$

where $\hat{H}$ is an entropy estimator, and a rejection region of the form $T < T_{crit}$ where T_{crit} is the appropriate critical value.

We consider two estimators of entropy, based on sample $m-$spacings of the form $x_{(i+m)} - x_{(i-m)}$ and $x_{(i+m)} - x_{(i)}$, where m is a positive integer such that $m \leq n/2$. Also, we define $x_{(i)} = x_{(1)}$ if $i < 1$ and $x_{(i)} = x_{(n)}$ if $i > n$.

3.4.1 Vasicek test

Vasicek defined the following estimator for entropy:

$$\hat{H}_{vas,\,m} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n}{2m}[x_{(i+m)} - x_{(i-m)}]\right) \tag{3.15}$$

using 3.15 and 3.14 the Vasicek test statistic [16] is defined as:

$$VAS_m = \frac{n}{2m\hat{\sigma}}\left[\prod_{i=1}^{n}(x_{(i+m)} - x_{(i-m)})\right]^{1/n} \tag{3.16}$$

After applying the location–scale transformation $y = \alpha x + \beta$ the statistic 3.16 is calculated as:

$$\begin{aligned}
VAS_{m,y} &= \frac{n}{2m\hat{\sigma}_y}\left[\prod_{i=1}^{n}(y_{(i+m)} - y_{(i-m)})\right]^{1/n} \\
&= \frac{n}{2m\alpha\hat{\sigma}}\left[\prod_{i=1}^{n}(\alpha x_{(i+m)} + \beta - \alpha x_{(i-m)} - \beta)\right]^{1/n} \\
&= \frac{n}{2m\alpha\hat{\sigma}}\left[\prod_{i=1}^{n}(\alpha x_{(i+m)} - \alpha x_{(i-m)})\right]^{1/n} \\
&= \frac{n}{2m\alpha\hat{\sigma}}\alpha\left[\prod_{i=1}^{n}(x_{(i+m)} - x_{(i-m)})\right]^{1/n} \\
&= \frac{n}{2m\hat{\sigma}}\left[\prod_{i=1}^{n}(x_{(i+m)} - x_{(i-m)})\right]^{1/n} \\
&= VAS_m
\end{aligned}$$

Then 3.16 is invariant to location–scale transformations.

3.4.2 Van Es test

It is based on the estimator proposed by Van Es [17]:

$$\hat{H}_m = \frac{1}{n-m}\sum_{i=1}^{n-m} \log\left(\frac{n+1}{m}(x_{(i+m)} - x_{(i)})\right) + \sum_{k=m}^{n}\frac{1}{k} + \log\frac{m}{n+1} \tag{3.17}$$

Using 3.14 with 3.17 results on the test statistic:

$$VAN_m = \frac{1}{\hat{\sigma}} \exp\left[\sum_{k=m}^{n} \frac{1}{k}\right] \left[\prod_{i=1}^{n-m} (x_{(i+m)} - x_{(i)})\right]^{1/(n-m)} \tag{3.18}$$

After applying the location–scale transformation $y = \alpha x + \beta$ the statistic 3.16 is calculated as:

$$\begin{aligned}
VAN_{m,y} &= \frac{1}{\hat{\sigma}_y} \exp\left[\sum_{k=m}^{n} \frac{1}{k}\right] \left[\prod_{i=1}^{n-m} (y_{(i+m)} - y_{(i)})\right]^{1/(n-m)} \\
&= \frac{1}{\alpha\hat{\sigma}} \exp\left[\sum_{k=m}^{n} \frac{1}{k}\right] \left[\prod_{i=1}^{n-m} (\alpha x_{(i+m)} + \beta - \alpha x_{(i)} - \beta)\right]^{1/(n-m)} \\
&= \frac{1}{\alpha\hat{\sigma}} \exp\left[\sum_{k=m}^{n} \frac{1}{k}\right] \left[\prod_{i=1}^{n-m} (\alpha x_{(i+m)} - \alpha x_{(i)})\right]^{1/(n-m)} \\
&= \frac{1}{\alpha\hat{\sigma}} \alpha \exp\left[\sum_{k=m}^{n} \frac{1}{k}\right] \left[\prod_{i=1}^{n-m} (x_{(i+m)} - x_{(i)})\right]^{1/(n-m)} \\
&= \frac{1}{\hat{\sigma}} \exp\left[\sum_{k=m}^{n} \frac{1}{k}\right] \left[\prod_{i=1}^{n-m} (x_{(i+m)} - x_{(i)})\right]^{1/(n-m)} \\
&= VAN_m
\end{aligned}$$

Then 3.18 is invariant to location–scale transformations.

A difficulty that both the Vasicek test and the Van Es test present is how to choose an appropriate value of m because there are no guidelines available. It is necessary then to investigate the power of the test for different values of m when n and the alternative distribution are changing to choose a value that maximizes it.

In chapter 5 we show the power of both Vasicek test and Van Es test, for values of m from 1 to 10 using different distributions. In all the cases studied we observe that there is a tendency and find an optimal value for each case.

CHAPTER 4
THE ENVELOPE TEST

The proposed test, called envelope test, is an extension of the normality plots reviewed in chapter 2. The procedure consists in adding a confidence band to the normality plot. For each sample quantile upper and lower limits are plotted such that the total coverage probability is a fixed value $1-\alpha$. Adding a confidence band to test distributional assumptions has been discussed before, one proposed method is inverting the Kolmogorov-Smirnov test, but overall the method has a poor performance [18].

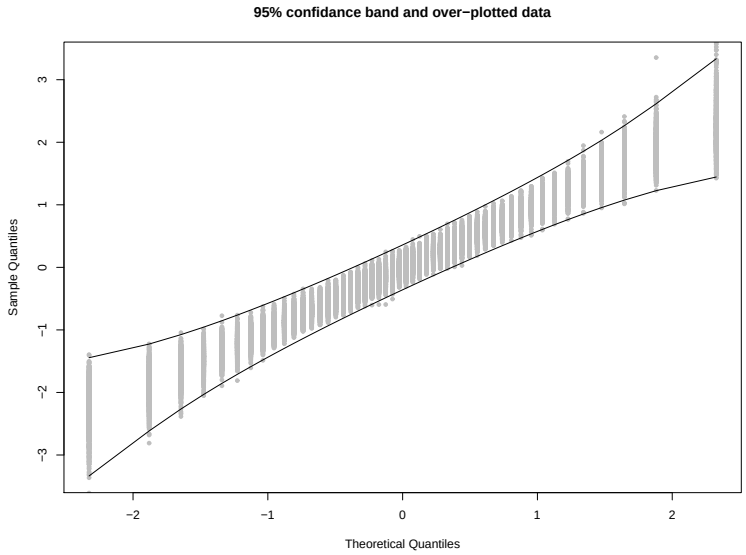

Figure 4–1: Coverage of the confidence band for $n = 50$ and $\alpha = 0.05$.

Figure 4–1 shows an example of a confidence band (solid lines) constructed with the envelope test, 1000 samples of size 50 were taken, for each one its normality plot was constructed (grey dots). The confidence band gives upper and lower limits of each sample quantile, the test would reject the null hypothesis if at least one observation of the sample falls outside of the confidence band.

4.1 Construction of the confidence band

Let $X_1, \ldots, X_n$ be an i.i.d. sample. To construct the base normality plot, we will use the points:

$$p_i = \begin{cases} \Phi^{-1}\left(\dfrac{i-0.375}{n+0.25}\right), & 1 \leq n \leq 10 \\ \Phi^{-1}\left(\dfrac{i-0.5}{n}\right), & n > 10 \end{cases}$$

as proposed by Blom [19].

Since our interest is testing for composite normality we need to standardize our sample. We then define the standardized order statistics:

$$\hat{Z}_{(i)} = \frac{X_{(i)} - \bar{X}}{S}$$

To obtain an overall level of $1-\alpha$ we need to find limits $\boldsymbol{l}$ and $\boldsymbol{u}$ such that

$$P(l_i \leq \hat{Z}_{(i)} \leq u_i;\ i = 1, \ldots, n) = 1 - \alpha \tag{4.1}$$

When data comes from a standard normal distribution, by lemma 3 of appendix A we have that $\Phi(X_{(k)}) \sim Beta(k, n-k+1)$, this gives us a starting point for estimation:

$$\begin{aligned} l_k(\alpha, n) &= \Phi^{-1}\left(\beta^{-1}(\alpha/2;\ k, n-k+1)\right) \\ u_k(\alpha, n) &= \Phi^{-1}\left(\beta^{-1}(1-\alpha/2;\ k, n-k+1)\right) \end{aligned} \tag{4.2}$$

One problem with this is that the ordered statistics are not independant and we do not have an explicit formula for their joint distribution or the joint distribution of the $\Phi(X_{(k)})$, also after standardizing the distribution of $\Phi(X_{(k)})$ is not $Beta(k, n-k+1)$, but the true distribution of the standardized statistics is free from the mean and standard deviation. Because of this we use to a Monte Carlo simulation to estimate appropriate limits to reach the desired coverage probability.

The procedure used to estimate the probability in 4.1, is based on the law of large numbers. We define a bernoulli random varible T that has value 1 if every standardized observation of a normal random sample lies within the confidence band and 0 otherwise. We want to find a value of p that yields to a probability of $1-\alpha$, that is, we need that the sucess rate of T to be $1-\alpha$.

We define $T_1,\ldots,T_k$, a sequence of i.i.d. T random variables. The mean of each T_i is its success rate, $1-\alpha$. The strong law of large numbers states that:

$$\lim_{k\to\infty}\frac{T_1+\cdots+T_k}{k}=1-\alpha \tag{4.3}$$

with probability 1.[23]

We define the function

$$\Psi(p)=P(l_i(p,n)\leq\hat{Z}_{(i)}\leq u_i(p,n);\ i=1,\ldots,n) \tag{4.4}$$

with limits of the form 4.2. We want to find a value of p such that $\Psi(p)=1-\alpha$ for a fixed α.

The function 4.4 is stricly decreasing in p, if $p\to 0$ the limits $l_k(\alpha,n)\to-\infty$ and $u_k(\alpha,n)\to+\infty$ and the probability is 1, when p increases each interval becomes smaller, thus reducing the overall probability. This assures us that there is a unique solution to $\Psi(p)=1-\alpha$. We will use a bisection algorithm to find the solution.

We start with $p_l=0$ and $p_h=1$, for $m=(p_l+p_h)/2$ estimate $\Psi(m)$. If it is greater than $1-\alpha$ we set $p_l=m$ or $p_h=m$ if it is less. Then repeat until we reach $p_h-p_l<e$ for a fixed e. The estimation of $\Psi(m)$ is done by generating k samples of size n and then using 4.3, that is the proportion of the number of samples which all of their standardized ordered observations $\hat{z}_i$ fell between l_i and u_i.

For $\alpha=0.05$ a fit of the form $p=an^b$ was found using linear fitting with data transformation. This is useful because the estimation of p using the Monte Carlo simulation requires some computational effort even for values of n which are not very big, this calculation is shown in section 5.3.

4.2 Differences between existing procedures to calculate a confidence band

The method developed will be compared to two existing procedures, one in the library "car" available in R that draws point-wise envelope band and the confidence band displayed in *Minitab* for a probability plot.

We will show three examples using the data corresponding to the measures of the depth of earthquakes from the data set "quakes" found in the datasets package of R [20].

We will use samples of 10, 20 and 50 observations of earthquakes with magnitude between 4 and 4.1.

Example. *For a sample of* $n = 10$*, figure 4–2 is the outcome of using first the function* `qqnorm` *to construct the normality plot and* `qqenvl` *to construct the proposed confidence band, figures 4–3 and 4–4 are the plots generated by the function* `qq.plot` *of the library car in R and the plot generated by Minitab respectively. Figure 4–5 is the result of the envelope test* `envl.plot` *(solid lines) and drawing the corresponding confidence band given by Minitab (dotted lines) after standardizing the data.*

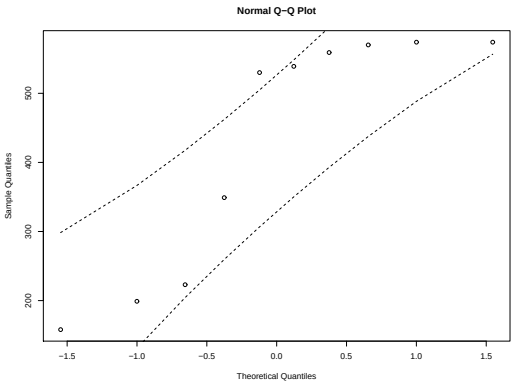

Figure 4–2: Envelope method, $n = 10$

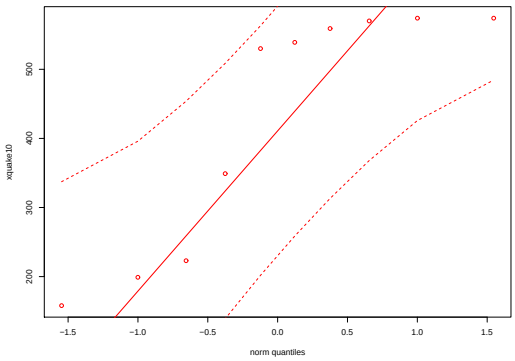

Figure 4–3: Confidence band from *car* library, $n = 10$

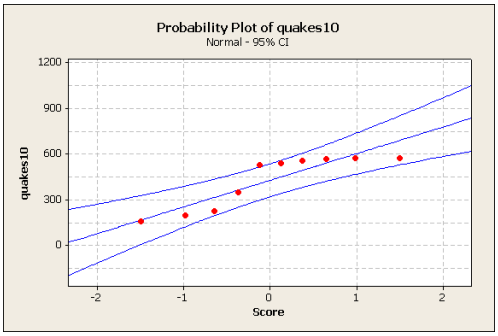

Figure 4–4: Confidence band from *Minitab*, $n = 10$

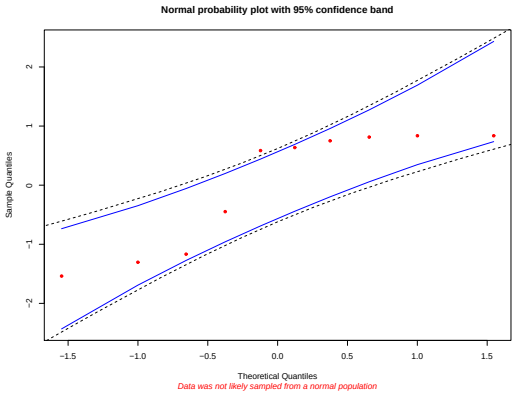

Figure 4–5: Envelope method with *Minitab* confidence band, $n = 10$

From the plots it can be easily seen that the confidence band from the car library is wider than that generated by the proposed method, which would result in a total coverage confidence of more than 95%. The confidence band drawn by Minitab is just a little wider than the proposed method's.

Example. *For a sample of $n = 20$, figure 4–6 is the outcome of using first the function* `qqnorm` *to construct the normality plot and* `qqenvl` *to construct the proposed confidence band, figures 4–7 and 4–8 are the plots generated by the function* `qq.plot` *of the library car in R and the plot generated by Minitab respectively. Figure 4–9 is the result of the envelope test* `envl.plot` *(solid lines) and drawing the corresponding confidence band given by Minitab (dotted lines) after standardizing the data.*

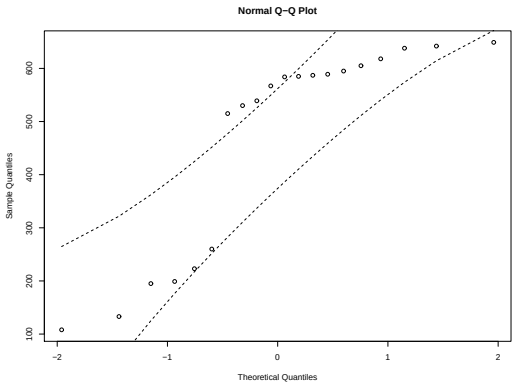

Figure 4–6: Envelope method, $n = 20$

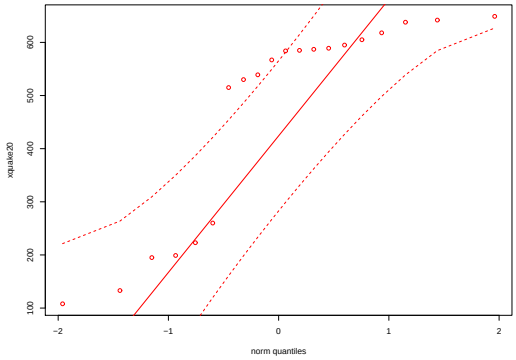

Figure 4–7: Confidence band from *car* library, $n = 20$

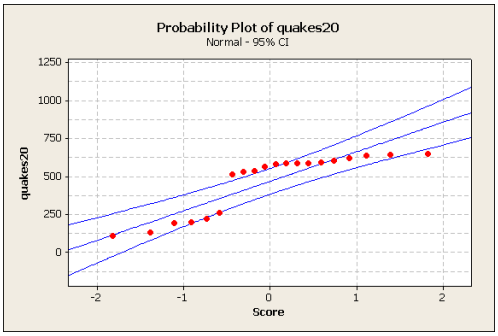

Figure 4–8: Confidence band from *Minitab*, $n = 20$

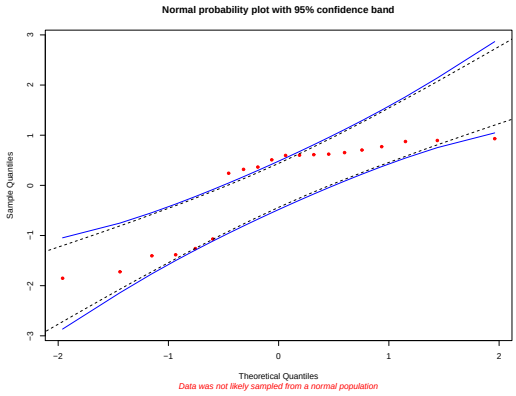

Figure 4–9: Envelope method with *Minitab* confidence band, $n = 20$

In the figures corresponding to $n = 50$ it can be seen that the confidence band corresponding to the car library is wider than both the Minitab and the proposed method's confidence band. The Minitab's and the proposed method's band are close to each other in the middle values and start to differ towards the extremes, also since Minitab's confidence band is contained in the proposed method's the total coverage confidence will likely be less than 95%.

Example. *For a sample of n=50, figure 4–10 is the outcome of using first the function* `qqnorm` *to construct the normality plot and* `qqenvl` *to construct the proposed confidence band, figures 4–11 and 4–12 are the plots generated by the function* `qq.plot` *of the library car in R and the plot generated by Minitab respectively. Figure 4–13 is the result of the envelope test* `envl.plot` *(solid lines) and drawing the corresponding confidence band given by Minitab (dotted lines) after standardizing the data.*

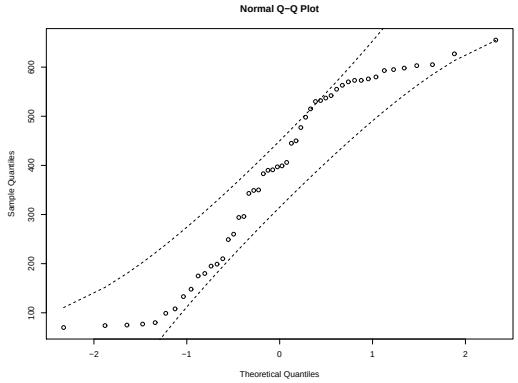

Figure 4–10: Envelope method, $n = 50$

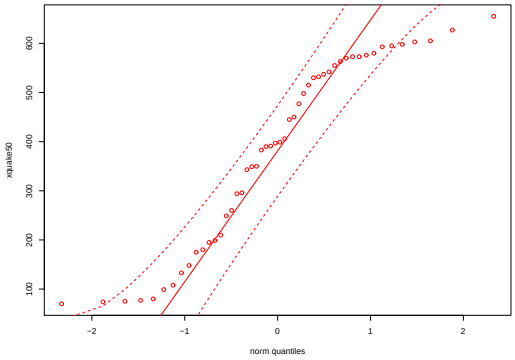

Figure 4–11: Confidence band from *car* library, $n = 50$

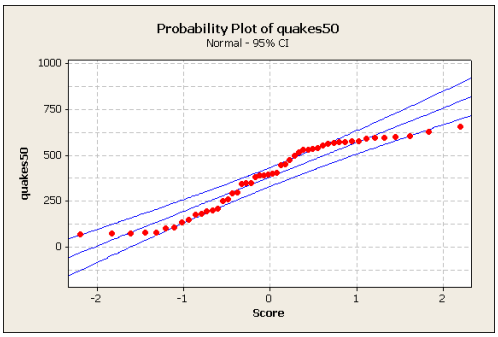

Figure 4–12: Confidence band from *Minitab*, $n = 50$

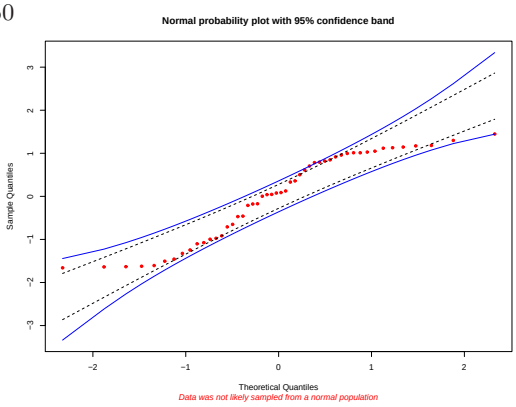

Figure 4–13: Envelope method with *Minitab* confidence band, $n = 50$

The confidence band for the car library is, once again, wider than Minitab's and the proposed method's confidence band. Also, the difference between the Minitab's and the proposed method's band is larger than what it was with a smaller sample size, Minitab's band is enclosed by the proposed method hence the total coverage confidence for the former will be less than 95%, the proposed method's confidence.

In table 5–1 of section 5.1 we see that the empirical value for α is close to 0.05 for sample sizes ranging from 20 to 200. We do not have a way to calculate the empirical

power of the *Minitab*'s confidence band, however, when *Minitab*'s band lies within the proposed band, the empirical α of the proposed test gives us evidence that the true α of *Minitab*'s band will be greater than 0.05 resulting in a total coverage confidence of less than 95% and will be greater than 95% if the lie outside. As a check, a 100 normal samples of size 50 were generated in *Minitab* and from these, 39 resulted being rejected as normal by a critirion based on the confidence band.

CHAPTER 5
SIMULATIONS

We are interested in comparing the power of the numerical tests described in Chapter 3 and the Envelope test described in Chapter 4. Critical values for $\alpha = 0.05$ were calculated for each test, testing is done using different alternative distributions to see how different distributions affect the power of the test and see what test has a greater power in each case. The alternative distributions were chosen by classifying them on five categories: infinite support, asymmetric with infinite support, semi–infinite support, support in $[0, 1]$ and a Normal distribution with an outlier. The distributions are:

1. Symmetric and infinite support:
   (a) $t(1)$
   (b) $t(2)$
   (c) $t(5)$
   (d) $t(10)$
   (e) $DExponential(1)$
2. Asymmetric with infinite support:
   (a) $Gumbel(0, 1)$
   (b) $Gumbel(1, 3)$
3. Semi–infinite support:
   (a) $exp(1)$
   (b) $Gamma(2, 1/2)$
   (c) $Gamma(5, 1/2)$
   (d) $LogNorm(0, 1)$
   (e) $LogNorm(0, 1/2)$
   (f) $Weibull(5, 2)$

4. Support $[0, 1]$
   (a) $Uniform$
   (b) $Beta(1/2, 1/2)$
   (c) $Beta(1, 1/2)$
   (d) $Beta(5, 2)$
   (e) $Triangular(0, 0.5, 1)$
5. Normal with Outlier
   (a) $N(0, 1)$ with outlier at $3s$
   (b) $N(0, 1)$ with outlier at $4s$

First For the Vasicek and Van Es test, different values of m were tested and are presented in section 5.2, the one that presented the highest power is the considered in the power calculation.

The power was approximated by simulating 100000 samples of different sizes from 20 to 200, then looking at the proportion of rejections. The critical values were calculated simulating a sample of 50000 and then choosing the corresponding ordered statistic for the desired level.

Table 5–1 show the empirical level α for each test in each sample size. This is the probability of rejecting the null hypothesis when it is true. The empirical level is close to the theoretical level $\alpha = 5\%$

Table 5–1: Empirical α for each test.

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	7	5.16	5.08	5.01	4.78	5.09	5.27	4.64	5.35	5.14	5.18
30	4.97	4.78	5.11	5.16	5.13	4.91	5.18	5.26	5.28	5.14	5.17
40	5.55	5.04	4.95	4.74	5.37	5.19	4.75	4.65	5.33	5.15	4.95
50	5.28	4.8	4.87	5.2	5.14	4.98	4.94	4.93	5.16	5.39	4.98
60	4.52	5.21	4.93	5.08	4.73	4.93	4.87	4.89	5.2	5.19	4.88
70	5.89	4.72	5.34	4.78	5.22	4.93	4.93	5.38	5.17	5.21	4.94
80	5.02	5.09	5.34	4.97	4.69	5.12	4.72	5.1	5.15	5.29	4.89
90	5.18	5.21	5.12	5.14	5.28	5.13	5.08	5.3	5.28	5.18	5.13
100	5.11	4.99	4.81	5.12	4.98	5.05	4.74	5.07	5.05	5.1	5.02
120	4.92	5.02	5.14	4.88	5.04	4.9	4.99	5.37	4.86	4.93	5.23
140	5.09	4.94	5.2	5.26	5.03	5.11	5.3	5.01	4.9	4.86	4.89
160	5.07	4.61	5.11	4.99	5.16	4.89	5.41	5	5.04	5	4.87
180	4.64	5.12	5.17	4.9	5.23	5.01	4.72	4.88	5.07	5.16	5.03
200	4.96	5.01	5.43	5.45	5.22	5.12	5.16	4.77	5.18	5.2	5.13

Table 5–2: Power against $t(1)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	80.38	84.92	88.09	88.2	85.99	85.67	86.93	88.02	76.78	91.71	86.56
30	90.7	94.28	96.37	96.5	95.58	94.85	95.9	96.72	90.96	98.06	95.47
40	96.26	98.14	98.96	98.99	98.55	98.17	98.71	99.18	96.71	99.55	98.35
50	98.46	99.31	99.65	99.7	99.58	99.37	99.64	99.76	98.95	99.87	99.45
60	99.32	99.83	99.93	99.93	99.84	99.76	99.88	99.95	99.64	99.98	99.8
70	99.78	99.93	99.98	99.98	99.96	99.92	99.97	99.98	99.88	99.99	99.92
80	99.9	99.98	99.99	*	99.98	99.97	99.99	99.99	99.97	100	99.98
90	99.96	99.99	100	*	99.99	99.99	100	100	99.99	100	99.99
100	99.98	100	100	*	100	100	100	100	100	100	100
120	100	100	100	*	100	100	100	100	100	100	100

* The Anderson-Darling test can not be computed for many samples of this size.

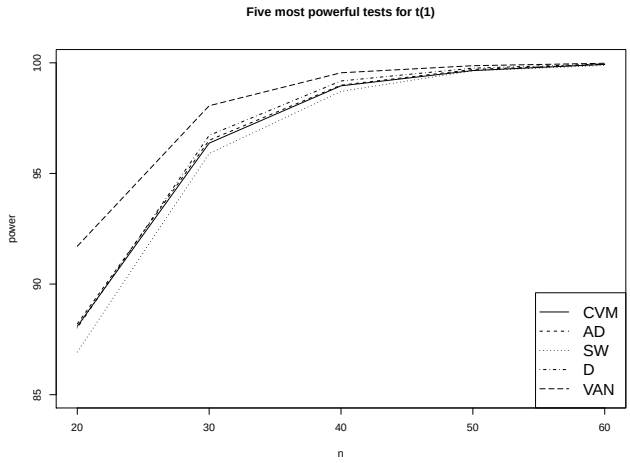

Figure 5–1: Better performing tests against $t(1)$

5.1 Power calculation

5.1.1 Symmetric distributions with infinite support

$t(1)$

Table 5–2 shows the approximated power for different sample sizes. For the $t(1)$ distribution all tests have good performance, the exception would be the Vasicek test presenting the lowest power for a sample size of 20. Also the Anderson-Darling can not be computed for large sizes but every test reach a power of more than 99% with a sample size of 60. This is not really surprising because of the heavy–tailness of a $t(1)$ we expect to see observations either too big or too small for a normal distribution. Figure 5–1 shows the five tests with best performance, that is the Cramér–von Mises test, the Anderson–Darling test, the Shapiro–Wilk test and Van Es test.

Table 5–3: Power against $t(2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	38.63	45.82	51.28	52.74	56.05	56.18	53.28	54.12	33.26	61.27	53.62
30	46.74	59.16	66.78	68.67	72.46	70.78	69.18	72.12	48.4	76.64	69.1
40	57.57	70.28	77.06	78.48	82.42	80.78	78.98	82.82	60.92	86.2	78.45
50	64.47	77.5	84.29	86.28	88.6	87.01	86.58	89.65	71.15	91.71	85.26
60	69.65	84.09	89.34	90.62	92.11	91	90.89	93.44	77.85	94.63	89.43
70	78.32	88.28	93.31	93.96	95.42	94.17	94.28	96.27	84.17	96.5	92.78
80	81.52	92.02	95.7	94.63	96.84	96.24	96.45	97.83	88.35	97.85	95.09
90	84.81	94.14	96.93	*	98.1	97.47	97.57	98.67	91.43	98.57	96.53
100	88.31	96.03	98.03	*	98.84	98.46	98.55	99.28	93.89	99.04	97.62
120	92.87	98.05	99.24	*	99.51	99.29	99.38	99.73	96.87	99.6	98.83
140	95.53	99.06	99.71	*	99.81	99.72	99.79	99.92	98.47	99.81	99.46
160	97.39	99.54	99.87	*	99.94	99.9	99.92	99.97	99.19	99.91	99.74
180	98.51	99.81	99.95	*	99.97	99.95	99.96	99.98	99.62	99.95	99.88
200	99.17	99.94	99.99	*	100	99.99	99.99	100	99.82	99.98	99.94

* The Anderson-Darling test can not be computed for many samples of this size.

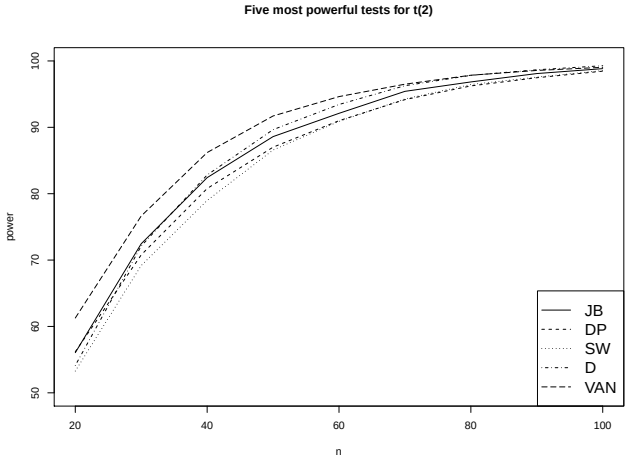

Figure 5–2: Better performing tests against $t(2)$

$t(2)$

For $t(2)$ we can see in table 5–3 that the power of each test is still good, with eight of them having more than 80% with a sample size of 50. Figure 5–2 shows the five most power test for this alternative, with the Van Es test showing the highest which is also the test that achieved highest power for $t(1)$.

$t(5)$

For $t(5)$, since as the degree of freedom rises, the t distribution approaches the normal distribution, the overall power decreases substantially, table 5–4 and figure 5–3 show the Jarque-Bera test is the most powerful for sample sizes up to 160, and then it is surpassed by the D'Agostino test. Also for small samples the Envelope test has a comparable power to the D'Agostino test and the Shapiro-Wilk test.

Table 5–4: Power against $t(5)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	12.28	13.55	15.75	17.11	22.3	22.81	18.95	18.11	8.27	22.56	19.93
30	10.95	15.55	20.09	22.24	30.87	29.9	25.24	25.99	10.98	30.34	27.26
40	12.83	19.03	23.64	25.73	38.22	36.67	30.18	32.97	13.07	37.49	32.98
50	13.25	20.83	26.94	30.96	43.31	41.31	35.34	38.75	15.63	43.28	37.7
60	12.82	24.37	30.43	34.72	47.59	46.33	40.4	44.64	17.96	47.96	42.04
70	16.42	25.33	34.31	37.02	53.3	50.3	44.32	49.85	20.63	52.85	45.73
80	16.15	28.76	37.89	41.45	56.43	54.94	48.28	54.65	22.97	56.78	49.68
90	16.83	31.76	40.53	45.17	61.46	58.55	52.63	59.5	25.16	59.59	53.43
100	18.39	33.39	42.65	48.41	64.3	61.67	55.72	63.17	27.65	62.12	56.07
120	19.85	37.91	49.64	54.23	70.67	67.58	62.66	70.44	32.34	67.04	61.8
140	21.69	41.62	54.52	59.8	75.49	72.82	68.43	76.09	35.77	71	65.92
160	23.82	44.44	58.96	64.08	80.58	77.57	74.44	80.57	40.99	74.43	70.51
180	25.58	50.96	64.73	69.29	83.94	81.26	77.34	84.23	44.73	77.24	74.12
200	28.33	54.44	68.92	73.54	86.91	84.44	81.29	87.34	48.13	79.36	77.45

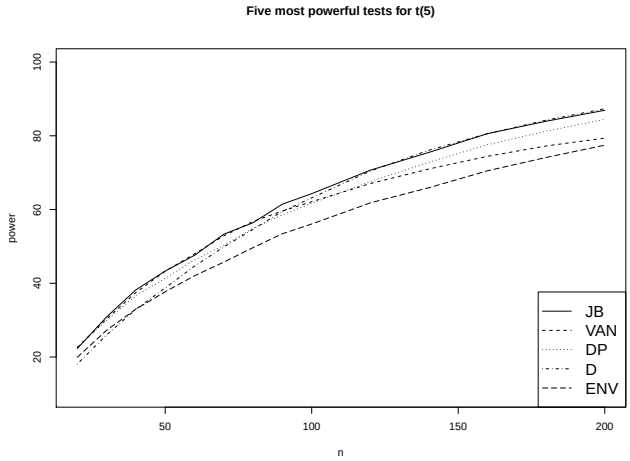

Figure 5–3: Better performing tests against $t(5)$

$t(10)$

Table 5–5 shows that for $t(10)$ the four test based on the E.D.F. have power significantly lower than the other tests. The power is higher in the moments–based tests, the Anderson–Darling and the Jarque-Bera test, followed by the D'Agostino test. The envelope test also presents a reasonable power, higher than the Shapiro–Wilk and close to the Van-Es. Figure 5–4 shows the power of the five most powerful tests.

When we have a t distribution as an alternative, a good tests are the Jarque-Bera test and the D'Agostino Pearson test, both based on sample moments, this is probably because the value of the kurtosis is high if the d.f. is not high, because the first moments are not well defined so the sample moments usually return high values. Also the Van Es shows a good result for an adequate choice of m. The envelope test's power falls between the

Table 5–5: Power against $t(10)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	8.45	7.6	8.22	8.76	11.9	12.45	10.16	9.3	5.41	11.74	10.43
30	6.39	7.65	9.37	10.22	15.57	15.05	12.24	12.01	5.74	14.3	13.4
40	7.24	8.39	9.98	10.79	18.78	18.1	13.59	14.16	6.16	17.05	15.77
50	6.77	8.73	10.36	12.36	20.44	19.4	15.12	15.96	6.62	19.5	17.26
60	5.98	9.75	11.48	13.39	22.28	21.98	17.01	18.14	6.8	20.66	19.24
70	7.74	9.42	12.71	13.42	25.75	23.97	18.64	20.48	7.27	22.95	21.07
80	6.78	10.23	13.37	14.65	26.19	25.71	19.56	22.09	7.54	24.61	22.68
90	6.75	10.62	13.4	15.49	28.95	27.21	21.36	23.91	7.93	26.13	24.08
100	7.2	10.86	13.53	16.18	30.39	28.92	22.47	25.71	8.54	27.17	25.02
120	7.15	11.73	15.61	17.7	34.01	31.71	25.3	29.33	9.28	29.29	27.67
140	7.21	11.97	16.69	19.58	37.25	35.52	29.04	33.18	9.73	30.96	30.2
160	7.62	12.18	17.83	20.93	40.65	37.74	32.01	35.89	10.45	32.9	32.47
180	7.68	14.64	20.12	23.16	44.28	41.08	33.35	39.54	11.36	34.25	35.05
200	8.44	15.12	21.74	25.27	46.98	43.88	36.58	42.48	12.11	34.97	37.22

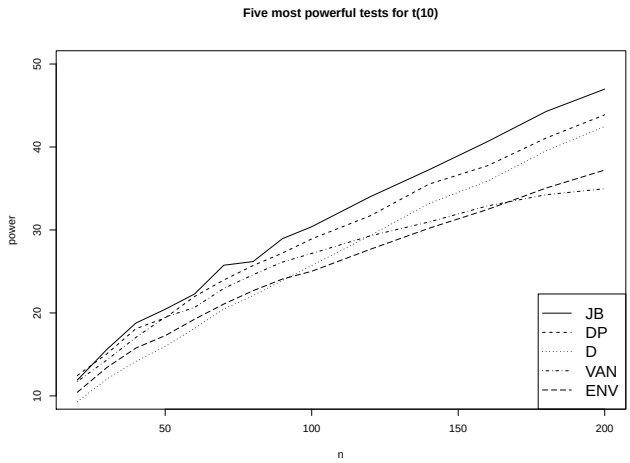

Figure 5–4: Better performing tests against $t(10)$

power of the D'Agostino test and Shapiro–Wilk, showing a higher power than the E.D.F. tests.

$Dexp(1)$

In table 5–6 we can see that against a double exponential distribution, almost every test shows a good power, for smaller sample sizes the Van Es test performs better while for bigger sample sizes the D'Agostino test is the one with the highest power, this can be seen in figure 5–5. The tests that perform poorly are the Pearson test and the Vasicek test. The envelope test has a lower power than most alternatives, up to a sample size of 50 it has a similar performance to other tests that do well in general.

Table 5–6: Power against $Dexp(1)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	18.23	22.48	26.35	26.9	29.3	29.65	26.43	26.96	11.3	38.02	27.49
30	18.95	28.85	36.57	37.51	40.29	38.33	36.18	39.79	16.83	51.74	36.9
40	24	36.75	45.17	45.28	49.19	46.54	43.35	50.68	22.65	61.65	43.63
50	27.49	42.93	53.31	55.15	55.97	52.54	52.11	60.05	29.05	68.45	50.36
60	28.73	50.35	60.65	61.91	60.93	58.11	58.45	67.39	35.09	73.33	55.15
70	37.73	54.7	68.29	67.64	67.81	63.13	64.93	74.15	40.2	77.93	59.9
80	39.67	61.6	74.4	74.25	71.78	68.87	70.73	79.91	45.52	81.47	65.05
90	42.42	66.45	78.17	78.65	76.31	71.93	75.48	83.95	51.17	84.32	68.41
100	47.51	71.07	82.22	83.34	80.22	76.32	79.59	87.65	55.96	86.26	72.22
120	54.24	78.23	88.53	88.53	85.38	81.51	85.82	92.36	64.41	89.91	77.66
140	60.57	83.88	92.63	92.94	89.78	86.97	90.87	95.32	71.8	92.8	82.13
160	66.91	87.72	95.38	95.42	93.13	90.52	94.39	97.16	77.65	94.22	86.26
180	72.03	91.94	97.17	97.16	95.24	93.22	95.91	98.32	82.16	96.04	89.53
200	77.25	94.16	98.43	98.46	96.77	95.29	97.62	99.02	86.34	96.97	91.86

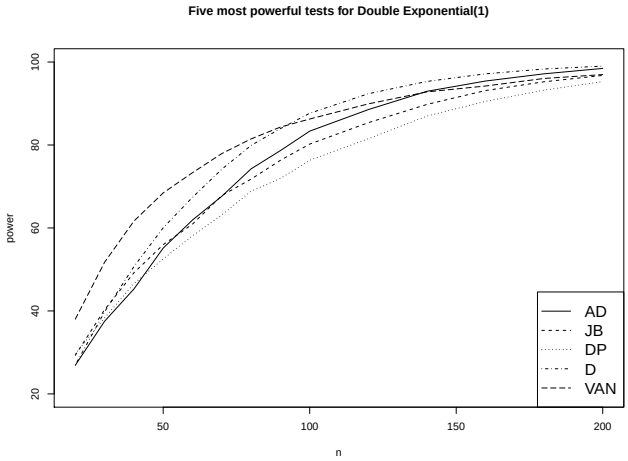

Figure 5–5: Better performing tests against $Dexp(1)$

5.1.2 Asymmetric distributions with infinite support

$Gumbel(0, 1)$

Table 5–7: Power against $Gumbel(0, 1)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	17.68	20.7	24.99	27.34	28.28	28.49	32.05	19.28	20.91	21.71	25.88
30	18.32	27.61	35.66	39.68	41.5	40.4	46.38	27.18	30.92	30.24	37.01
40	24.22	36.64	45.33	49.31	53.13	51.94	57.42	33.49	40.21	38.32	46.61
50	27.14	42.97	54.01	60.79	62.48	61.38	68.71	39.11	48.22	45.65	55.49
60	29.4	51.82	62.71	68.82	68.79	69.61	76.83	44.61	56.77	52.18	63.11
70	38.49	57.28	70.8	75.1	78.12	76.97	83.25	50.29	63.87	59.05	70.17
80	41.07	63.89	76.71	81.06	81.93	83.1	87.98	54.34	69.15	64.04	76.05
90	44.77	68.94	80.53	85.49	87.21	87.18	91.44	58.88	74.59	68.92	80.89
100	50.14	73.49	84.24	89.24	90.65	91.09	94	62.49	78.42	73.51	84.81
120	57.49	81.03	90.74	93.7	95.18	95.33	97.2	69.02	85.44	80.46	90.86
140	65.29	86.34	94.58	96.83	97.68	97.97	98.89	74.96	90.38	85.9	94.57
160	72.14	90.01	96.75	98.27	99.03	99.09	99.54	79.15	93.7	90.12	96.96
180	77.55	93.88	98.28	99.16	99.6	99.63	99.78	83.15	95.83	93.17	98.31
200	83.21	95.91	99.1	99.57	99.82	99.83	99.9	86.04	97.32	95.12	99.07

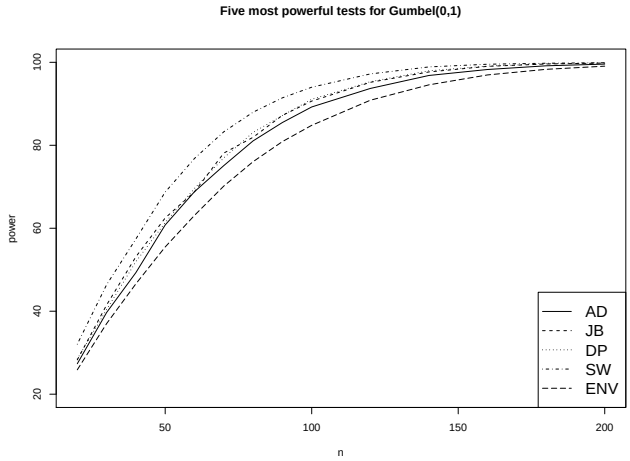

Figure 5–6: Better performing tests against $Gumbel(0, 1)$

In table 5–7 and figure 5–6 we can see that the test that performs better against a $Gumbel(0, 1)$ is the Shapiro–Wilk test, followed by the Jarque–Bera, D'Agostino–Pearson and Anderson-Darling tests. The envelope test follows these last tests performing better than both the entropy–based tests, the D'Agostino test, the Pearson test and the Lilliefors test.

$Gumbel(1, 3)$

In table 5–8 and figure 5–7 we can see that the results for the $Gumbel(1, 3)$ and the $Gumbel(0, 1)$ do not vary significantly.

Table 5–8: Power against $Gumbel(1,3)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	17.42	20.9	25.15	27.47	28.05	28.36	31.99	19.13	21.13	21.73	25.85
30	18.48	28.02	35.84	39.89	41.62	40.46	46.57	27.03	31.04	30.81	37.1
40	24.07	36.55	45.39	49.47	53.31	52.19	57.63	33.15	40.34	38.23	46.62
50	27.39	43.65	54.45	61.01	62.31	61.2	68.82	39.34	48.48	45.61	55.66
60	29.57	51.58	62.49	68.63	68.89	69.8	76.58	44.7	56.71	52.32	63.03
70	38.48	56.96	70.68	75.33	78.25	77.04	83.77	49.97	63.69	58.95	70.19
80	41.24	63.93	76.6	81.26	82.1	83.35	88.12	54.43	69.3	63.74	76.41
90	44.89	69.15	80.59	85.46	87.43	87.41	91.66	58.29	74.82	69.06	80.68
100	49.95	73.42	84.24	89.09	90.68	91.1	94.09	62.21	78.7	73.43	84.8
120	57.38	80.76	90.8	93.79	95.5	95.65	97.38	69.19	85.39	80.41	90.93
140	65.23	86.3	94.42	96.73	97.65	97.96	98.76	74.7	90.23	85.69	94.43
160	72.14	89.93	96.77	98.23	99.03	99.1	99.54	78.83	93.65	90.11	96.89
180	77.69	93.76	98.29	99.13	99.59	99.61	99.77	83.03	95.86	93.01	98.18
200	83.19	95.87	99.12	99.6	99.86	99.86	99.94	86.36	97.41	95.19	99.18

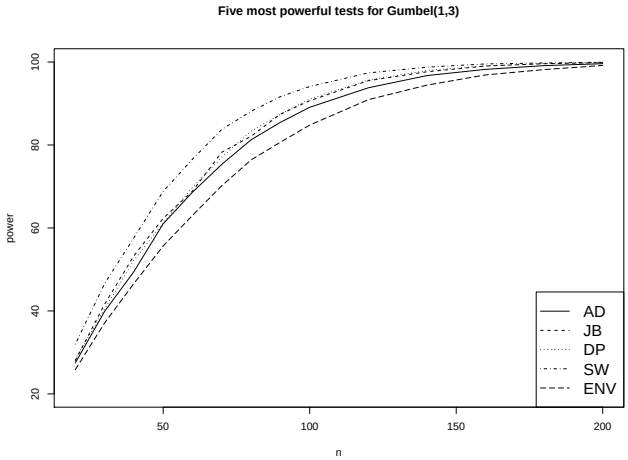

Figure 5–7: Better performing tests against $Gumbel(1,3)$

5.1.3 Distributions with semi–infinite support

$Exp(1)$

Table 5–9: Power against $Exp(1)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	69.2	59.16	72.76	77.64	61.65	59.48	84.31	50.99	85.66	66.51	69.6
30	84.87	77.62	89.65	93.33	82.47	78.3	96.73	68.46	97.5	85.87	92.39
40	95.16	90.22	96.53	98.19	93.52	90.98	99.41	79.94	99.68	94.76	98.88
50	98.41	95.95	98.94	99.7	97.93	96.94	99.93	87.54	99.97	98.21	99.87
60	99.13	98.7	99.73	99.95	99.14	99.07	99.99	92.11	100	99.53	99.99
70	99.9	99.54	99.92	99.98	99.9	99.85	100	95.44	100	99.87	100
80	99.96	99.89	99.99	100	99.98	99.98	100	97.28	100	99.97	100
90	99.97	99.96	99.99	100	100	99.99	100	98.35	100	100	100
100	100	99.99	100	*	100	100	100	99.08	100	100	100

* The Anderson-Darling test can not be computed for many samples of this size.

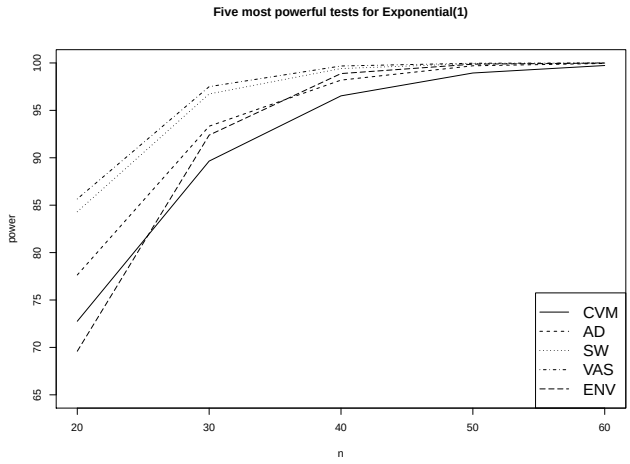

Figure 5–8: Better performing tests against $Exp(1)$

In table 5–9 and figure 5–8 we see that for an exponential distribution most of the classical tests have good performance, three of them and the envelope test exceed a power 90% with a sample size of 30.

$Gamma(2, 1/2)$

Table 5–10 and figure 5–9 shows that when moving from a exponential to a gamma the power starts to decrease, but the same tests that do better for the exponential case do better for this case.

$Gamma(5, 1/2)$

In table 5–10 and figure 5–9 we see that for a greater shape parameter the power decreases even more, this behavior is expected because as the shape parameter increases the distribution starts approximating to a normal distribution. And for a shape value of 5

Table 5–10: Power against $Gamma(2, 1/2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	32.52	33.06	41.47	45.74	39.35	38.81	53.54	27.66	47.49	33.39	38.82
30	41	45.85	59.91	66.48	58.22	55.33	75.44	39.53	69.61	49.42	58.89
40	55.48	59.62	73.66	79.6	73.26	69.87	87.52	48.67	84.04	62.23	75.47
50	65.18	69.36	83.49	89.41	83.21	80.45	94.92	56.89	92.24	72.66	87.6
60	72.75	78.83	90.18	94.61	89.11	88.89	97.91	63.53	96.49	81.54	94.33
70	82.82	84.2	94.61	97.09	95.1	93.93	99.29	69.71	98.44	87.45	97.86
80	87.89	89.77	97.02	98.63	97.01	97.25	99.72	75.15	99.31	91.8	99.22
90	91.7	93.01	98.32	99.48	98.85	98.72	99.93	79.18	99.72	94.83	99.78
100	94.76	95.45	99	99.72	99.5	99.47	99.97	82.75	99.91	96.87	99.91
120	97.94	98.03	99.79	99.95	99.94	99.93	100	88.35	99.98	98.81	99.99
140	99.31	99.31	99.96	99.99	100	100	100	92.07	99.99	99.63	100
160	99.79	99.7	99.99	99.99	100	100	100	94.46	100	99.91	100

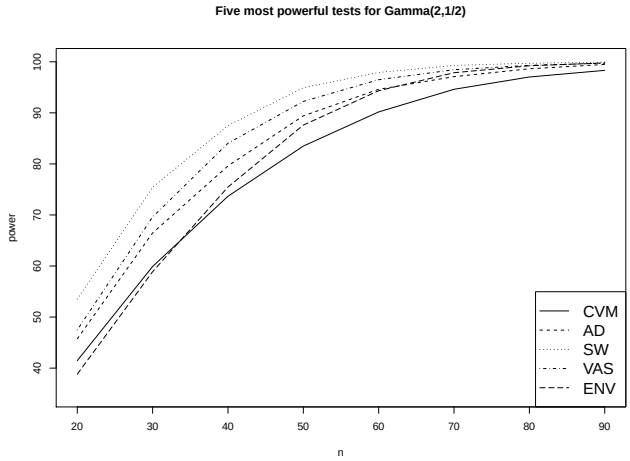

Figure 5–9: Better performing tests against $Gamma(2, 1/2)$

even having a relative big sample of 200 not all tests have achieved a power of more than 90%. Some of the same tests that performed better in the last two cases keep performing better in this case, including the envelope test.

$Lognormal(0, 1)$

Against a $lognormal(0, 1)$ all classical tests have a good performance, see table 5–12 and figure 5–11, achieving a power of 90% with a sample size of 30. Also with this sample size the envelope test has a better performance or comparable performance to all the tests.

$Lognormal(0, 1/2)$

With a smaller shape parameter the lognormal distribution starts taking a bell–form, resembling a normal distribution, and overall, the tests become less powerful, see table 5–13 and figure 5–12. Still the envelope test has a similar performance to the classical tests with higher power.

Table 5–11: Power against $Gamma(5, 1/2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	13.9	15.67	18.5	20.42	20.57	20.95	24.59	13.37	17.55	15.15	19.07
30	13.99	21.11	26.93	30.65	31.03	30.02	37.32	18.31	25.9	20.49	27.63
40	18.21	27.47	34.48	38.62	40.99	39.76	47.59	22.05	34.03	25.95	35.36
50	20.98	32.93	42.23	49.15	49	47.64	58.92	25.13	42.16	31.32	43.53
60	22.21	40.13	49.6	56.88	54.41	55.4	67.49	28.29	50.44	37.05	50.37
70	29.46	44.04	57.91	63.34	65.15	63.38	75.38	31.68	57.39	41.56	58.13
80	31.45	50.61	64.03	70.29	69.13	70.91	81.33	35.07	62.55	46.54	65.1
90	34.79	55.87	68.4	75.79	77.03	76.92	86.44	37.65	68.47	51.05	71.5
100	38.76	60.45	73.07	80.96	81.31	81.91	89.94	40.52	73.16	55.58	76.38
120	45.86	68.43	81.94	87.55	89.37	89.52	95.13	45.48	81.01	63.81	84.98
140	52.95	74.46	87.66	92.68	93.79	94.44	97.66	50.05	86.83	71.03	90.65
160	60.17	79.96	91.98	95.59	97.2	97.27	99.12	54.09	91.06	77.18	94.66
180	65.69	85.79	94.83	97.56	98.77	98.8	99.56	58.57	94.21	82.36	97.1
200	71.99	89.11	97.02	98.71	99.41	99.46	99.81	62.05	95.98	86.36	98.49

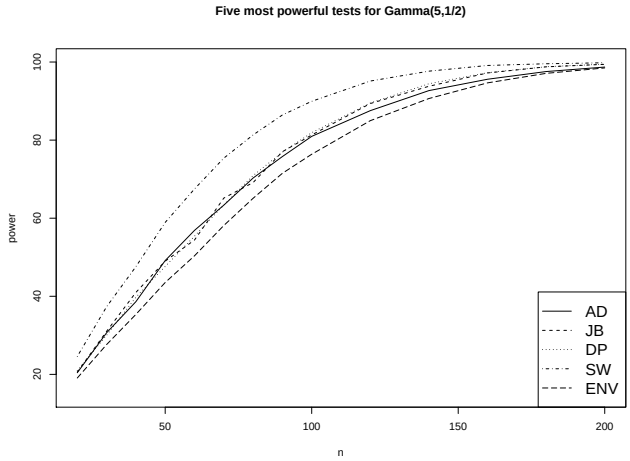

Figure 5–10: Better performing tests against $Gamma(5, 1/2)$

$Weibull(5, 2)$

In table 5–14 and figure 5–13 we see that when tested against a $weibull(5, 2)$, the tests show overall, a poorly performance, the two best tests are the Shapiro–Wilk test and the Anderson–Darling test but neither has achieved a power of 30% with a sample of 200.

Overall, the best performing test is Shapiro–Wilk test, followed by the Anderson–Darling test, but this test has the problem that sometimes for distributions with heavy tails it can not be calculated for big sample sizes. To these the Vasicek test follows but this has the problem that we need to find an adequate value for m to have its maximum performance. For most of the alternatives presented a value of m around 4 was used. This is shown with more detail in the next section. The envelope test has a similar performance to the Cramér–von Mises test, and most of the time it is better than the rest of the tests.

Table 5–12: Power against $Lognormal(0, 1)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	85.14	79.74	88.31	90.69	81.56	79.84	93.5	75.75	92.96	83.85	85.65
30	95.02	93.07	97.53	98.42	95.25	93.55	99.24	90.2	99.1	95.6	97.71
40	98.87	98.18	99.45	99.7	99.05	98.57	99.89	96.02	99.89	98.97	99.71
50	99.71	99.55	99.93	99.98	99.81	99.72	100	98.41	100	99.81	99.98
60	99.87	99.89	99.99	99.99	99.98	99.97	100	99.43	100	99.96	100
70	99.99	99.99	100	100	100	100	100	99.8	100	100	100
80	99.99	99.99	100	*	100	100	100	99.93	100	100	100
90	100	100	100	*	100	100	100	99.98	100	100	100

* The Anderson-Darling test can not be computed for many samples of this size.

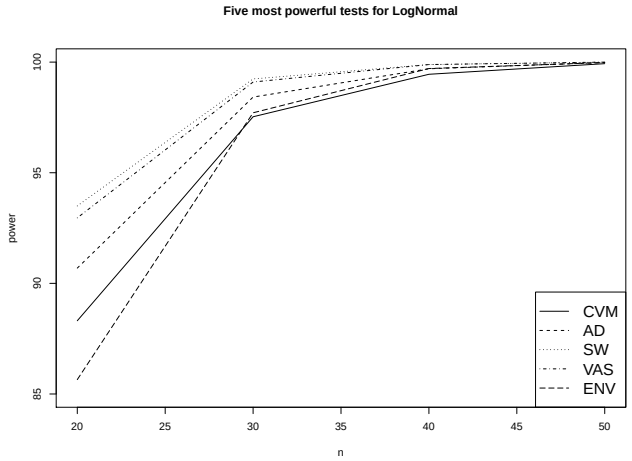

Figure 5–11: Better performing tests against $LogN(0, 1)$

Table 5–13: Power against $Lognormal(0, 1/2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	32.02	34.91	42.9	46.87	43.75	43.4	53.13	32.42	41.99	36.51	41.7
30	38.57	47.84	60.27	65.49	62.13	60.06	72.64	45.4	60.85	51.37	58.97
40	50.79	61.18	73.32	77.76	76.02	73.74	84.51	55.51	74.82	63.83	72.48
50	58.82	70.38	82.37	87.45	84.8	83.16	92.39	64.21	84.51	74.03	82.82
60	64.48	79.27	88.92	92.72	90.35	90.29	96.17	70.85	90.79	81.17	89.65
70	75.92	84.56	93.64	95.76	95.48	94.69	98.32	77.1	94.54	86.75	94.27
80	80.09	89.52	96.26	97.8	97.08	97.32	99.19	81.65	96.81	90.85	96.74
90	84.57	92.86	97.75	98.82	98.74	98.66	99.67	85.24	98.13	93.69	98.44
100	89.09	94.99	98.68	99.42	99.36	99.38	99.85	88.28	98.98	95.86	99.19
120	93.95	97.84	99.59	99.67	99.85	99.85	99.97	92.58	99.66	98.08	99.81
140	96.98	98.99	99.89	99.69	99.99	99.99	100	95.54	99.91	99.2	99.96
160	98.68	99.54	99.96	*	100	99.99	100	97.11	99.95	99.7	99.99

* The Anderson-Darling test can not be computed for many samples of this size.

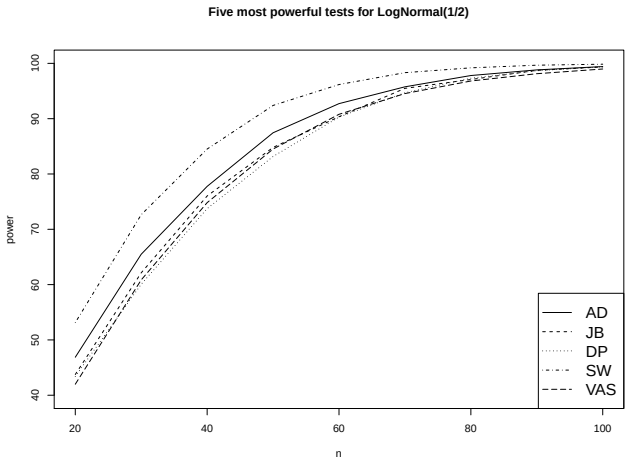

Figure 5–12: Better performing tests against $LogN(0, 1/2)$

Table 5–14: Power against $Weibull(5, 2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	7.46	6.08	5.84	5.89	4.68	4.97	6.24	4.63	6.73	5.15	5.4
30	5.83	6.35	6.88	7.12	5.27	5.15	7.2	5.1	7.29	5.51	5.86
40	6.5	6.93	7.36	7.27	5.69	5.6	7.43	4.61	8.25	5.38	5.69
50	6.3	7.37	7.92	8.88	5.98	5.85	8.58	4.84	8.95	5.57	6.04
60	5.56	8.66	9.02	9.79	5.64	6.29	9.45	4.79	9.79	5.59	6.37
70	7.43	8.62	10.43	10.25	7.08	6.91	10.61	5.33	10.37	5.86	6.81
80	6.95	10.19	11.52	11.65	6.56	7.68	11.46	5.26	11.25	6.01	7.63
90	6.95	10.41	11.62	12.37	7.75	8.1	12.56	5.3	11.93	5.9	7.84
100	7.46	11.15	12.11	13.6	8.29	9.14	13.75	5.13	11.88	6.31	8.78
120	7.66	12.48	14.39	15.23	9.97	10.56	16.16	5.21	12.55	6.28	9.49
140	8.12	13.41	16.12	17.91	11.17	12.78	19.32	5.07	13.9	6.81	10.81
160	8.69	14.41	17.87	19.54	13.68	14.28	22.74	5.23	14.79	7.12	11.96
180	8.66	17.06	20.26	21.94	16.22	16.88	23.28	5.32	16.01	7.46	13.55
200	9.92	18.09	22.55	24.82	18.59	19.5	27.03	5.13	17.51	7.72	14.84

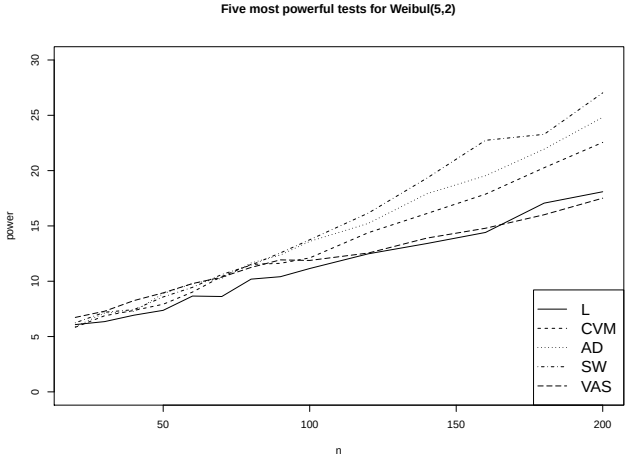

Figure 5–13: Better performing tests against $Weibull(5, 2)$

5.1.4 Distributions with support [0,1]

$Uniform$

Table 5–15: Power against $Uniform$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	11.12	10.13	14.02	16.68	0.22	0.2	20.78	8.21	47.21	13.6	8.82
30	10.92	13.76	23.43	30.12	0.12	0.06	39.19	21.98	76.16	20.05	12.41
40	15.66	20.17	32.81	42.4	0.15	0.03	56.31	37.1	91.27	28.51	17.89
50	20.09	25.49	43.42	58.95	1.24	0.07	74.82	54.99	97.19	37.13	26.22
60	22.76	33.24	53.93	70.16	3.41	0.11	86.26	70.08	99.34	46.14	37.75
70	29.3	38.46	64.65	78.62	23.68	0.31	93.77	82.38	99.84	54.96	52.27
80	34.66	46.62	72.56	86.5	30.48	0.71	97.12	88.8	99.96	64.08	67.12
90	42.31	53.42	78.64	91.76	59.31	1.29	99	93.04	99.99	70.68	80.41
100	45.7	59.42	83.65	95.22	73.6	2.63	99.6	95.79	100	77.37	88.96
120	57	70.28	92.02	98.33	92.73	9.46	99.97	98.5	100	87.6	97.62
140	68.48	78.55	96.12	99.58	98.33	37.93	100	99.36	100	93.45	99.64
160	79.47	85.06	98.35	99.87	99.79	67.88	100	99.78	100	96.97	99.97
180	82.71	91.49	99.3	99.97	99.98	90.64	100	99.95	100	98.72	100
200	90.17	94.63	99.72	100	100	97.84	100	99.98	100	99.43	100

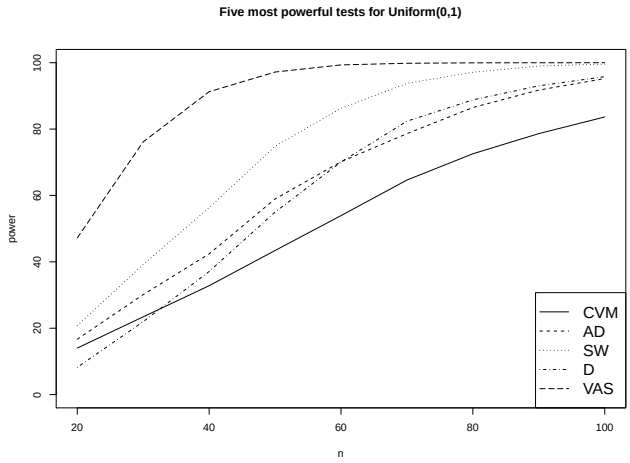

Figure 5–14: Better performing tests against $Uniform(0, 1)$

Against an $uniform(0, 1)$ distribution as alternative, most of the test have a low performance for small sample sizes, with only the Vasicek test having a power of more than 50% at a sample size of 30. The Vasicek test is followed by the Shapiro–Wilk test, see table 5–15 and figure 5–14.

$Beta(1/2, 1/2)$

Table 5–16: Power against $Beta(1/2, 1/2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	31.36	32.62	50.51	61.6	0.5	0.31	73.51	2.84	47.88	13.47	37.09
30	39.94	49.25	74.37	86.02	0.49	0.11	94.39	5.12	76.17	20.35	60.51
40	56.92	67.14	88.42	95.76	11.39	0.35	99.17	7.45	91.23	28.31	81.6
50	71.72	79.62	95.23	99.13	45.17	0.84	99.92	11.66	97.2	37.26	95.09
60	83.84	89.62	98.59	99.87	70.5	1.99	100	15.62	99.28	46.11	99.27
70	88.5	93.82	99.55	99.97	96.62	4.62	100	22.42	99.81	54.97	99.93
80	95.08	97.3	99.89	100	98.67	11.56	100	27.42	99.96	64.3	99.99
90	98.41	98.71	99.97	100	99.89	23.9	100	32.42	99.99	70.86	100
100	98.95	99.5	99.98	100	99.97	53.64	100	37.7	100	77.47	100
120	99.9	99.93	100	100	100	93.77	100	48.3	100	100	100
140	99.99	99.99	100	100	100	99.93	100	55.72	100	100	100

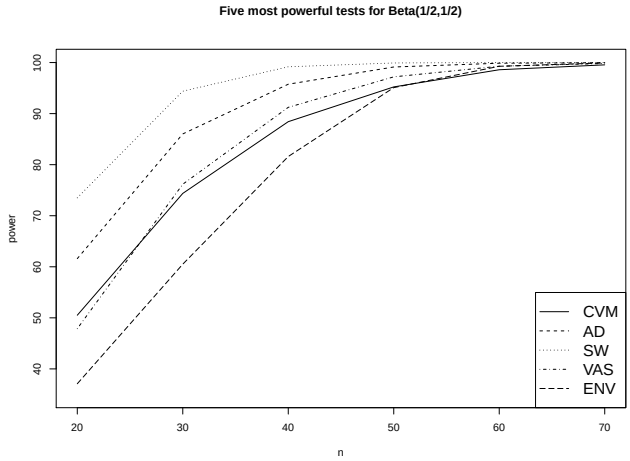

Figure 5–15: Better performing tests against $Beta(1/2, 1/2)$

Table 5–16 and figure 5–15 show that using a $beta(1/2, 1/2)$ as alternative we found that again the Shapiro–Wilk test has the higher power, this time followed by the Anderson–Darling test. After these two, the Cramér–von Mises test, the Vasicek test and the envelope test follow, and around a sample size of 50 all the previous tests have reached a power of at least 90%.

$Beta(1, 1/2)$

Table 5–17: Power against $Beta(1, 1/2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	53.11	43.52	57.16	64.44	11.24	9.08	73.14	10.65	59.83	22	49.62
30	71.17	62.01	79.36	86.85	18.5	11.09	93.21	12.02	85.3	34.41	82.12
40	86.71	78.89	90.96	95.69	35.81	22.52	98.64	12.65	96.32	46.94	96.2
50	95.09	88.64	96.55	99.06	58.04	39.16	99.86	13.57	99.15	59.14	99.49
60	98.52	95.21	98.93	99.8	74.36	60.2	99.98	14.14	99.89	70.96	99.95
70	99.27	97.93	99.7	99.96	95.11	78.2	100	15.3	99.97	80.66	99.99
80	99.82	99.24	99.9	99.99	97.52	90.6	100	15.33	100	87.32	100
90	99.97	99.75	99.98	100	99.67	95.83	100	16.19	100	91.4	100
100	99.98	99.92	99.99	100	99.93	98.65	100	17.21	100	94.56	100
120	100	100	100	100	100	99.91	100	18.72	100	100	100
140	100	100	100	100	100	100	100	20.09	100	100	100
160	100	100	100	100	100	100	100	20.84	100	100	100
180	100	100	100	100	100	100	100	22.27	100	100	100
200	100	100	100	100	100	100	100	23.71	100	100	100

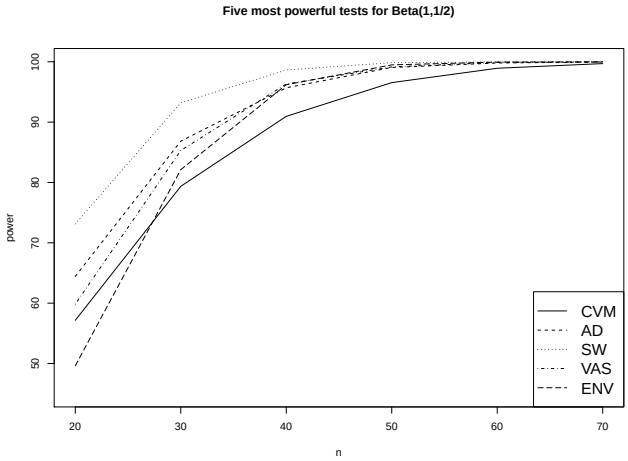

Figure 5–16: Better performing tests against $Beta(1, 1/2)$

Against a $beta(1, 1/2)$, we can see in table 5–17 and figure 5–16 that the performance of the tests varies by a large amount, for small sample sizes the Shapiro–Wilk reaches a power of 90% with a sample of 30 while both moment based tests and the D'Agostino test are below 20%. The envelope test has a good performance comparable to other classical tests, attaining a power of more 95% with a sample size of 40.

$Beta(5, 2)$

Using a $beta(5, 2)$ as alternative, the best performing tests are the Shapiro–Wilk and the the Vasicek test, the later being a little better up to sample sizes of 70. The envelope test has a similar performance to the Cramér–von Mises test and to the Anderson–Darling

Table 5–18: Power against $Beta(5, 2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	12.05	11.87	13.51	14.62	9.45	9.61	17.41	6.73	19.24	9.12	10.75
30	11.75	15.42	19.84	22.61	13.69	12.61	27.87	8.04	30.52	11.25	15.53
40	15.96	20.82	26.12	29.63	18.67	16.85	37.32	7.92	42.55	14.29	21.15
50	18.74	25.19	32.37	39.88	22.96	20.86	49.55	8.15	53.82	16.94	29.09
60	21.1	31.54	40.26	48.66	26.45	26.71	60.85	8.21	63.11	20.43	39.31
70	27.51	35.14	48.3	55.3	37.45	33.55	70.48	8.51	70.92	24.67	49.96
80	30.69	41.25	54.78	63.14	40.05	42.03	77.8	8.63	77.71	28.43	60.97
90	35.74	46.66	60.22	70.62	52.4	50.42	84.97	8.68	83.47	31.69	71.63
100	39.57	50.5	64.34	76.06	58.85	58.61	89.29	8.33	87.13	35.43	79.27
120	48.08	59.58	75.51	84.77	74.32	72.65	95.63	8.55	93.06	42.97	91.1
140	56.69	66.54	82.77	91.28	84.85	85.26	98.42	8.49	96.42	52.18	96.53
160	65.92	72.24	88.41	94.89	92.87	91.92	99.49	8.47	98.19	59.89	98.91
180	71.26	79.64	92.51	97.32	96.8	96.28	99.85	8.65	99.15	67.04	99.72
200	78.95	84.17	95.49	98.77	98.6	98.35	99.93	8.3	99.64	74.19	99.92

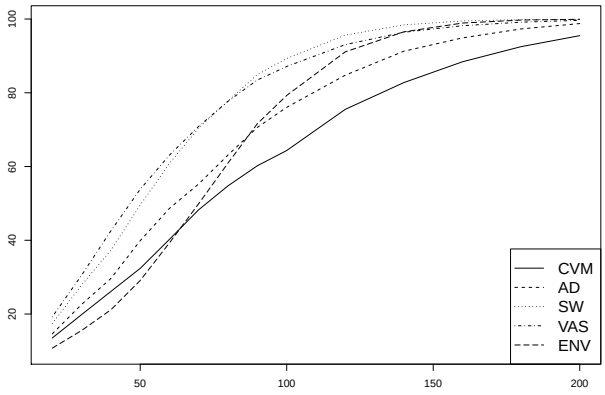

Figure 5–17: Better performing tests against $Beta(5, 2)$

test for small samples and by a sample size of 90 it performs better than those. See table 5–18 and figure 5–17.

$Triangular(0, 1, 1/2)$

We can see in table 5–19 and figure 5–18 that against a $triangular(0, 1, 1/2)$, all tests have a low performance, with the best being Vasicek test reaching a power of more 50% with a sample size of 140.

Table 5–19: Power against $Triangular(0, 1, 1/2)$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	6.43	4.31	4.09	3.86	0.83	0.89	3.46	4.5	8.57	5.26	2.74
30	4.51	3.73	4.01	4.13	0.44	0.41	3.67	6.7	10.99	5.63	2.32
40	5.36	4.21	4.19	4.14	0.23	0.2	3.58	8.17	14.2	5.59	2.09
50	5.27	4.14	4.34	5.1	0.11	0.09	4.54	10.97	17.96	6.04	2.16
60	4.55	4.68	4.7	5.51	0.08	0.09	5.32	13.43	23.13	6.12	2.15
70	5.8	4.19	5.22	5.52	0.08	0.06	6.61	17.62	27.43	6.28	2.22
80	5.53	4.64	5.52	6.27	0.06	0.05	7.14	20.57	31.92	6.63	2.33
90	5.66	4.97	5.73	7.03	0.2	0.08	8.99	23.01	37.65	6.68	2.67
100	5.78	4.92	5.6	7.71	0.24	0.09	10.18	26.45	39.58	6.89	2.81
120	6.05	5.16	6.69	8.53	0.65	0.13	13.95	32.85	48.01	7.42	3.39
140	6.38	5.2	6.95	10.19	1.3	0.18	19.36	37.19	54.77	8.01	4.14
160	6.82	5.19	7.73	11.36	3.38	0.31	26	44.25	60.91	8.49	5.18
180	6.63	6.19	9.07	13.53	6.27	0.45	28.39	49.05	66.12	9.28	6.92
200	7.51	6.39	10.02	16.13	9.45	0.81	36.82	53.74	71.49	10.09	8.93

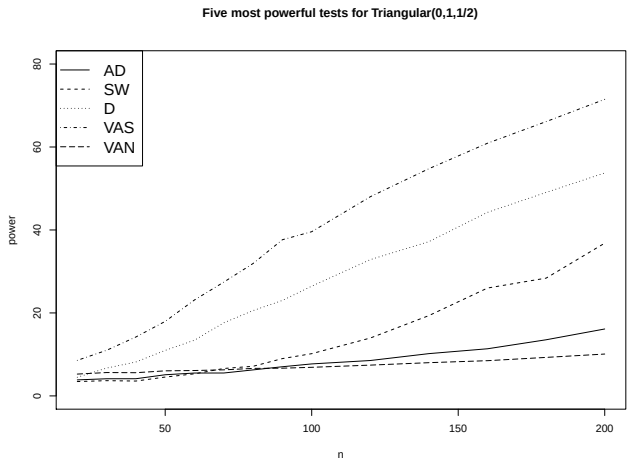

Figure 5–18: Better performing tests against $Triangular(0, 1/2, 1)$

Overall, against this type of alternatives the best tests are the Shapiro–Wilk test and the Vasicek test. The envelope test has a performance comparable to the Cramér–von Mises test. Tests that are better to avoid using against these types of alternatives, specially with small samples, are the moment based tests and the D'Agostino test.

5.1.5 Standard normal with outlier

Outlier $3s$

Table 5–20: Power against SN with outlier $3s$

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	7.98	6.76	6.91	6.77	3.77	3.93	5.9	3.15	3.3	10.2	4.92
30	5.86	6.24	7.17	7.44	5.86	5.38	6.84	4.28	3.55	10.97	6.85
40	6.33	6.39	6.6	6.45	6.67	6.17	6.07	4.69	3.91	10.93	7.6
50	5.67	6.02	6.26	6.95	6.81	6.39	6.35	4.9	4.08	10.27	8.29
60	4.66	6.06	6.14	6.65	6.27	6.57	6.24	5.02	4.12	10.32	8.29
70	6.13	5.68	6.47	6.12	7.45	6.7	6.18	5.1	4.27	10.25	8.23
80	5.37	5.85	6.55	6.51	6.61	7.04	6.03	5.02	4.34	9.76	8.13
90	5.25	5.86	5.88	6.22	7.16	6.8	6.03	4.99	4.42	9.45	7.61
100	5.51	6.07	5.79	6.52	7.05	6.99	5.89	5.13	4.5	9.3	7.11
120	5.29	5.54	5.91	5.85	6.76	6.51	5.78	4.98	4.49	8.88	6.63
140	4.98	5.48	5.74	6.09	6.77	6.86	6.05	5.1	4.41	8.21	6.45
160	5.1	4.81	5.58	5.67	7.02	6.69	6.2	4.79	4.6	8.38	6.73
180	4.83	5.48	5.65	5.53	7.12	6.71	5.33	4.68	4.71	7.76	6.67
200	5.42	5.49	6.12	6.18	7.09	6.79	5.72	4.64	4.65	7.27	6.98

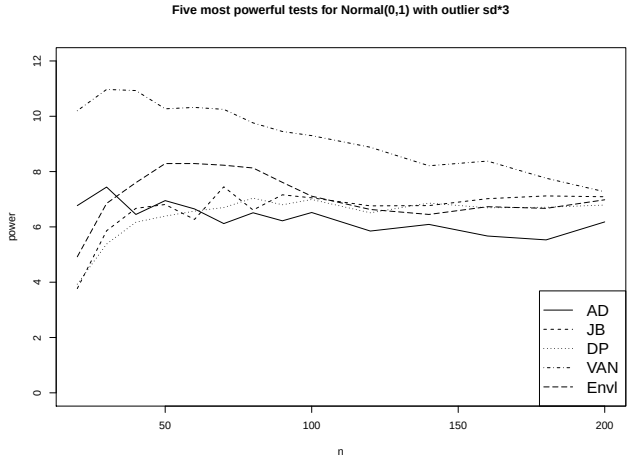

Figure 5–19: Better performing tests against $Normal + Outlier(3s)$

In table 5–20 and figure 5–19 we see that an outlier of three times the standard deviation of the sample most the tests do not detect non-normality, only the Van Es test reaches a power of 10%.

Outlier $4s$

With an outlier of 4 times the sample standard deviation we see in table 5–21 and figure 5–20 that the best performing tests are the Jarque–Bera test, the D'Agostino–Pearson test and for sample sizes of less than 120, the envelope test has a comparable performance to these two.

Table 5–21: Power against SN with outlier 4s

n	P	L	CVM	AD	JB	DP	SW	D	VAS	VAN	ENV
20	12.12	13.86	16.25	18.55	84.12	87.41	34.91	29.67	6.42	37.95	57.02
30	8.66	11.66	15.37	18.59	93.03	91.77	41.52	38.09	6.87	36.44	79.74
40	8.82	11.2	13.73	15.78	94.44	92.8	39.27	39.09	6.83	37.49	84.86
50	7.6	9.77	12.12	15.33	93.29	90.37	39.69	37.22	6.95	37.96	86.85
60	6.05	9.82	11.4	13.96	89.75	87.74	37.63	35.46	6.73	37.13	85.9
70	7.81	8.39	11.07	12.06	89.7	84.13	36.37	34.03	6.82	36.46	84.52
80	6.63	9.05	11.24	12.52	83.74	81.45	34	32.56	6.56	35.2	82.69
90	6.35	8.35	9.91	11.57	82.55	76.84	33.11	30.9	6.45	33.98	79.77
100	6.25	8.26	9.24	11.25	77.57	72.71	30.84	29.29	6.58	33.51	74.55
120	5.89	7.52	9.13	9.96	71.21	64.64	29.12	27.33	6.19	30.6	63.25
140	5.55	7.02	8.47	9.58	63.87	59.83	28.62	25.28	5.83	28.04	48.78
160	5.68	6.35	7.9	8.74	59.64	53.49	27.81	22.61	6.09	26.81	36.55
180	5.46	6.92	7.94	8.54	55.66	50	23.37	21.38	6.06	25.43	26.71
200	5.67	6.8	7.87	8.69	50.99	46.22	23.54	20.12	5.74	24.09	20.2

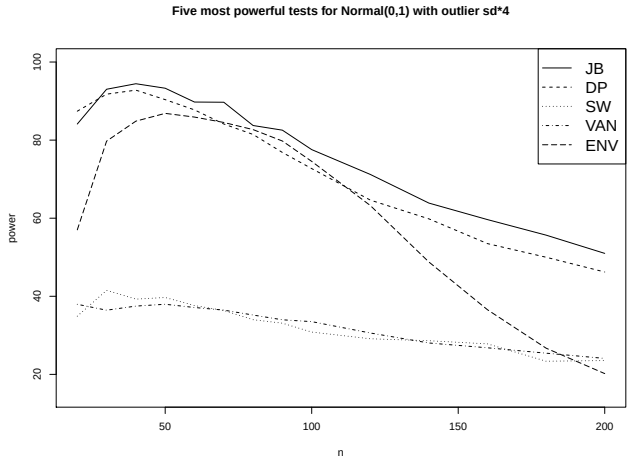

Figure 5–20: Better performing tests against $Normal + Outlier(4s)$

In general, it can be seen that not one test is conclusively better than any other, nevertheless from the E.D.F. group the Pearson test and the Lilliefors test present overall, the lowest performance. The envelope test, although not as powerful as other classical tests, like the Shapiro–Wilk test, shows in most cases a comparable performance to other classical tests.

5.2 Choosing the best m

When it comes to the test based on entropy, we need to choose a value m for the test, the power of the test will change for each value of m. The possible values of m depend on the sample size n, we can choose from 1 to $n/2$, so with a large n there are many values of m to be considered. For this study we considered only values of m up to 10 for each sample size, for each one the corresponding power was calculated for different sample sizes, giving us a reasonable value for m for each sample size.

5.2.1 Symmetric distributions with infinite support

Figures 5–21 through 5–30 show that for distributions with infinite support the power for Vasicek will be higher for values of m between 3 and 4. For the Van Es test the power of test increases as the value of m increases as well.

$t(1)$

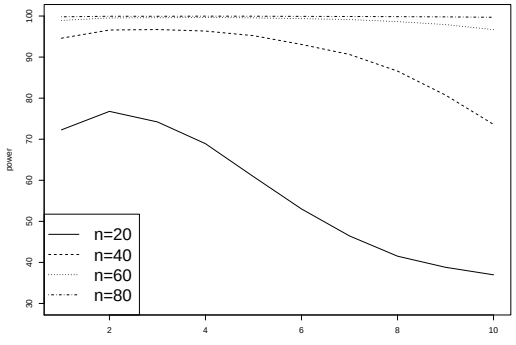

Figure 5–21: Power by m, Vasicek test, $t(1)$

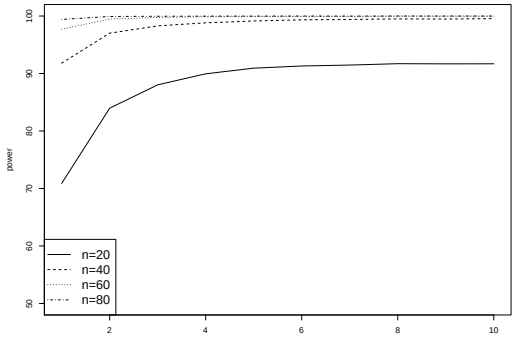

Figure 5–22: Power by m, Van Es test, $t(1)$

$t(2)$

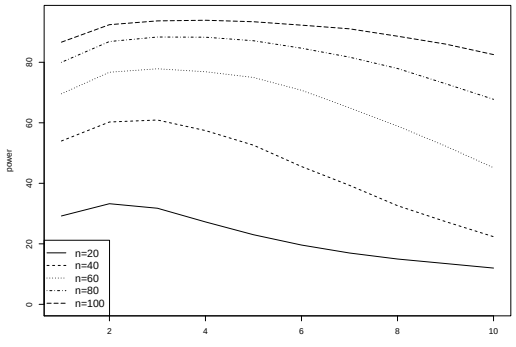

Figure 5–23: Power by m, Vasicek test, $t(2)$

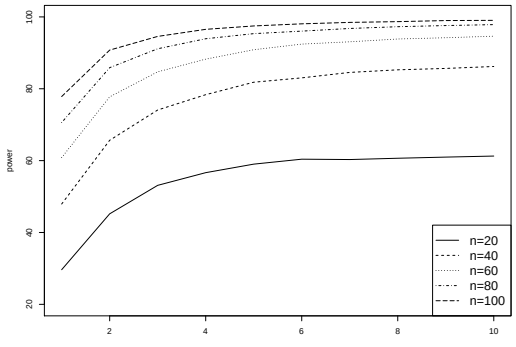

Figure 5–24: Power by m, Van Es test, $t(2)$

$t(5)$

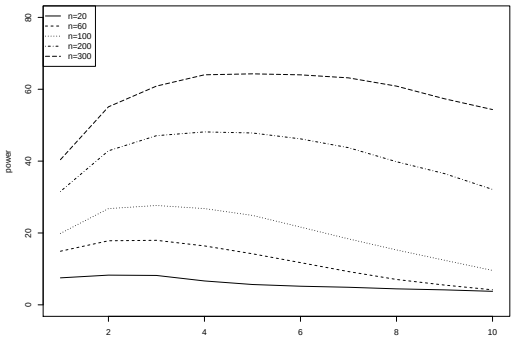

Figure 5–25: Power by m, Vasicek test, $t(5)$

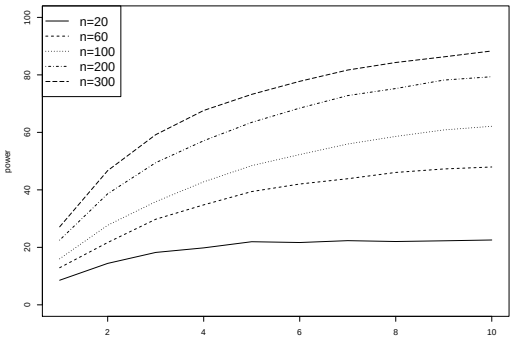

Figure 5–26: Power by m, Van Es test, $t(5)$

$t(10)$

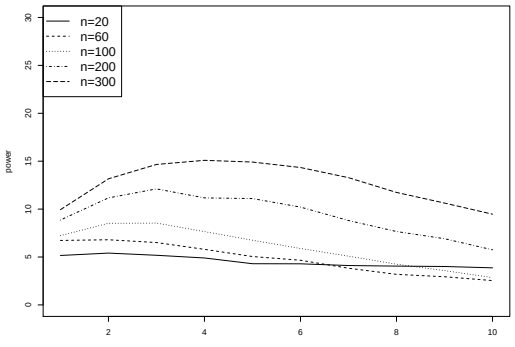

Figure 5–27: Power by m, Vasicek test, $t(10)$

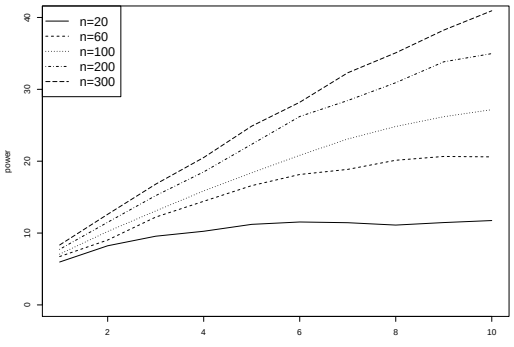

Figure 5–28: Power by m, Van Es test, $t(10)$

$Dexp(1)$

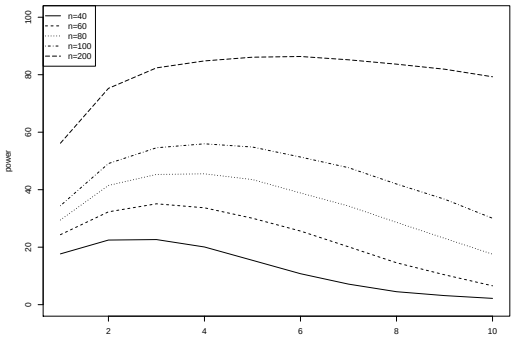

Figure 5–29: Power by m, Vasicek test, $Dexp(1)$

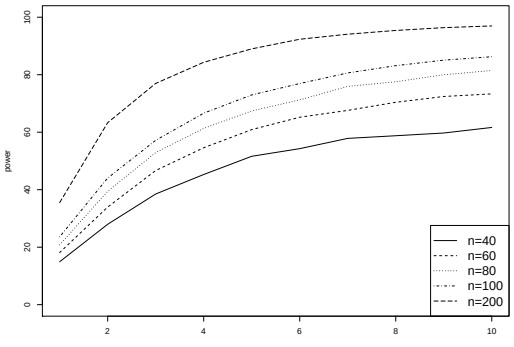

Figure 5–30: Power by m, Van Es test, $Dexp(1)$

5.2.2 Asymmetric distributions with infinite support

For asymmetric distributions with infinite support, figures 5–31 through 5–34 show that both tests have the behavior that they had against the symmetric alternatives, a value of m between 3 and 4 for the Vasicek test and a high value of m for the Van Es test will yield to maximum power.

$Gumbel(0, 1)$

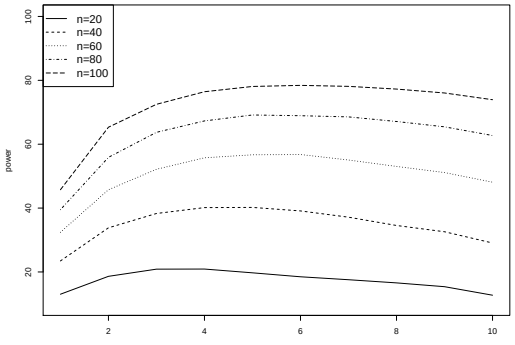

Figure 5–31: Power by m, Vasicek test, $Gumbel(0, 1)$

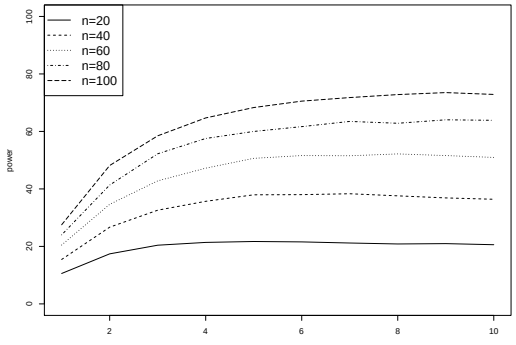

Figure 5–32: Power by m, Van Es test, $Gumbel(0, 1)$

$Gumbel(1, 3)$

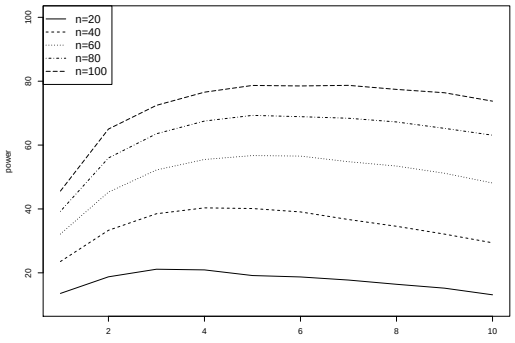

Figure 5–33: Power by m, Vasicek test, $Gumbel(0, 1)$

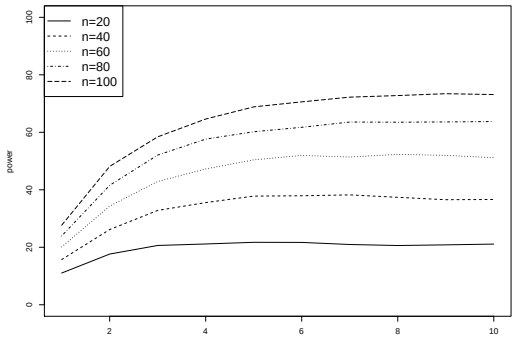

Figure 5–34: Power by m, Van Es test, $Gumbel(0, 1)$

5.2.3 Distributions with semi–infinite support

Against an exponential alternative, figures 5–35 and 5–36 show that the maximum power for both tests is attained for a value of m between 3 and 4.

$Exp(1)$

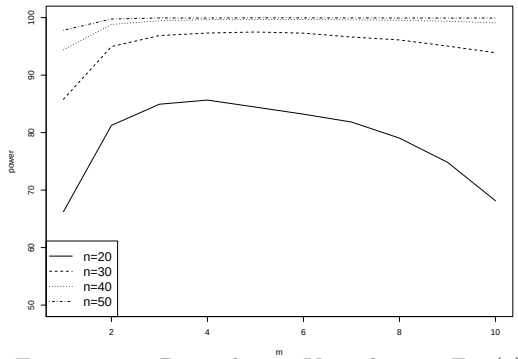

Figure 5–35: Power by m, Vasicek test, $Exp(1)$

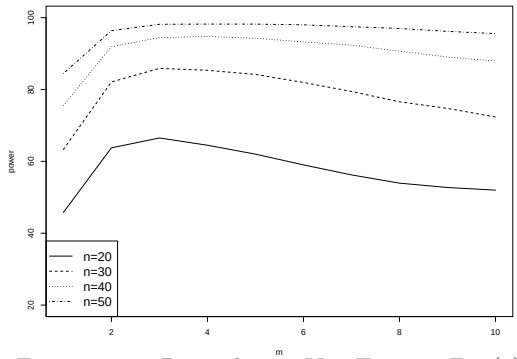

Figure 5–36: Power by m, Van Es test, $Exp(1)$

Against the used Gamma alternatives, figures 5–37 thorugh 5–40 show that the best value of m lays between 4 and 6.

$Gamma(2, 1/2)$

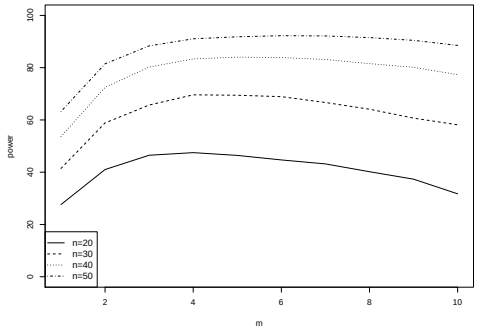

Figure 5–37: Power by m, Vasicek test, $Gamma(2, 1/2)$

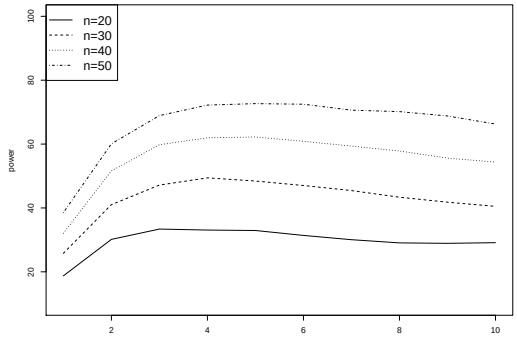

Figure 5–38: Power by m, Van Es test, $Gamma(2, 1/2)$

In figures 5–41 through 5–44 it can be seen that against a Lognormal distribution the best value of m for both tests is a low value between 3 and 4.

Against the weibull distribution used, figures 5–45 and 5–45 show that for the Vasicek test, the power rises with the value of m; the Van Es test presents a very small power, but it tends to fall as the value of m decreases.

$Gamma(5, 1/2)$

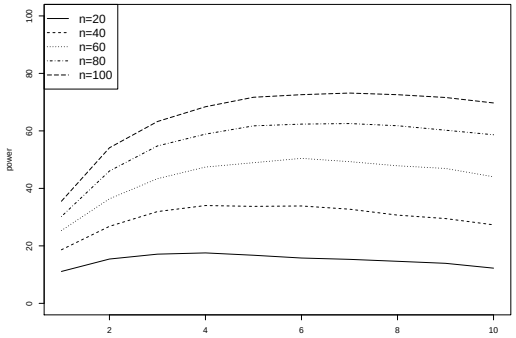

Figure 5–39: Power by m, Vasicek test, $Gamma(5, 1/2)$

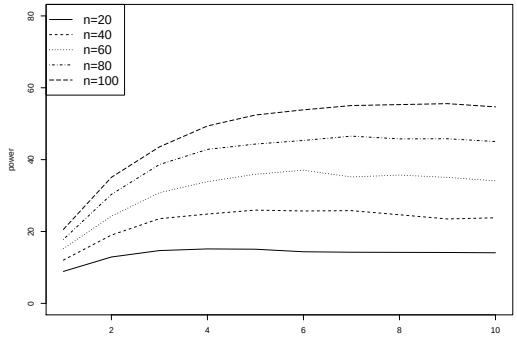

Figure 5–40: Power by m, Van Es test, $Gamma(5, 1/2)$

$Lognormal(0, 1)$

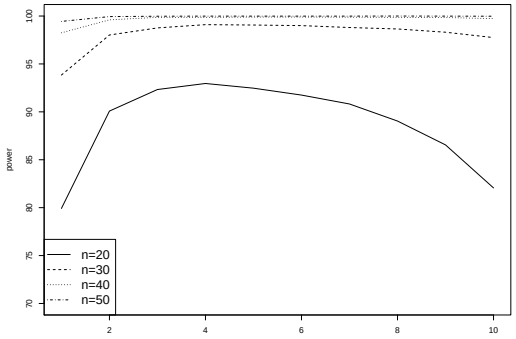

Figure 5–41: Power by m, Vasicek test, $Lognormal(0, 1)$

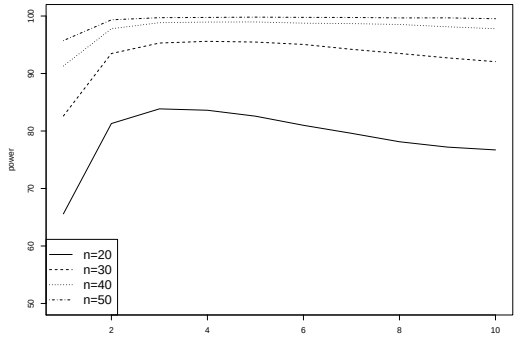

Figure 5–42: Power by m, Van Es test, $Lognormal(0, 1)$

$Lognormal(0, 1/2)$

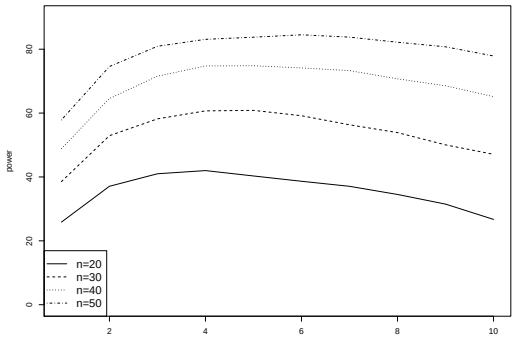

Figure 5–43: Power by m, Vasicek test, $Lognormal(0, 1/2)$

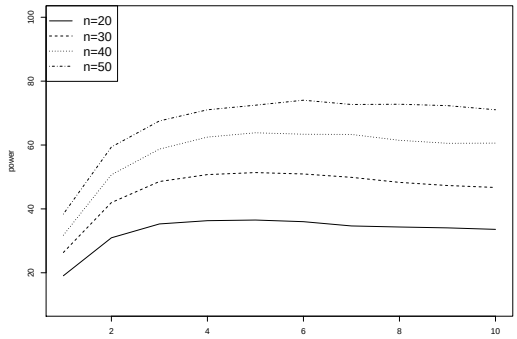

Figure 5–44: Power by m, Van Es test, $Lognormal(0, 1/2)$

$Weibull(5, 2)$

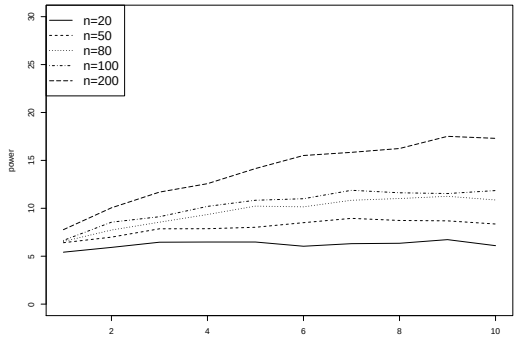

Figure 5–45: Power by m, Vasicek test, $Weibull(5, 2)$

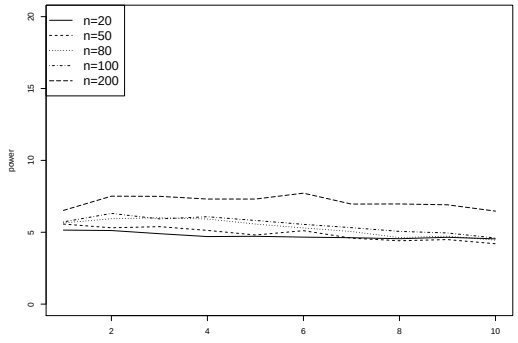

Figure 5–46: Power by m, Van Es test, $Weibull(5, 2)$

According to the alternatives presented, the best value of m for this alternatives will be a low value, usually a value between 3 and 4, against gamma distributions, this value tends to slightly increase as the shape parameter increases.

5.2.4 Distributions with support [0,1]

$Uniform(0, 1)$

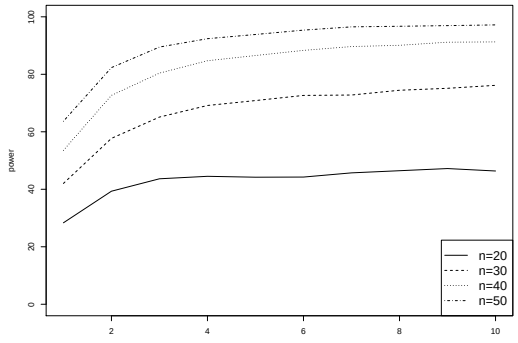

Figure 5–47: Power by m, Vasicek test, $Uniform(0, 1)$

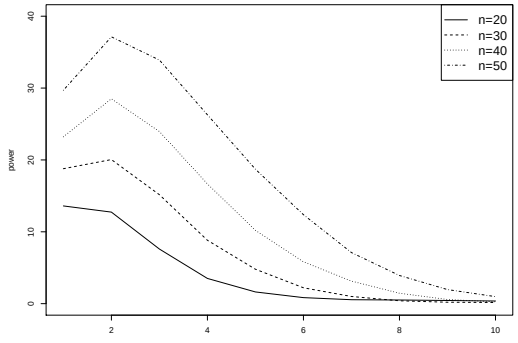

Figure 5–48: Power by m, Van Es test, $Uniform(0, 1)$

$Beta(1/2, 1/2)$

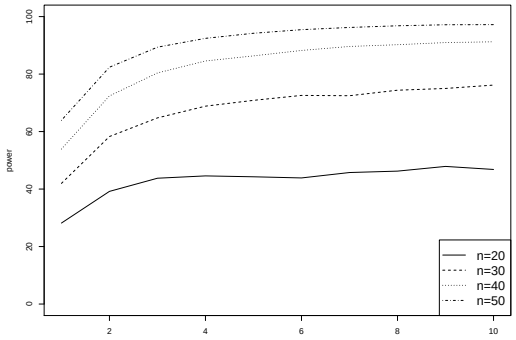

Figure 5–49: Power by m, Vasicek test, $Beta(1/2, 1/2)$

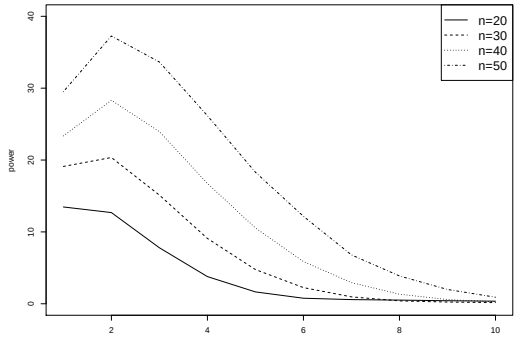

Figure 5–50: Power by m, Van Es test, $Beta(1/2, 1/2)$

$Beta(1, 1/2)$

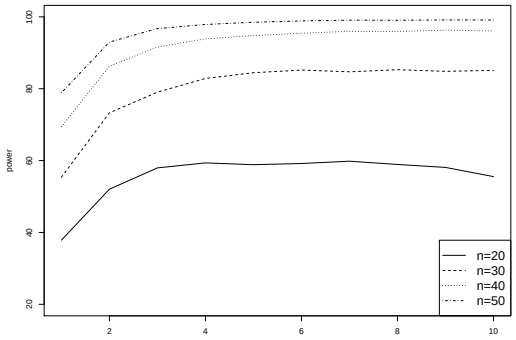

Figure 5–51: Power by m, Vasicek test, $Beta(1, 1/2)$

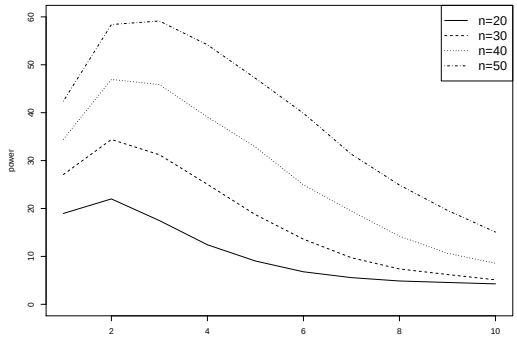

Figure 5–52: Power by m, Van Es test, $Beta(1, 1/2)$

$Beta(5, 2)$

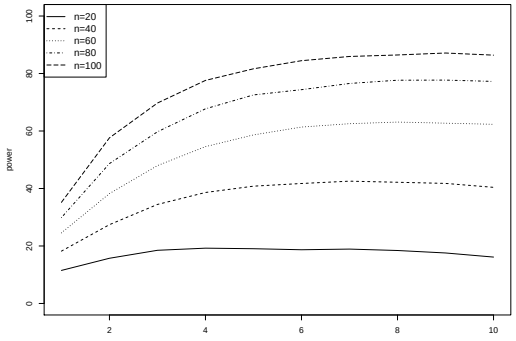

Figure 5–53: Power by m, Vasicek test, $Beta(5, 2)$

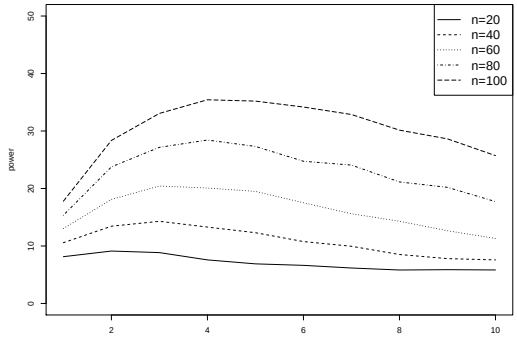

Figure 5–54: Power by m, Van Es test, $Beta(5, 2)$

$Triangular(0, 1, 1/2)$

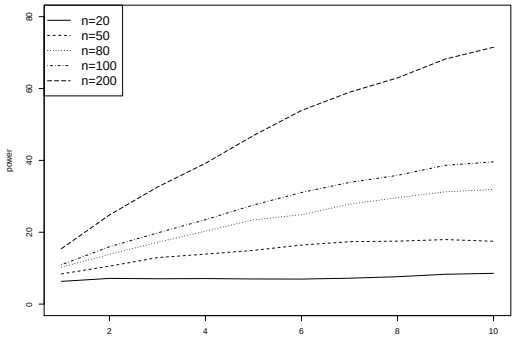

Figure 5–55: Power by m, Vasicek test, $Triangular(0, 1, 1/2)$

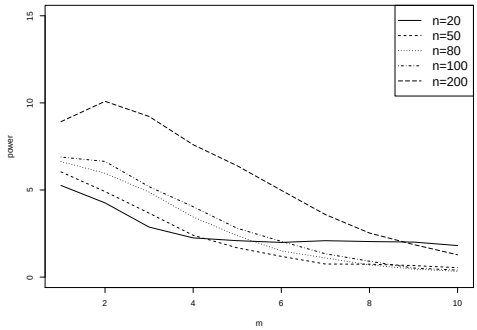

Figure 5–56: Power by m, Van Es test, $Triangular(0, 1, 1/2)$

For this type of distribution, from figure 5–47 to figure 5–56 we can see that the tests show the same behavior through all the alternatives. Vasicek test's power increases with the value of m and Van-Es test's power reached its maximum for values of m between 2 and 4.

5.2.5 Standard normal with outlier

Standard normal with outlier $3s$

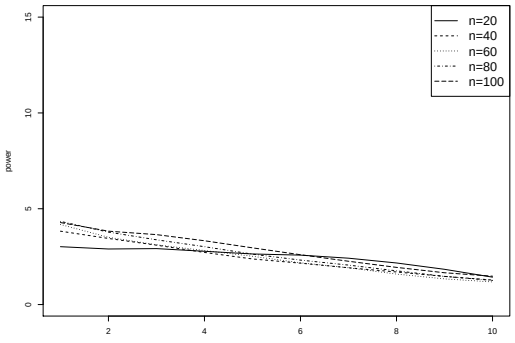

Figure 5–57: Power by m, Vasicek test, $SN + outlier(3s)$

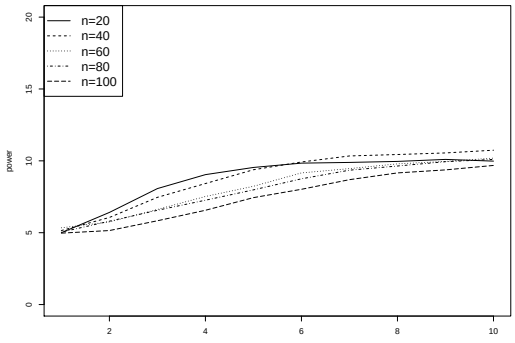

Figure 5–58: Power by m, Van Es test, $SN + outlier(3s)$

Standard normal with outlier $4s$

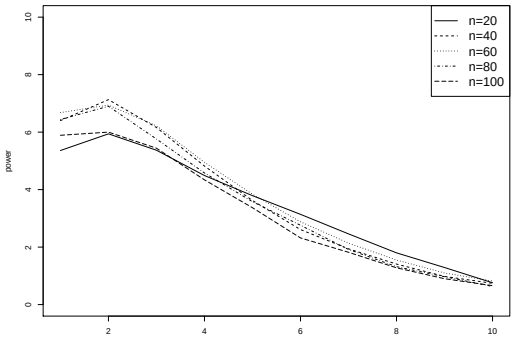

Figure 5–59: Power by m, Vasicek test, $SN + outlier(4s)$

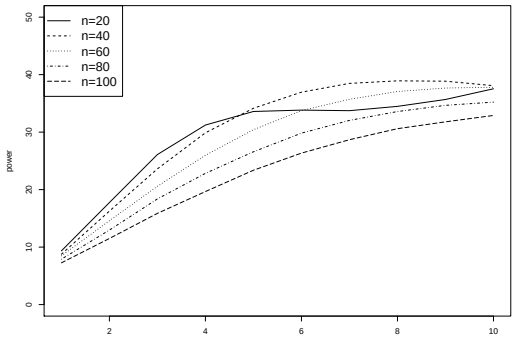

Figure 5–60: Power by m, Van Es test, $SN + outlier(4s)$

With outliers present, Vasicek test's power is higher for low values of m and Van Es test's power increases with as value of m increases.

5.3 Envelope test calculations

Using the algorithm described in section 4.1 implemented in C++ with an interface for R, the calculation of each value of p using simulation can take more than a couple of seconds even for sample sizes that are not really big. For example, to calculate the corresponding p for $n = 50$ using 10000 repetitions takes about 3 seconds on a PC running at 1600 MHz. Therefore we are interested in obtained a function for values of p. For different sample sizes from 6 to 3000 and using $\alpha = 0.05$, values for p were calculated.

Figure 5–61 shows the resulting points and suggests that a model of the form $p = an^b$ is adequate. Applying a logarithmic transformation we obtained a linear model of the form $\log p = \log a + b \log n$, this returns the values $\log a = -0.6095$ and $b = -0.6454$ figure 5–62 shows the transformed data with the regression line.

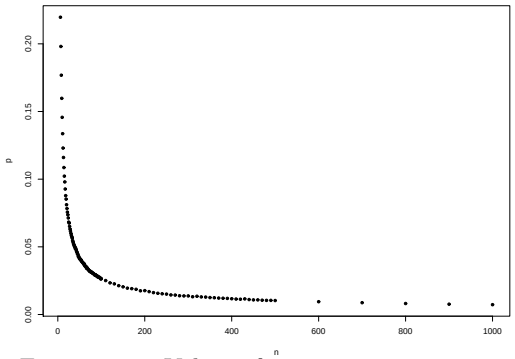

Figure 5–61: Values of p against n up to 1000

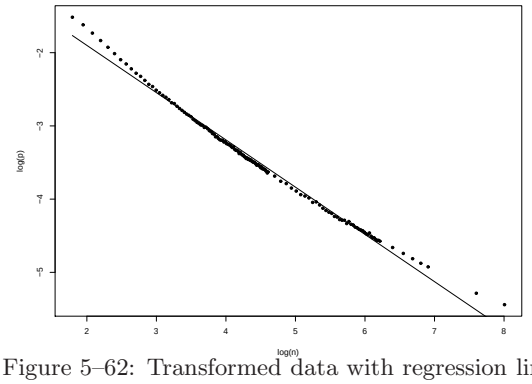

Figure 5–62: Transformed data with regression line

As can be seen, the regression line does not have a good fit for small or big values, to help this situation we are going to use two linear regressions instead, one for values of n up to 75 and other for the rest. This approach yields to the following regression coefficients: $\log a = -0.219938$ and $b = -0.754644$ for the first regression and $\log a = -1.000823$ and $b = -0.573086$ for the second.

The transformed data and the new regression lines are show in figure 5–63 and figure 5–64 shows the original data with the fitting curves.

Using this fitting, we obtain a function to quickly approximate p at a level $\alpha = 0.05$ given by:

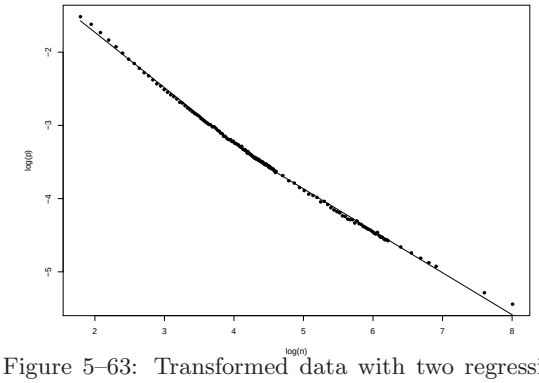

Figure 5–63: Transformed data with two regression lines

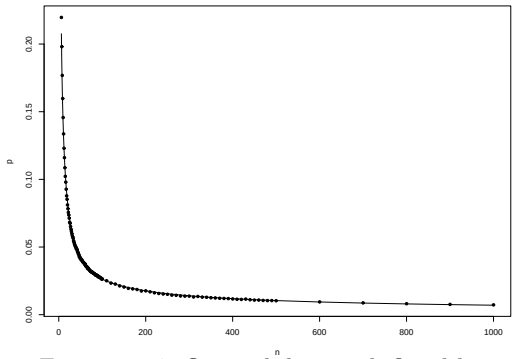

Figure 5–64: Original data with fitted line

$$p(x) = \begin{cases} 0.8025685\, n^{-0.7546436} & 6 \leq n \leq 75 \\ 0.3675769\, n^{-0.573086} & 76 \leq n \leq 3000 \end{cases}$$

CHAPTER 6
CONCLUSIONS

6.1 Normality tests power study

The task of efficiently determining if a data set was sampled from a normal population is not a trivial one. We have seen that there is not a superior normality test, instead, most of the tests' power varies with the alternative distribution. However there are still some tests that regularly do better than others. Specifically we see a poor performance from the older tests, the Pearson χ^2 test and the Lillifors test.

Especially for sample sizes smaller than 50 the following recommendations can be made: For the first group of alternative distributions, the symmetric distributions with infinite support, the Van Es test using a large value for m has a good performance against t distribution with a low d.f. value and the double exponential distribution. It's power is only exceeded by the Jarque–Bera test and the D'Agostino–Pearson test's power. Both of these tests belong to the moment based tests.

In the asymmetric distributions with infinite support category the Shapiro–Wilk test shows a higher power than any other test. Followed by both moment based tests.

For the next group, the asymmetric distributions with semi–infinite support, the Shapiro–Wilk test has a very good performance. It is only exceeded by the Vasicek test against an exponential alternative. Even in this case the recommended test would still be the Shapiro–Wilk because the difference between is less than 2% and finding the right m value for the Vasicek test is not straightforward.

In the group of distributions with support $[0, 1]$ we have again the Shapiro-Wilk test and the Vasicek test as the best performing tests. Against an uniform distribution and a triangular distribution the Vasicek test using $m = 10$ overwhelms all other tests for small sample sizes. The Shapiro–Wilk test has a higher power for two of the Beta distributions,

$\alpha = .5$, $\beta = .5$ and $\alpha = 1$, $\beta = .5$. For the third Beta distribution studied, the Vasicek test's power is about 2% to 4% higher than the Shapiro–Wilk test's power.

In the case of data with outliers, the Van Es test has a power of about 10% and this is the highest for the case $3s$. For the case $4s$ both moment based tests have the best performance. These are followed by the envelope test's performance.

6.2 Envelope test

With the envelope test we constructed a procedure which allows us to eliminate the subjectivity in a widely used graphical test for normality. This method is comparable, and in some cases better, to other classical normality tests.

Although we do not know an analytical form of the curve of the confidence band we find an approximation to it for the most commonly used confidence level and a Monte Carlo algorithm to find the band for other confidence levels.

Also we showed that other similar procedures do not take account of a total coverage confidence. In the case of Minitab, the empirical evidence shows that its confidence band does not attain the specified level.

APPENDICES

APPENDIX A
PROPOSITIONS

Proposition 1. *The standardization of the sample observations is invariant to location–scale transformations.*

Proof. Let $X_1, \ldots, X_n$ and i.i.d. sample from a random variable X, the standardization of an observation is defined as $\hat{z}_i = \dfrac{x_i - \bar{x}}{s}$. Let the location–scale transformation of X be $Y = \alpha X + \beta$, where α and β are positive constants. The standardization of an observation of Y is

$$\hat{z_i}^* = \frac{y_i - \bar{y}}{s_y} = \frac{\alpha x_i + \beta - (\alpha\bar{x} + \beta)}{\alpha s} = \frac{\alpha(x_i - \bar{x})}{\alpha s} = \frac{x_i - \bar{x}}{s} = \hat{z}_i$$

Therefore, the standardized transformed observations are equal to the original standardized observations. □

Proposition 2. *The sample skewness and kurtosis of a sample are invariant to location–scale transformations.*

Proof. Let $X_1, \ldots, X_n$ and i.i.d. sample from a random variable X with $\sqrt{b_1}$ and b_2 as its sample Kurtosis. Let the location–scale transformation of X be $Y = \alpha X + \beta$, where α and β are positive constants.

The new observations are calculated as $y_i = \alpha x_i + \beta$

The sample skewness of Y is:

$$\sqrt{b_{1y}} = \frac{\sqrt{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(\alpha x_i + \beta - (\alpha\bar{x} + \beta))^3}{\left(\sum_{i=1}^{n}(\alpha x_i + \beta - (\alpha\bar{x} + \beta))^2\right)^{3/2}}$$

$$
\begin{aligned}
&= \frac{\sqrt{n}\sum_{i=1}^{n}(\alpha x_i - \alpha\bar{x})^3}{\left(\sum_{i=1}^{n}(\alpha x_i - \alpha\bar{x})^2\right)^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}\alpha^3(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}\alpha^2(x_i - \bar{x})^2\right)^{3/2}} \\
&= \frac{\alpha^3\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\alpha^3\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{3/2}} \\
&= \sqrt{b_1}
\end{aligned}
$$

The sample kurtosis of Y is:

$$
\begin{aligned}
b_{2y} &= \frac{n\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^2} = \frac{n\sum_{i=1}^{n}(\alpha x_i + \beta - (\alpha\bar{x} + \beta))^4}{\left(\sum_{i=1}^{n}(\alpha x_i + \beta - (\alpha\bar{x} + \beta))^2\right)^2} \\
&= \frac{n\sum_{i=1}^{n}(\alpha x_i - \alpha\bar{x})^4}{\left(\sum_{i=1}^{n}(\alpha x_i - \alpha\bar{x})^2\right)^2} = \frac{n\sum_{i=1}^{n}\alpha^4(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}\alpha^2(x_i - \bar{x})^2\right)^2} \\
&= \frac{\alpha^4 n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\alpha^4\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \\
&= b_2
\end{aligned}
$$

□

Proposition 3. *Let $X_1, \ldots, X_n$ an i.i.d. sample from a standard normal distribution, then $\Phi(X_{(k)})$ is distributed $Beta(k, n-k+1)$ for $k = 1, \ldots, n$*

Proof. Let $Y = \Phi(X_{(k)})$ and ϕ denote the p.d.f. of a standard normal distribution.

$$
\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P\left(\Phi(X_{(k)}) \leq y\right) \\
&= P\left(X_{(k)} \leq \Phi^{-1}(y)\right), \qquad \text{since } \Phi^{-1}(x) \text{ is increasing}
\end{aligned}
$$

$$= \int_{-\infty}^{\Phi^{-1}(y)} \frac{n!}{(k-1)!(n-k)!}\, \phi(t)\ (\Phi(t))^{k-1}\, (1-\Phi(t))^{n-k}\, dt$$

Now we derivate to find $f_Y(y)$:

$$\begin{aligned} f_Y(y) &= \frac{dF_y(y)}{dy} \\ &= \frac{n!}{(k-1)!(n-k)!}\, \phi(\Phi^{-1}(y))\ \left(\Phi(\Phi^{-1}(y))\right)^{k-1}\ \left(1-\Phi(\Phi^{-1}(y))\right)^{n-k}\ \frac{d\Phi^{-1}(y)}{dy} \end{aligned}$$

Rewriting $n! = \Gamma(n+1)$, $(k-1)! = \Gamma(k)$ and $(n-k)! = \Gamma(n-k+1)$ and using the derivative of the inverse function $\left[f^{-1}\right]'(a) = \frac{1}{f'[f^{-1}(a)]}$ we have that

$$\begin{aligned} f_Y(y) &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\, \phi(\Phi^{-1}(y))\, y^{k-1}\, (1-y)^{n-k}\, \frac{1}{\phi(\Phi^{-1}(y))} \\ &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\, y^{k-1}\, (1-y)^{n-k+1-1} \end{aligned}$$

Which is the p.d.f. of a $Beta(k, n-k+1)$. □

REFERENCE LIST

[1] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–174, 1900.

[2] H. C. Thode. *Testing for Normality*. Marcel Dekker, third edition, 2002.

[3] A. N. Kolmogorov. Sulla determinazione empirica di une legge di distribuzione. *Giornale dell'Intituto Italiano degli Attuari*, 4:83–91, 1933.

[4] R. B. D'Agostino and M. A. Stephens. *Goodness of Fit Techniques*. Marcel Dekker, first edition, 1986.

[5] H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.*, 62:399–402, 1967.

[6] R. von Mises. *Wahrscheinlichkeitsrechnung und ihre anwendung in der statistik and theoretischen Physik*. Leipzig: Deuticke, 1931.

[7] H. Cramér. On the composition of elementary errors. *Skandinavisk Aktaurietidskrift*, 11:141–180, 1928.

[8] T. W.Anderson and D. A. Darling. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *The Annals of Mathematical Statistics*, pages 193–212, 1952.

[9] A. K. Bera and C. M. Jarque. A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2):163–172, 1987.

[10] G. Poitras. More on the correct use of omnibus tests for normality. *Economics Letters*, 90:304–309, 2006.

[11] R. B. D'Agostino and E. S. Pearson. Testing for departures from normality. epirical results for the distribution of b_2 and $\sqrt{b}_1$. *Biometrika*, 60(3):613–622, 1973.

[12] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman & Hall/CRC, third edition, 2003.

[13] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[14] R. B. D'Agostino. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2), 1971.

[15] C. E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423, July 1948.

[16] O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society*, 38:54–59, 1975.

[17] B. van Es. Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, 19:61–72, 1992.

[18] P. J. Bickel and K. A. Doksum. *Mathematical Statistics.* Holden-day, 1977.

[19] G. Blom. *Statistical Estimates and Transformed Beta Variables.* John Wiley and Sons, New York., first edition, 1958.

[20] J. Woodhouse. Harvard PRIM-H, Dept of Geophysics, Harvard University.

[21] A. Buja and W. Rolke. Calibration for simultaneity: (re) sampling methods for simultaneous inference with applications to function estimation and functional data. *Unpublished manuscript*, http://charma.uprm.edu/∼rolke/simulinf.pdf.

[22] W. Rolke. An extension of the normal probability plot. *Unpublished manuscript*, http://charma.uprm.edu/∼rolke/envelope.pdf.

[23] S. M. Ross. *Simulation.* Academic Press, 2006.

A STUDY OF NORMALITY TESTS AND AN EXTENSION TO THE NORMALITY PLOT

Felipe H. Acosta Archila
Department of Mathematics
Chair: Wolfgang Rolke
Degree: Master of Science
Graduation Date: May 2010

We are considering a number of existing normality tests and study their power against different alternative hypotheses. We also develop an extension to the normality plot by adding a fixed confidence band. This procedure is named the envelope test. We use a Monte Carlo simulation to do a power study among all normal tests considered and in the development of the envelope test. The results show that although there was no best normality test, we can provide guidelines to improve their efficiency. With the envelope test we construct a method that eliminates the subjectivity that the normality plot carries within.