On goodness-of-fit tests for the Rayleigh distribution based on the Stein characterisation

E Bothma

orcid.org 0000-0002-8604-0753

Dissertation accepted in partial fulfilment of the requirements for the degree *Master of Science in Mathematical Statistics* at the North-West University

Supervisor: Dr GL Grobler

Co-supervisor: Prof JS Allison

Graduation May 2020

26071134

Acknowledgements

I wish to express appreciation and gratitude to the following individuals:

- Prof James Allison and Dr Gerrit Grobler, my supervisors, for their unwavering support, patience and guidance throughout this process. Thank you for sharing your knowledge of and passion for statistics. This would not have been possible without the two of you and for that I am eternally grateful.
- Prof Leonard Santana for your words of encouragement and for insightful discussions.
- Shawn Liebenberg for discussions regarding the code.
- My mother, brother, and sister-in-law for their unconditional love and support. Also for lending an ear when I needed it. Especially my mom, for the privilege of an excellent education, without her love and encouragement I would not have made it this far.
- My friends, for always being there whenever I needed you.

And to God for the privilege and opportunity to be able to do what I love without His strength, guidance and grace I would not be here.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Abstract

In this mini-dissertation, two new goodness-of-fit tests for the Rayleigh distribution are proposed. These tests are developed by exploiting the Stein characterisation of the Rayleigh distribution. The newly suggested tests are compared with the traditional tests as well as with some more modern tests by making use of a Monte Carlo simulation. The traditional tests include the Kolmogorov-Smirnov, Anderson-Darling and Cramér-von Mises tests. A test based on the empirical Laplace transform and a test based on the cumulative residual entropy are the two modern tests considered. When the powers of the respective tests are compared it can be seen that the newly proposed tests are not only feasible but also very competitive. The results further indicate that the new tests outperform the other tests for most of the alternatives considered in the study. We also provide a proof of the consistency of one of our new tests, as well as a theoretical justification for the choice of our weight function.

Key words: *Goodness-of-fit, Stein characterisation, Rayleigh distribution, Monte Carlo simulation, Asymptotics*

Uittreksel

In hierdie mini-verhandeling word twee nuwe passingstoetse vir die Rayleigh verdeling voorgestel. Hierdie toetse word ontwikkel deur gebruik te maak van die Stein karakterisering. Die nuut voorgestelde toetse word vegelyk met tradisionele toetse asook met 'n paar meer morderne toetse deur gebruik te maak van 'n Monte Carlo simulasie. Die trandisionele toetse wat ingesluit word is die Kolmogorov-Smirnov, Anderson-Darling en Cramér-von Mises toetse. Die moderne toetse wat oorweeg is in die studie is 'n toets wat gebaseer is op die empiriese Laplace transformasie en 'n toets gebaseer op die kumulatiewe residuele entropie. Wanneer die onderskeidingsvermoë van die verskillende toetse vergelyk word is dit duidelik dat die nuutvoorgestelde toetse nie net haalbaar is nie, maar ook mededingend. Verder dui die resultate daarop dat die nuwe toetse die ander toetse oortref, vir die meerderheid van die alternatiewe verdelings wat oorweeg is in die studie. 'n Bewys vir die konsekwentheid van een van ons nuwe toetse word ingesluit asook 'n teoretiese regverdiging vir ons keuse van gewigsfunksie.

Sleutel woorde: *Passingstoetse, Stein karakterisering, Rayleigh verdeling, Monte Carlo simulasie, Asimptotiese teorie*

Table of Contents

Chapter 1

Introduction

1.1 Overview

In 1880 an acoustics problem gave rise to a distribution that plays a prominent role in research fields such as reliability theory, life testing and survival analysis. The Rayleigh distribution was developed by Rayleigh (1880) while undertaking a study regarding the resultant of a great number of sound waves with differing phases. Dyer & Whisenand (1973) and Polovko (1968) demonstrated the importance of the Rayleigh distribution in communication engineering and electro-vacuum devices, respectively. The distance between individuals in a spatial Poisson process tends to follow a Rayleigh distribution. It is also used in oceanography to model the height of waves. Brummer, Mersereau, Eisner & Lewine (1993) found that the Rayleigh distribution has clinical applications, specifically estimating the noise variance of Magnetic Resonance Images (MRI). Sijbers, Poot, den Dekker & Pintjens (2007) found that this estimation can be done by fitting the density function of the Rayleigh distribution to the partial histogram of the MRI. Rajan, Poot, Juntu & Sijbers (2010) improved this estimation with the use of background segmentation. The density function of the Rayleigh distribution is then fitted to the histogram of the segmented background in order to estimate the noise variance. The estimation of the noise forms a crucial part in efficiently denoising the MRI as well as in the quality assessment of these images. The interested reader is referred to Brummer et al. (1993), Sijbers et al. (2007) and Rajan et al. (2010).

For any of the above mentioned applications to be relevant it is crucial to test the hypothesis that the observed data follows a Rayleigh distribution. Since the square of a Rayleigh distributed variable is exponentially distributed, goodness-of-fit tests designed for the exponential distribution can be used to test for the Rayleigh distribution. Even though the applications of the Rayleigh distribution increased significantly over the past few decades, literature on testing the goodness-of-fit for the Rayleigh distribution is relatively scarce. The main aim of this mini-dissertation is to contribute towards this literature

1.2 Objectives

The main objectives of this mini-dissertation can be summarised as follows:

- Provide an overview of the Rayleigh distributions along with some of its properties.
- Provide a discussion of existing goodness-of-fit tests for the Rayleigh distribution.
- Develop two new tests for the Rayleigh distribution using the Stein characterisation.
- Present a proof of consistency for one of our tests.
- Use a Monte Carlo study to analyse the finite sample performance of the newly proposed tests.

1.3 Dissertation outline

In Chapter 2 the Rayleigh distribution will be discussed in much greater detail, this includes the different parameterisations of the Rayleigh distribution as well as its probability structure. It further provides different characterisations of the Rayleigh distribution. The chapter concludes with a discussion on the two parameter estimation methods considered.

Chapter 3 provides a discussion on some already existing goodness-of-fit tests for the Rayleigh distribution. We also introduce some notation and state the hypothesis to be tested formally.

Chapter 4 introduces two new tests that can be used to test whether or not an observed data set is realised from a Rayleigh distribution. We also provide a proof of consistency for one of these tests. In order to develop these new tests we provide an overview of the Stein characterisation along with proofs that the Rayleigh distribution complies with the required properties.

In Chapter 5 the feasibility as well as the finite sample performance of the new tests will be analysed through the use of a Monte Carlo simulation study. We will also provide some conclusions regarding our newly proposed tests in this chapter.

Chapter 2

Rayleigh distribution

In this section the Rayleigh distribution will be discussed, including its different parameterisations as well as its probability structure. Different characterisations of the Rayleigh distribution will also be elaborated on. Lastly, in this chapter we will consider how parameter estimation is done.

2.1 Characteristics of the Rayleigh distribution

Let X be a non-negative random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, X is Rayleigh distributed with parameter θ ($X \sim Ral(\theta)$) if X has the density function

$$f(x) = \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}} \text{ for } x \geq 0 \text{ and } \theta > 0. \tag{2.1}$$

The distribution function of X has a closed form expression and is given by

$$F(x) = 1 - e^{-\frac{x^2}{2\theta^2}} \text{ for } x \geq 0 \text{ and } \theta > 0. \tag{2.2}$$

There also exists an alternative parameterisation of the Rayleigh distribution, where the density function is given as

$$f(x) = \frac{2x}{\Theta^2} \exp\left(\frac{-x^2}{\Theta^2}\right) \text{ for } x \geq 0 \text{ and } \Theta > 0.$$

However, for the remainder of the mini dissertation we will be working with the parameterisation given in equation (2.1). The utility of θ is clear from Figure 2.1 that illustrates the density function of the Rayleigh distribution for various values of θ.

Properties of the Rayleigh distribution will now be discussed. The mean and variance of the Rayleigh distribution, denoted by μ and σ^2 respectively, are

$$\mu = \sqrt{\frac{\pi}{2}}\,\theta \tag{2.3}$$

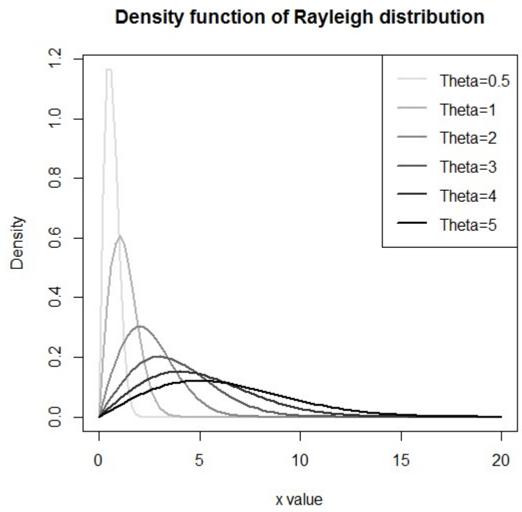

Fig. 2.1: Rayleigh densities for $\theta \in \{0.5, 1, 2, 3, 4, 5\}$.

and

$$\sigma^2 = \frac{4-\pi}{2}\theta^2.$$

Since the Rayleigh distribution is skewed to the right, it can be used to analyse skewed data accurately and effectively. The skewness coefficient is given by

$$\gamma = \frac{2\sqrt{\pi}(\pi-3)}{\sqrt{(4-\pi)^3}} \approx 0.631.$$

This easily computable skewness, which does not depend on the value of the parameter θ, allows one to identify data that could possibly be Rayleigh distributed. This is done by comparing this population skewness value to the sample skewness. A closed form expression for the inverse distribution function of the Rayleigh distribution can be derived from equation (2.2) and is given by

$$F^{-1}(p) = \sqrt{-2\theta^2 \log(1-p)}, \text{ where } 0 < p < 1. \tag{2.4}$$

Due to this relatively simple form it is easy to simulate data from the Rayleigh distribution using the inverse transform method. Quantiles for the Rayleigh distribution can easily be calculated using equation (2.4), for example the median is defined as

$$\nu = F^{-1}\left(\frac{1}{2}\right) = \theta\sqrt{2\log(2)}.$$

The Rayleigh distribution's probability structure can further be explained using either its survival function, $S(x)$, or its hazard rate, $h(x)$. These have the form

$$S(x) = P(X > x) = 1 - F(x) = \exp\left(-\frac{x^2}{2\theta^2}\right) \tag{2.5}$$

and

$$h(x) = \frac{f(x)}{S(x)} = \frac{x}{\theta^2}.$$

The Rayleigh distribution thus has a linearly increasing hazard rate. Furthermore, the mode of the Rayleigh distribution is easily seen to be θ. Therefore the maximum density value is given by

$$f(\theta) = \frac{1}{\theta}\exp\left(-\frac{1}{2}\right) \approx \frac{0.606}{\theta}.$$

This property of the Rayleigh density function can be observed from Figure 2.1. More details about the properties and probability structure of the Rayleigh distribution can be found in Hirano (1986).

There exists a relationship between the density function of the Rayleigh distribution and a few other well-known distributions. These include, but are not limited to, distributions such as the normal distribution, the chi-squared distribution, the gamma distribution and the exponential distribution. These relationships can be described as follows:

- If X and Y are independent normal random variables with mean zero and common variance σ^2 then
  $R = \sqrt{X^2 + Y^2}$ has a Rayleigh distribution with parameter $\theta = \sigma$.
- If $R \sim Ral(1)$, then R^2 has a chi-squared distribution with 2 degrees of freedom.
- If $R_1, R_2, ..., R_N$ are independent and identically $Ral(\theta)$ distributed random variables, then $\sum_{i=1}^{N} R_i^2$ follows a gamma distribution with parameters N and $2\theta^2$.
- If $R \sim Ral(\theta)$ then R^2 is exponentially distributed with parameter θ^2.

Similar to the case of the exponential distribution, the Rayleigh class of distributions is closed under scaled transformations. Therefore, if $X \sim Ral(\theta)$, then for any $k > 0$, the random variable $Y = kX$ is Rayleigh distributed with parameter $k\theta$:

$$F_Y(y) = P(Y \leq y) = P(kX \leq y) = P\left(X \leq \frac{y}{k}\right) = 1 - \exp\left(-\frac{y^2}{2(k\theta)^2}\right). \tag{2.6}$$

By choosing $k = \frac{1}{\theta}$, i.e. $Y = \frac{X}{\theta}$, it is possible to transform any Rayleigh distributed random variable with parameter θ to a Rayleigh distributed random variable with parameter 1. This scale invariance property of the Rayleigh distribution will be used in later chapters when constructing a new goodness-of-fit test for the Rayleigh distribution.

2.2 Characterisations of the Rayleigh distribution

In this section some characterisations of the Rayleigh distribution will be discussed. These include characterisations based on conditional expectations, order statistics, failure rates as well as entropy. Another characterisation, based on the Laplace transform, will be discussed in a later chapter.

2.2.1 Characterisations based on conditional expectations

Ahsanullah & Shakil (2013) provided two characterisations for the Rayleigh distribution based on conditional expectations. The first of these characterisations states that a Rayleigh random variable, X, with parameter θ can be characterised by the following conditional expectation

$$E(X^{2k}|X > t) = \sum_{i=0}^{k} 2^i \theta^{2i} k^{(i)} t^{2(k-i)}, \tag{2.7}$$

with $k^{(i)} = k(k-1)...(k-i+1)$ and $k^{(0)} = 1$ for any $k \geq 1$.

We further have that a random variable X follows a Rayleigh distribution with parameter θ if, and only if,

$$\begin{aligned} E(X^{2k-1}|X > t) = &\sum_{i=0}^{k-1} \frac{(2k-1)!!}{(2k-1-2i)!!(1/\theta^2)^i} t^{2k-1-2i} \\ &+ \frac{(2k-1)!!}{(1/\theta^2)^k} \sqrt{\pi/2\theta^2} \left(1 - \operatorname{erf}\left(\sqrt{\frac{1}{2\theta^2}}\, t\right)\right) e^{t^1/\theta^2} \end{aligned} \tag{2.8}$$

where $(2k-1)!! = 1, 3, ..., (2k-1), k \geq 1$ and

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t,$$

denotes the error function.

These two characterisations based on conditional expectation can be combined to find expressions for the moments of X conditional on a tail event, $\{X > t\}$. For example, from equation (2.8) we have for $k = 1$ the first moment of X conditioned on $\{X > t\}$:

$$E(X|X > t) = t + \theta\sqrt{\frac{\pi}{2}}\left(1 - \text{erf}\left(\sqrt{\frac{1}{2\theta^2}}\, t\right)\right)\exp(t^2\theta^2). \tag{2.9}$$

From equation (2.7) we have for $k = 1$ the second moment of X conditioned on $\{X > t\}$:

$$E(X^2|X > t) = t^2 + 2\theta^2. \tag{2.10}$$

Notice that as $t \to 0^+$, the conditional expectation in equation (2.9) approaches $\theta\sqrt{\frac{\pi}{2}}$, which corresponds with the mean in equation (2.3) and the conditional expectation in equation (2.10) corresponds with the second moment of X.

2.2.2 Characterisation based on order statistics

Ahsanullah & Shakil (2013) further provided a characterisation for the Rayleigh distribution, this time based on order statistics. The characerisation states that a Rayleigh random variable, X, with parameter θ can be characterised using the following order statistic results

$$E(X_{(i)}^{2m}|X_{(i-1)} = t) = \sum_{i=0}^{m} \frac{m!}{(m-i)!}\left(\frac{2\theta^2}{n-i+1}\right)^i t^{2(m-i)}, \tag{2.11}$$

for some $n \geq 1, m \geq 1$, where $X_{(i)}$ is the i^{th} order statistic in a sample of size n.

From (2.11) it is easy to derive the special case for $m = 1$,

$$E(X_{(i)}^2|X_{(i-1)} = t) = t^2 + \frac{\theta^2}{n-i+1}.$$

Proofs for the three characterisations in equations (2.7), (2.8) and (2.11) can be found in Ahsanullah & Shakil (2013).

2.2.3 Characterisation based on the failure rate

Nanda (2010) proved the following characterisation of the Rayleigh distribution based on the hazard rate

$h(x)$.

For any non-negative random variable X,

$$E\left[\frac{h(X)}{X}\right] = \frac{2}{\mu^2(1+c^2)},$$

if, and only if, X has a Rayleigh distribution with mean μ and variance σ^2, with $c^2 = \frac{\sigma^2}{\mu^2}$.

2.2.4 Characterisation based on cumulative residual entropy

Rao, Chen, Vemuri & Wang (2004) introduced an alternative to the well-known Shannon entropy which they called the cumulative residual entropy (CRE). The CRE is based on the survival function and is defined as

$$CRE = -\int_0^\infty S(x)\log S(x)\mathrm{d}x.$$

Since the survival function of the Rayleigh distribution has a closed form expression, see equation (2.5), it can easily be seen that the CRE of the $Ral(\theta)$ distribution is

$$\begin{aligned} CRE &= -\int_0^\infty S(x)\log S(x)\mathrm{d}x \\ &= \int_0^\infty \frac{x^2}{2\theta^2}\exp\left(\frac{x^2}{2\theta^2}\right)\mathrm{d}x \\ &= E\left(\frac{X}{2}\right) = \theta\frac{\sqrt{2\pi}}{4}. \end{aligned}$$

Based on the CRE Baratpour & Khodadadi (2012) proved the following characterisation:

The random variable X attains maximum CRE among all non-negative, absolutely continuous random variables Y subject to the restrictions $E(Y) = \upsilon$ and $E(Y^3) = \omega$, where $\sigma^2 = \frac{\omega}{3\upsilon}$ if, and only if, $X \sim Ral(\theta)$.

2.3 Parameter estimation

Suppose we have a random sample $X_1, ..., X_n$ from the Rayleigh distribution with unknown parameter θ. We will now discuss two methods for estimating θ, namely the method of moments and maximum likelihood estimation.

2.3.1 Method of moments estimators

It is easy to see that the first moment of the Rayleigh distribution with parameter θ is given by

$$\mu_1 = \theta\sqrt{\frac{\pi}{2}}.$$

Now estimating μ_1 by

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

one finds that the method of moments estimate for θ is

$$\tilde{\theta}_n = \bar{X}\sqrt{\frac{2}{\pi}}.$$

This estimate is unbiased and has variance

$$\mathrm{var}(\tilde{\theta}_n) = \frac{\theta^2}{n}(\frac{4-\pi}{\pi}).$$

2.3.2 Maximum likelihood estimators

Maximum likelihood estimation (MLE) is an estimation method that obtains parameter estimates by maximising a likelihood function. Then the maximum likelihood estimate is the value in the parameter space that maximises the likelihood function. In this case the likelihood function, $L_n(\theta)$, has the form

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \frac{X_i}{\theta^2} e^{-\frac{X_i^2}{2\theta^2}}.$$

Sometimes it is more practical to use the log-likelihood function which is just the natural logarithm of the likelihood function. The first derivative of the log-likelihood is set equal to zero and solved for θ to obtain $\hat{\theta}$, the maximum likelihood estimate. When this is done for the Rayleigh distribution one obtains the maximum likelihood estimate for θ^2, which is an unbiased estimator for θ^2, and is given by

$$\hat{\theta}_n^2 = \frac{1}{2n}\sum_{i=1}^{n} X_i^2.$$

The maximum likelihood estimate for θ is consistent but is not an unbiased estimate for θ and is given by

$$\hat{\theta}_n = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} X_i^2}. \tag{2.12}$$

The Fisher information, $I(\theta)$, can be used to approximate the variance of $\hat{\theta}_n$, where $I(\theta) = \frac{4}{\theta^2}$. Thus the variance for $\hat{\theta}_n$ is approximately

$$\mathrm{var}(\hat{\theta}_n) \approx \frac{1}{nI(\theta)} \approx \frac{\theta^2}{4n}.$$

Since $\frac{4-\pi}{\pi} > \frac{1}{4}$ one can conclude that the variance for the maximum likelihood estimate is smaller than that of the method of moments estimate. In what follows we will only consider $\hat{\theta}_n$, the maximum likelihood estimate of θ.

In this chapter different parameterisations and the probability structure of the Rayleigh distribution were discussed. Furthermore, we provided different characterisations of the Rayleigh distribution. Lastly, we discussed two parameter estimation methods along with a motivation for the use of the maximum likelihood estimate throughout this mini-dissertation.

Chapter 3

Goodness-of-fit tests for the Rayleigh distribution

In this section some already existing goodness-of-fit tests for the Rayleigh distribution are discussed. We chose traditional tests as well as some modern tests. We chose the modern tests, since they outperformed the frequently used traditional tests, including a test based on the Laplace transform. This test makes use of a weight function with a tuning parameter, which is comparable to our test. Before commencing the discussion we introduce some notation and state the formal hypothesis to be tested.

Let $X_1, X_2, ..., X_n$ be independent and identically distributed copies of a random variable X with unknown continuous distribution function F. The composite goodness-of-fit hypothesis to be tested is

$$H_0: \text{ the distribution of } X \text{ is } Ral(\theta), \tag{3.1}$$

for some $\theta > 0$, against general alternatives. All the test statistics that will be considered (including our new tests introduced in Chapter 4) are based on the scaled observations $Y_j = \frac{X_j}{\hat{\theta}_n}, j = 1, ..., n$, where $\hat{\theta}_n$ is the maximum likelihood estimate defined in equation (2.12). The motivation for using the scaled observations is given in Chapter 2.1. Denote the order statistics of $X_1, ..., X_n$ and $Y_1, ..., Y_n$ by $X_{(1)} < X_{(2)} < ... < X_{(n)}$ and $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$, respectively.

3.1 Traditional tests based on the empirical distribution function

The traditional tests that we consider are all formulated in terms of the distances between the empirical distribution function of the scaled data $Y_1, ..., Y_n$

$$G_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(Y_j \leq t),$$

and the theoretical distribution function of the standard (i.e. $\theta = 1$) Rayleigh distribution

$$G(t) = 1 - \exp\left(-\frac{t^2}{2}\right).$$

We consider the following three test statistics: The Kolmogorov-Smirnov (KS) test

$$KS_n = \sup_{t \geq 0} |G_n(t) - G(t)|,$$

the Cramér-von Mises (CM) test

$$CM_n = \int_0^\infty (G_n(t) - G(t))^2 dG(t)$$

and the Anderson-Darling (AD) statistic

$$AD_n = \int_0^\infty \frac{[G_n(t) - G(t)]^2}{G(t)[1 - G(t)]} dt.$$

Easily calculable formulae for these three statistics exist and are given by

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

where

$$KS_n^+ = \max_{1 \leq j \leq n} \left[\frac{j}{n} - G(Y_j)\right],$$

$$KS_n^- = \max_{1 \leq j \leq n} \left[G(Y_j) - \frac{j-1}{n}\right],$$

$$CM_n = \frac{1}{12n} + \sum_{j=1}^{n} \left[G(Y_{(j)}) - \frac{2j-1}{n}\right]^2$$

and

$$AD_n = -n - \sum_{j=1}^{n} \frac{2j-1}{n} \left[\log G(Y_{(j)}) - Y_{(n+1-j)}\right].$$

All three of these tests reject H_0 for large values.

The only difference between the CM_n and AD_n tests is the weight function $[G(t)\,(1 - G(t))]^{-1}$ used in the AD_n test statistic. Since the function $w(t) = (G(t)(1 - G(t))$ tails off to zero as t approaches infinity, this choice of the weight function puts more emphasis on observations in the tail of the distribution, and thus, might be more powerful for certain alternatives.

3.2 A test based on the empirical Laplace transform

In general, the Laplace transform (LT) of a random variable X is given by

$$\phi(t) = E(e^{-tX}) = \int_{-\infty}^{\infty} e^{-tx} dF(x).$$

If F is now estimated by

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(X_i \leq x),$$

the empirical distribution function, we obtain the empirical Laplace transform (ELT)

$$\phi_n(t) = \int_{-\infty}^{\infty} e^{-tx} dF_n(x) = \frac{1}{n}\sum_{j=1}^{n} e^{-tX_j}.$$

Meintanis & Iliopoulos (2003) developed a test based on the ELT of the scaled data

$$\varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{-tY_j}.$$

The LT of a standard Rayleigh distribution is

$$\varphi(t) = 1 - \frac{\sqrt{\pi}}{2} t e^{\frac{t^2}{4}} \left[1 - \operatorname{erf}\left(\frac{t}{2}\right)\right],$$

where

$$\operatorname{erf}(x) = \frac{2}{\pi}\int_0^x e^{-t^2} \mathrm{d}t.$$

the approach of Meintanis & Iliopoulos (2003) is justified by the fact that $\varphi(t)$ is the unique solution of the differential equation

$$ty'(t) - \left[1 + \frac{t^2}{2}\right] y(t) + 1 = 0,$$

subject to $\lim_{t\to\infty} y(t) = 0$.

In the spirit of L_2-type tests (see e.g., Baringhaus & Henze (1991)) they propose the test statistic

$$EL_{n,a} = n\int_0^{\infty} D_n^2(t) w(t) \mathrm{d}t,$$

where

$$D_n(t) = t\varphi_n^{'}(t) - \left(1 + \frac{t^2}{2}\right)\varphi_n(t) + 1$$

and $w(t) = e^{-at}$ is a weight function with $a > 0$ a constant tuning parameter. For this choice of weight function the test statistic has the easily calculable form

$$EL_{n,a} = \frac{n}{a} + \frac{1}{n}\sum_{j,k=1}^{n}\left[\frac{1}{(Y_j + Y_k + a)} + \frac{Y_j + Y_k}{(Y_j + Y_k + a)^2} + \frac{2Y_jY_k + 2}{(Y_j + Y_k + a)^3}\right]$$
$$+ \frac{1}{n}\sum_{j,k=1}^{n}\left[\frac{3(Y_j + Y_k)}{(Y_j + Y_k + a)^4} + \frac{6}{(Y_j + Y_k + a)^5}\right] - 2\sum_{j=1}^{n}\left[\frac{1}{(Y_j + a)} + \frac{Y_j}{(Y_j + a)^2} + \frac{1}{(Y_j + a)^3}\right]$$

The test rejects the null hypothesis for large values of the test statistic. In their paper Meintanis & Iliopoulos (2003) derived the limiting null distribution of their test and also investigated the consistency of the test.

3.3 A test based on cumulative residual entropy

The last test we consider is based on the cumulative Kullback-Leibler (CKL) divergence introduced by Baratpour & Khodadadi (2012).

If V_1 and V_2 are two non-negative continuous random variables with survival functions H and Q respectively, then the CKL between H and Q is defined as

$$CKL(H, Q) = \int_0^{\infty} H(x)\log\frac{H(x)}{Q(x)}\mathrm{d}x - [E(V_1) - E(V_2)].$$

Baratpour & Khodadadi (2012) uses the fact that if the null hypothesis is true (i.e. if $X \sim Ral(\theta)$ with survival function $S_0(x) = \exp\left(-\frac{x^2}{2\theta^2}\right)$) then $CKL(S, S_0) = 0$, where S is the unknown survival function of $X_1, X_2, ..., X_n$. Now, if one writes the CKL measure in terms of the CRE measure (discussed earlier), and estimating the unknown quantities, the authors arrive at the following test statistic for testing the hypothesis in (3.1)

$$CR_n = \frac{1}{\bar{X}}\left(\sum_{i=1}^{n-1}\left(\frac{n-i}{n}\right)\left(\log\left(\frac{n-i}{n}\right)\right)\left(X_{(i+1)} - X_{(i)}\right) + \sqrt{\frac{\pi}{2}}\sqrt{\frac{\sum_{i=1}^{n} X_i^3}{3\sum_{i=1}^{n} X_i}}\right),$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The test statistic rejects the null hypothesis for large values. Baratpour & Khodadadi (2012) showed that the test is consistent, and in an extended Monte Carlo study they found the test to be relatively powerful.

3.4 Other tests

Other existing tests for the Rayleigh distribution, which do not form part of this study, include a test based on the Hellinger distance proposed by Jahanshahi, Rad & Fakoor (2016) and a test based on a lesser-known characteristic of the Rayleigh distribution by Liebenberg & Allison (2018). Zamanzade & Mahdizadeh (2017) also proposed a test based on the phi-divergence. Ahrari, Baratpour, Habibirad & Fakoor (2019) suggested a test based on quantiles. Safavinejad, Jomhoori & Alizadeh Noughabi (2015) proposed a density-based empirical likelihood ratio goodness-of-fit test for the Rayleigh distribution. Moment-based goodness-of-fit tests were introduced by Best, Rayner & Thas (2010). Recalling the relationship between the Rayleigh distribution and the exponential distribution, it is possible to perform any goodness-of-fit tests for the exponential distribution by using squared data, $X_1^2, X_2^2, ..., X_n^2$, where $X \sim Ral(\theta)$. Tests based on this transformation was done by Meintanis & Iliopoulos (2003) and Gulati (2011). Meintanis (2008) suggested another test based on transformations, this time using a transformation that results in testing for uniformity.

In this chapter we discussed already existing goodness-of-fit tests for the Rayleigh distribution. We also introduced some notation that will be used in the remainder of this mini-dissertation. The formal hypothesis that will be tested was also stated in this chapter.

Chapter 4

A new test for the Rayleigh distribution based on Stein's characterisation

In this chapter the newly proposed tests will be defined along with a proof of consistency for one of these tests. Before this can be done, the Stein characterisation for the Rayleigh distribution needs to be discussed, since the new goodness-of-fit tests are based on this characterisation.

4.1 The Stein characterisation

The goodness-of-fit tests that are going to be introduced are based on the Stein characterisation in combination with the zero-bias transformation. Goldstein & Reinert (1997) stated that Stein's method permits numerous types of dependence structures to be treated that results in computable bounds for the approximation error. Stein's method achieves this by making use of difference or differential equations that characterise the selected distribution. A well known result in probability theory (see Stein (1972)) is the standard Stein characterisation of the normal distribution, which states that Z is standard normal if, and only if,

$$\mathbb{E}[g'(Z) - Zg(Z)] = 0 \tag{4.1}$$

is true for all absolute continuous functions g for which the expectation exist. Some applications, such as goodness-of-fit tests, based on equation (4.1) is complicated since the results depend on the choice of g. Therefore the function g needs to be chosen carefully. Instead of using this relationship, Betsch & Ebner (2018b) characterised the standard normal distribution based on the zero bias distribution. A real valued random variable X^* is said to have a X zero-bias distribution if

$$\mathbb{E}[g'(X^*)] = \mathbb{E}[Xg(X)]$$

holds for all absolutely continuous functions g for which the expectation exist. If $\mathbb{E}X = 0$ and $Var(X) = 1$, the X zero-bias distribution exists and is unique. It also has the distribution function:

$$T^X(t) = \mathbb{E}[X(X-t)\mathbb{1}\{X \leq t\}], t \in \mathbb{R}.$$

Using this distribution function it can be shown that Z is standard normal if, and only if, the distribution function of Z is given by $T^X(t)$. As a result of this characterisation of the standard normal distribution a goodness-of-fit test for normality can be developed. The test statistic for such a test is based on a distance measure between the theoretical distribution function T^X and an empirical version of T^X (Betsch & Ebner 2018b).

A paper by Betsch & Ebner (2018a) generalised this method to construct a goodness-of-fit test for a wide range of continuous distributions by generalising Stein's characterisation. The characterisation for the standard normal distribution in equation (4.1) is generalised to a continuous random variable X with distribution F by the relationship

$$\mathbb{E}\left[g'(X) + \frac{f'(X)}{f(X)}g(Z)\right] = 0,$$

where f is the density function of X. Betsch & Ebner (2018a) also showed that, provided some conditions on the density and distribution function of X, a characterisation of the distribution of X can be obtained. If X has support $[0, \infty)$, then it has distribution F if, and only if, the distribution function of X is given by

$$T^X(t) = \mathbb{E}\left[-\frac{f'(X)}{f(X)}min\{X, t\}\right], t \in (0, \infty).$$

In the rest of this section we will state the conditions under which the result above is true and we will show that the Rayleigh distribution conforms to these conditions. We will also formulate the characterisation result formally in terms of the Rayleigh distribution, before commencing with the formulation of two new goodness-of-fit tests, based on the Stein characterisation, for the Rayleigh distribution.

Theorem 1 *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space and* f *a density function supported by the interval* $[0, \infty)$*. Denoting by* F *the distribution function associated with* f*, we state the following regularity conditions:*

1. f *is continuously differentiable on* $[0, \infty)$*,*

2. $f(x) > 0$ *for every* $x \in [0, \infty)$*,*

3. *for* $\kappa_f(x) = \left|\frac{f'(x)min\{F(x),1-F(x)\}}{f^2(x)}\right|$ *we have* $sup_{x\in[0,\infty]}\kappa_f(x) < \infty$,

4. $\int_0^\infty (1+|x|)|f'(x)|\mathrm{d}x < \infty$,

5. $lim_{x\to 0}\frac{F(x)}{f(x)} = 0$,

6. $lim_{x\to\infty}\frac{1-F(x)}{f(x)} = 0$.

It can easily be seen that conditions (1), (2), (5) and (6) hold for the Rayleigh distribution. For $X \sim Ral(\theta)$, κ_f in condition (3) becomes:

$$\kappa_f(x) = \frac{\theta^2 e^{x^2/2\theta^2}}{x^2} min\{F(x), 1-F(x)\}\left|1-\frac{x^2}{\theta^2}\right|$$

and for $x^2 > \theta^2$; $\left|1-\frac{x^2}{\theta^2}\right| = \frac{x^2}{\theta^2} - 1$. For x large enough we have that $1-F(x) < F(x)$, thus

$$\begin{aligned}
\lim_{x\to\infty}\kappa_f(x) &= \theta^2\lim_{x\to\infty}\left(\frac{\exp(x^2/2\theta^2)}{x^2}\right)(1-F(x))\left(\frac{x^2}{\theta^2}-1\right)\\
&= \theta^2\lim_{x\to\infty}\left(\frac{x^2}{\theta^2}-1\right)\frac{1}{x^2}\\
&= -1.
\end{aligned}$$

For x small enough we have that $F(x) < 1-F(x)$, thus with the help of L'Hopital's rule we have

$$\begin{aligned}
\lim_{x\to 0}\kappa_f(x) &= \theta^2\lim_{x\to 0}\exp(-x^2/2\theta^2)\left(1-\frac{x^2}{\theta^2}\right)\frac{\left(1-e^{x^2/2\theta^2}\right)}{x^2}\\
&= \theta^2\lim_{x\to 0}\frac{\exp(-x^2/2\theta^2)\left(2x/2\theta^2\right)}{2x}\\
&= \frac{1}{2\theta^2}\theta^2 = \frac{1}{2}.
\end{aligned}$$

Since $\kappa_f(x)$ is continuous with limits 1 and $\frac{1}{2}$ as x tends to infinity and zero, respectively, it implies that $\sup_{x\in[0,\infty)}\kappa_{f(x)} < \infty$.

The integral in condition (4) can be written in terms of expectations:

$$\int_0^\infty (1+|x|)\left(1-\frac{x^2}{\theta^2}\right)\left(\frac{1}{\theta^2}\right)e^{-x^2/2\theta^2}\mathrm{d}x = \mathbb{E}\left(\{1+X\}\left\{1-\frac{X^2}{\theta^2}\right\}\right),$$

where X is Rayleigh distributed. The finite moments of the Rayleigh distribution exist, i.e. $\mathbb{E}(X^k) < \infty,\ k\in\mathbb{N}$. Therefore,

$$\int_0^\infty (1+|x|)|f'(x)|\mathrm{d}x < \infty.$$

We have now shown that all conditions stated are true, which enables us to derive a characterisation of the Rayleigh distribution. Betsch & Ebner (2018a) provides the following theorem to define the fixed point characterisation of a distribution with support $[0, \infty)$:

Theorem 2 *Assume that f is a density function with $spt(f) = [0, \infty]$ that satisfies the conditions 1-5 stated in Theorem 1. Let $X : \Omega \to (0, \infty)$ be a random variable with distribution function F and $\mathbb{E}\left|\frac{f'(X)}{f(X)}X\right| < \infty$. Define $T^X : \mathbb{R} \to \mathbb{R}$ by*

$$T^X(t) = \mathbb{E}\left[-\frac{f'(X)}{f(X)} min\{X, t\}\right], \ t \in (0, \infty),$$

and $T^X = 0, \ t \in (-\infty, 0)$. Then $X \sim \mathcal{F}$ if, and only if, $T^X(t) = F(t)$ for every $t \in \mathbb{R}$.

Since, for the Rayleigh distributed X we have

$$\frac{f'(x)}{f(x)} = \frac{\frac{1}{\theta^2}e^{-x^2/2\theta^2}(-\frac{x^2}{\theta^2} + 1)}{\frac{x}{\theta^2}e^{-x^2/2\theta^2}} = \left(-\frac{x^2}{\theta^2} + 1\right) x^{-1},$$

the condition in the above theorem holds:

$$\mathbb{E}\left|\frac{f'(X)}{f(X)}X\right| = \mathbb{E}\left|1 - \frac{X^2}{\theta^2}\right| < \infty.$$

Therefore, X is Rayleigh distributed if, and only if, the distribution function of X is given by

$$T^X(t) = E\left[\left(\frac{X^2}{\theta^2} - 1\right) X^{-1} \min\{X, t\}\right], \ t \in [0, \infty).$$

This fixed point characterisation will be used in the next section to define new goodness-of-fit tests designed specifically for the Rayleigh distribution.

4.2 New tests

In this section two new goodness-of-fit tests for the Rayleigh distribution will be discussed. As in Chapter 3, let $X_1, X_2, ..., X_n$ be independent and identically distributed copies of a random variable X with unknown continuous distribution F^X. The composite goodness-of-fit hypothesis to be tested is

$$H_0: \text{ the distribution of } X \text{ is } Ral(\theta),$$

for some $\theta > 0$, against general alternatives.

Without loss of generality we assume that $E(X) = \sqrt{\frac{\pi}{2}}$, which implies that $X \sim Ral(1)$ if H_0 is true. With this notation we also have that the maximum likelihood estimator $\hat{\theta}_n$ converges almost surely to one. As in Chapter 3 we will use the scaled observations $Y_i = \frac{X_i}{\hat{\theta}_n}$, which renders $Y_1, ..., Y_n$ dependent observations. Also, under H_0 the scaled observations is approximately Rayleigh distributed with parameter $\theta = 1$ for large n.

A test statistic to test the hypothesis above can be formulated in terms of the empirical version of the distribution function T^Y, given by

$$T_n^Y(t) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \frac{1}{Y_j}\right)\min(Y_j, t),\ t \geq 0,$$

as well the empirical distribution function (EDF) given by

$$F_n^Y(t) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \leq t),\ t \geq 0.$$

The two tests we are about to formulate is based on a distance measure between T_n^Y and F_n^Y, which will be approximately equal to the corresponding distance between T^Y and F^Y for large n. We will prove a consistency result in the next section for one of the proposed tests. In general, a weighted distance measure between T_n^Y and F_n^Y is given by

$$D_n = \int_0^\infty [T_n^Y(t) - F_n^Y(t)]^2 w(t)\mathrm{d}t,$$

for some weight function $w(t)$. In Appendix B we show that the inclusion of a weight function is necessary in order to ensure that D_n is finite. Without an appropriate weight function the integral above does not exist.

For the first test statistic we choose $w(t) = e^{-at}f^Y(t)$, where $f^Y(t)$ is the density function corresponding to the unknown distribution function F^Y. This results in

$$D_n = \int_0^\infty [T_n^Y(t) - F_n^Y(t)]^2 w(t)\mathrm{d}t$$

$$
\begin{aligned}
&= \int_0^\infty [T_n^Y(t) - F_n^Y(t)]^2 e^{-at} f(t) \mathrm{d}t \\
&= E\left[T_n^Y(T) - F_n^Y(T)\right]^2 e^{-aT},
\end{aligned}
$$

where T has distribution F^Y. Approximating the quantity above using observations $Y_1, Y_2, ..., Y_n$ results in the test statistic

$$
S_{n,a} = \sum_{k=1}^{n} [T_n^Y(Y_k) - F_n^Y(Y_k)]^2 e^{-aY_k}.
$$

Note that the test includes an unknown tuning parameter, a, as part of the weight function. The choice of parameter a influences the performance of the test, which will be analysed in the Monte Carlo study in Chapter 5.

For the second test we choose $w(t) = e^{-at}$, which results in the following test statistic

$$
R_{n,a} = n \int_0^\infty [T_n^Y(t) - F_n^Y(t)]^2 e^{-at} \mathrm{d}t. \tag{4.2}
$$

This equation for $R_{n,a}$ will be used to prove consistency of this test statistic in the next section, but to apply this formula in a Monte Carlo simulation study requires numerical integration. However, to avoid this we derive a calculable form of $R_{n,a}$. Note that we will use ordered statistics $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$ in the calculation of $R_{n,a}$. Therefore, we provide new formulations for the empirical version of the distribution function T^Y

$$
T_n^Y(t) = \frac{1}{n} \sum_{j=1}^{n} \left(Y_{(j)} - \frac{1}{Y_{(j)}} \right) \min(Y_{(j)}, t), \ \ t \geq 0,
$$

as well for the empirical distribution function (EDF) given by

$$
F_n^Y(t) = \frac{1}{n} \sum_{j=1}^{n} I(Y_{(j)} \leq t), \ \ t \geq 0.
$$

The calculable form of $R_{n,a}$ is then

$$
\begin{aligned}
R_{n,a} &= n \int_0^\infty [T_n^Y(t) - F_n(t)]^2 e^{-at} \mathrm{d}t \\
&= n \int_0^\infty \left(\frac{1}{n} \sum_{j=1}^{n} \left\{ Y_{(j)} - \frac{1}{Y_{(j)}} \right\} \min\left\{ Y_{(j)}, t \right\} - \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\left\{ Y_{(j)} \leq t \right\} \right)^2 e^{-at} \mathrm{d}t
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\int_0^\infty \left(\sum_{j=1}^n \left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}\min\left\{Y_{(j)}, t\right\}\right)^2 e^{-at}\mathrm{d}t \\
&\qquad - \frac{2}{n}\int_0^\infty \left(\sum_{j=1}^n \left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}\min\left\{Y_{(j)}, t\right\}\right)\left(\sum_{j=1}^n \mathbb{1}\left\{Y_{(j)} \le t\right\}\right) e^{-at}\mathrm{d}t \\
&\qquad\qquad + \frac{1}{n}\int_0^\infty \left(\sum_{j=1}^n \mathbb{1}\left\{Y_{(j)} \le t\right\}\right)^2 e^{-at}\mathrm{d}t \\
&= \frac{1}{n}\int_0^\infty \sum_{j=1}^n \Bigg(\left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}^2 \min\left\{Y_{(j)}, t\right\}^2 - 2\left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}\min\left\{Y_{(j)}, t\right\}\mathbb{1}\left\{Y_{(j)} \le t\right\} \\
&\qquad\qquad + \mathbb{1}\left\{Y_{(j)} \le t\right\}\Bigg) e^{-at}\mathrm{d}t \\
&+ \frac{2}{n}\int_0^\infty \sum_{1\le j<k\le n} \Bigg(\left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}\left\{Y_{(k)} - \frac{1}{Y_{(k)}}\right\}\min\left\{Y_{(j)}, t\right\}\min\left\{Y_{(k)}, t\right\} \\
&\qquad\qquad - \left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}\min\left\{Y_{(j)}, t\right\}\mathbb{1}\left\{Y_{(k)} \le t\right\} \\
&\qquad\qquad\qquad - \left\{Y_{(k)} - \frac{1}{Y_{(k)}}\right\}\min\left\{Y_{(k)}, t\right\}\mathbb{1}\left\{Y_{(j)} \le t\right\} \\
&\qquad\qquad\qquad\qquad + \mathbb{1}\left\{Y_{(j)} \le t\right\}\mathbb{1}\left\{Y_{(k)} \le t\right\}\Bigg) e^{-at}\mathrm{d}t \\
&= \frac{1}{n}\sum_{j=1}^n \left(-\frac{1}{a}e^{-aY_{(j)}}\left[\left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}^2\left\{\frac{2}{a}Y_{(j)} + \frac{2}{a^2}\right\} + 2Y_{(j)}^2 - 3\right] + \frac{2}{a^3}\left[Y_{(j)}^2 - 2 + \frac{1}{Y_{(j)}^2}\right]\right) \\
&+ \frac{2}{n}\sum_{1\le j<k\le n} \Bigg(\left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}\left\{Y_{(k)} - \frac{1}{Y_{(k)}}\right\}\left(-\frac{1}{a}e^{-aY_{(j)}}\left\{\frac{1}{a}Y_{(j)} + \frac{2}{a^2}\right\} + \frac{2}{a^3} - \frac{Y_{(j)}}{a^2}e^{-aY_{(k)}}\right) \\
&\qquad\qquad + \left\{Y_{(j)} - \frac{1}{Y_{(j)}}\right\}\left\{-\frac{Y_{(j)}}{a}e^{-aY_{(k)}}\right\} \\
&\qquad\qquad\qquad + \left\{Y_{(k)} - \frac{1}{Y_{(k)}}\right\}\left\{\frac{1}{a^2}e^{-aY_{(k)}} - \frac{1}{a}e^{-aY_{(j)}}\left(Y_{(j)} + \frac{1}{a}\right)\right\} + \frac{1}{a}e^{-aY_{(k)}}\Bigg).
\end{aligned}
$$

The last step follows from tedious integration that is shown in Appendix A.

4.3 A consistency result

In this section we show that for the test defined in (4.2) the weighted distance between T_n^Y and F_n^Y, which will be approximately equal to the corresponding distance between T^Y and F^Y for large n. More formally, to prove consistency we want to show that

$$
\frac{R_{n,a}}{n} \xrightarrow{\mathrm{P}} \int_0^\infty [T^X(t) - F^X(t)]^2 e^{-at}\mathrm{d}t,
$$

where $T^X(t)$ is given by

$$T^X(t) = E\left[\left(X^2 - 1\right) X^{-1} \min\{X, t\}\right], \; t \in [0, \infty),$$

$F(t)$ is the c.d.f. of X and $\xrightarrow{\mathrm{P}}$ denotes convergence in probability.

The first step is to show that $T_n^Y(t) \xrightarrow{\mathrm{as}} T^Y(t) = T^X(t)$, where $\xrightarrow{\mathrm{as}}$ denotes almost sure convergence. This is a direct result of the classical law of large numbers if our observations $Y_1, Y_2, ..., Y_n$ were independent. However, the scaled observations are not independent. In general, to use limit theorems to prove consistency we rewrite our test statistic $R_{n,a}$ in equation (4.2) in terms of the unscaled observations X_i by using the substitution $t = s/\hat{\theta}_n$:

$$\begin{aligned}
R_{n,a} &= n \int_0^\infty \left(\frac{1}{n} \sum_{j=1}^n \left\{ \frac{X_{(j)}}{\hat{\theta}_n} - \frac{\hat{\theta}_n}{X_{(j)}} \right\} \min\left\{ \frac{X_{(j)}}{\hat{\theta}_n}, t \right\} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}\left\{ \frac{X_{(j)}}{\hat{\theta}_n} \leq t \right\} \right)^2 e^{-at} \mathrm{d}t \\
&= \frac{n}{\hat{\theta}_n} \int_0^\infty \left(\frac{1}{n\hat{\theta}_n} \sum_{j=1}^n \left\{ \frac{X_{(j)}}{\hat{\theta}_n} - \frac{\hat{\theta}_n}{X_{(j)}} \right\} \min\left\{ X_{(j)}, s \right\} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}\left\{ X_{(j)} \leq s \right\} \right)^2 e^{-as/\hat{\theta}_n} \mathrm{d}s \\
&= \frac{1}{\hat{\theta}_n} \int_0^\infty \left(\sqrt{n} \left\{ \hat{T}_n^X(s) - F_n^X(s) \right\} \right)^2 e^{-as/\hat{\theta}_n} \mathrm{d}s,
\end{aligned}$$

where

$$\hat{T}_n^X(s) = \frac{1}{n\hat{\theta}_n^2} \sum_{j=1}^n \left(X_{(j)} - \frac{\hat{\theta}_n^2}{X_{(j)}} \right) \min\left\{ X_{(j)}, s \right\}. \tag{4.3}$$

The first step is to show that $\hat{T}_n^X(s) \xrightarrow{\mathrm{as}} T^X(s)$:

Lemma 1. *Let $\hat{T}_n^X(s)$ be defined as in (4.3), then $\hat{T}_n^X(s) \xrightarrow{as} T^X(s)$.*

Proof. The idea in this proof is to rewrite $\hat{T}_n^X(t)$ into two components of which one component is a function of $T_n^X(t)$. This is important since $T_n^X(t) \to^{as} T^X(t)$ by the strong law of large numbers. Therefore,

$$\begin{aligned}
\hat{T}_n^X(s) &= \frac{1}{n\hat{\theta}_n^2} \sum_{j=1}^n \left(X_{(j)} \min\left\{ X_{(j)}, s \right\} - \frac{\hat{\theta}_n^2}{X_{(j)}} \min\left\{ X_{(j)}, s \right\} \right) \\
&= \frac{1}{n\hat{\theta}_n^2} \sum_{j=1}^n \left[\left(X_{(j)} - \frac{1}{X_{(j)}} \right) \min\left\{ X_{(j)}, s \right\} + \left(\frac{1}{X_{(j)}} - \frac{\hat{\theta}_n^2}{X_{(j)}} \right) \min\left\{ X_{(j)}, s \right\} \right] \\
&= \frac{1}{\hat{\theta}_n^2} T_n^X(s) + \frac{1 - \hat{\theta}_n^2}{n\hat{\theta}_n^2} \sum_{j=1}^n \frac{1}{X_{(j)}} \min\left\{ X_{(j)}, s \right\}.
\end{aligned}$$

Therefore, by applying the strong law of large numbers and a continuous mapping theorem we have

$$\hat{T}_n^X(s) \xrightarrow{\text{as}} T^X(s) + 0 \times \mathbb{E}\left[\frac{1}{X}\min\{X, s\}\right].$$

Since $X \geq 0$ and

$$\frac{1}{X}\min\{X,s\} = \underbrace{\begin{cases} 1 & \text{if} \quad X \leq s \\ \frac{s}{X} & \text{if} \quad X > s \end{cases}}_{\leq 1},$$

we have

$$0 \leq \mathbb{E}\left[\frac{1}{X}\min\{X,s\}\right] \leq 1,$$

which completes the proof. □

We have now shown that $\hat{T}_n^X(t)$ converges almost surely to a distribution function $T^X(t)$. Therefore, our next step is to show that $\hat{T}_n^X(t)$ is in fact (similar to the empirical cumulative distribution function $F_n(t)$) a distribution function for each fixed n.

Lemma 2. *The random function $\hat{T}_n^X(t)$, as defined in* (4.3), *is a distribution function for each fixed n.*

Proof. First we note that the function $\min\{X_{(j)}, t\}$ can be written as

$$\min\{X_{(j)}, t\} = \int_0^t \mathbb{1}\{X_{(j)} \geq s\}\,\mathrm{d}s.$$

As a result, $\hat{T}_n^X(t)$ can be written as

$$\begin{aligned}
\hat{T}_n^X(t) &= \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)} - \frac{\hat{\theta}_n^2}{X_{(j)}}\right)\min\{X_{(j)}, t\}.\\
&= \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)} - \frac{\hat{\theta}_n^2}{X_{(j)}}\right)\int_0^t \mathbb{1}\{X_{(j)} \geq s\}\,\mathrm{d}s\\
&= \int_0^t \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)} - \frac{\hat{\theta}_n^2}{X_{(j)}}\right)\mathbb{1}\{X_{(j)} \geq s\}\,\mathrm{d}s.
\end{aligned}$$

The proof will be complete if we can show that the function $f_{\hat{T}_n^X}(t)$, defined by

$$f_{\hat{T}_n^X}(t) = \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)} - \frac{\hat{\theta}_n^2}{X_{(j)}}\right)\mathbb{1}\{X_{(j)} \geq t\},$$

is a density function for each fixed n.

The total mass of the function $f_{\hat{T}_n^X}$ is one:

$$\begin{aligned}\int_0^\infty f_{\hat{T}_n^X}(t)\mathrm{d}t &= \int_0^\infty \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)}-\frac{\hat{\theta}_n^2}{X_{(j)}}\right)\mathbb{1}\left\{X_{(j)}\geq t\right\}\mathrm{d}t\\ &= \int_0^{X_{(j)}} \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)}-\frac{\hat{\theta}_n^2}{X_{(j)}}\right)\mathrm{d}t\\ &= \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)}^2-\hat{\theta}_n^2\right)\\ &= \frac{1}{n\hat{\theta}_n^2}\left(2n\hat{\theta}_n^2-n\hat{\theta}_n^2\right)=1,\end{aligned}$$

where the MLE of $\hat{\theta}^2$ is given by $\hat{\theta}_n^2=\frac{1}{2n}\sum_{j=1}^n X_{(j)}^2$. To show that $f_{\hat{T}_n^X}\geq 0$, we have

$$\begin{aligned}f_{\hat{T}_n^X}(t) &= \frac{1}{n\hat{\theta}_n^2}\sum_{j=1}^n\left(X_{(j)}-\frac{1}{2n}\left(\sum_{k=1}^n X_{(k)}^2\right)\frac{1}{X_{(j)}}\right)\mathbb{1}\left\{X_{(j)}\geq t\right\}\\ &= \frac{1}{n\hat{\theta}_n^2}\left(\left\{1-\frac{1}{2n}\right\}\sum_{j=1}^n\left(X_{(j)}\mathbb{1}\left\{X_{(j)}\geq t\right\}\right)-\frac{1}{2n}\sum_{j\neq k}^n\frac{X_{(j)}^2}{X_{(k)}}\mathbb{1}\left\{X_{(k)}\geq t\right\}\right),\end{aligned}\tag{4.4}$$

since

$$\sum_{j=1}^n X_{(j)}^2\sum_{j=1}^n\frac{1}{X_{(j)}}\mathbb{1}\left\{X_{(j)}\geq t\right\}=\sum_{j=1}^n X_{(j)}\mathbb{1}\left\{X_{(j)}\geq t\right\}+\sum_{j\neq k}^n\frac{X_{(j)}^2}{X_{(k)}}\mathbb{1}\left\{X_{(k)}\geq t\right\}.$$

Now assume that $(n-l)$ values of X_j is greater than fixed t, then we have

$$\begin{aligned}-\frac{1}{2n}\sum_{j\neq k}^n\frac{X_{(j)}^2}{X_{(k)}}\mathbb{1}\left\{X_{(k)}\geq t\right\} &= -\frac{1}{2n}\sum_{j\neq k}^n\frac{X_{(j)}^2}{X_{(k)}}\mathbb{1}\left\{-\frac{1}{X_{(k)}}\geq-\frac{1}{t}\right\}\\ &= \frac{1}{2n}\sum_{j=1}^n\sum_{j\neq k=l+1}^n\frac{X_{(j)}^2}{-X_{(k)}}\\ &\geq -\frac{1}{2n}\frac{n-l-1}{t}\sum_{j=1}^n X_{(j)}^2\\ &\geq -\frac{1}{2n}\frac{n-l-1}{t}\sum_{j=l+1}^n X_{(j)}^2\\ &\geq -\frac{1}{2n}t(n-l)(n-l-1),\end{aligned}$$

and

$$\left\{1-\frac{1}{2n}\right\}\sum_{j=1}^n X_{(j)}\mathbb{1}\left\{X_{(j)}\geq t\right\}=\left\{1-\frac{1}{2n}\right\}\sum_{j=l+1}^n X_{(j)}\geq\left\{1-\frac{1}{2n}\right\}t(n-l).$$

Therefore, combining these results into equation (4.4) we have

$$f_{\hat{T}_n^X}(t) \geq \frac{1}{n\hat{\theta}_n^2}\left(\left\{1 - \frac{1}{2n}\right\} t(n-l) - \frac{1}{2n} t(n-l)(n-l-1)\right) = \frac{t(n-l)(n+l)}{2n^2\hat{\theta}_n^2} \geq 0.$$

□

The last result we need before we can prove consistency is that the rescaled weight function does not have an influence on the asymptotic behaviour of our test statistic $R_{n,a}$ (see Betsch & Ebner (2019) for a similar result). Therefore, we assume

$$\begin{aligned}
&\int_0^\infty \left(\sqrt{n}\left\{\hat{T}_n^X(t) - F_n^Y(t)\right\}\right)^2 e^{-at/\hat{\theta}_n} \mathrm{d}t \\
&= \int_0^\infty \left(\sqrt{n}\left\{\hat{T}_n^X(t) - F_n^Y(t)\right\}\right)^2 e^{-at} \mathrm{d}t + o_{\mathbb{P}}(1).
\end{aligned}$$

To formalise the results we note that the test statistic $R_{n,a}$ can be written in terms of the norm $||\cdot||_{\mathcal{H}}$ in a Hilbert space $\mathcal{H}$, where the random functions $\hat{T}_n^X(t)$ and $F_n^Y(t)$ are random elements of $\mathcal{H}$. Therefore, we write

$$\frac{R_{n,a}}{n} = \frac{1}{\hat{\theta}_n}\int_0^\infty \left\{\hat{T}_n^X(t) - F_n^X(t)\right\}^2 e^{-at}\mathrm{d}t + o_{\mathbb{P}}(1) = \frac{1}{\hat{\theta}_n}||\hat{T}_n^X(t) - F_n^X(t)||_{\mathcal{H}}^2 + o_{\mathbb{P}}(1). \tag{4.5}$$

Theorem 1. *As $n \to \infty$, we have*

$$\frac{R_{n,a}}{n} \xrightarrow{\mathbb{P}} ||T^X(t) - F^X(t)||_{\mathcal{H}}^2.$$

Proof. We first note that according to the classical Glivenko-Cantelli theorem we have

$$\sup_t |F_n^X(t) - F^X(t)| \xrightarrow{\text{as}} 0,$$

and since we have shown that $\hat{T}_n^X(t) \xrightarrow{\text{as}} T^X(t)$ and $\hat{T}_n^X(t)$ is a distribution function for each fixed n, the Glivenko-Cantelli result for $\hat{T}_n^X(t)$ and $T^X(t)$ also holds:

$$\sup_t |\hat{T}_n^X(t) - T^X(t)| \xrightarrow{\text{as}} 0.$$

By the triangle inequality we have

$$||\hat{T}_n^X(t) - F_n^X(t)||_{\mathcal{H}}^2 \leq ||\hat{T}_n^X(t) - T^X(t)||_{\mathcal{H}}^2 + ||T^X(t) - F^X(t)||_{\mathcal{H}}^2 + ||F_n^X(t) - F^X(t)||_{\mathcal{H}}^2.$$

Therefore, we have

$$
\begin{aligned}
&\Big| ||\hat{T}_n^X(t) - F_n^X(t)||^2_{\mathcal{H}} - ||T^X(t) - F^X(t)||^2_{\mathcal{H}} \Big| \\
&\leq ||\hat{T}_n^X(t) - T^X(t)||^2_{\mathcal{H}} + ||F_n^X(t) - F^X(t)||^2_{\mathcal{H}} \\
&\leq \int_0^\infty \sup_t |\hat{T}_n^X(t) - T^X(t)|^2 e^{-at}\mathrm{d}t + \int_0^\infty \sup_t |F_n^X(t) - F^X(t)|^2 e^{-at}\mathrm{d}t \\
&= \int_0^\infty \left(\sup_t |\hat{T}_n^X(t) - T^X(t)| \right)^2 e^{-at}\mathrm{d}t + \int_0^\infty \left(\sup_t |F_n^X(t) - F^X(t)| \right)^2 e^{-at}\mathrm{d}t \\
&= \left\{ \left(\sup_t |\hat{T}_n^X(t) - T^X(t)| \right)^2 + \left(\sup_t |F_n^X(t) - F^X(t)| \right)^2 \right\} \int_0^\infty e^{-at}\mathrm{d}t \xrightarrow{\text{as}} 0.
\end{aligned}
$$

Applying a continuous mapping theorem and taking the limit in (4.5) we obtain the result, since almost sure convergence implies convergence in probability. $\square$

In this chapter the Stein characterisation as well as the zero bias transformation were discussed. These were used to define our newly proposed tests. We also provided an asymptotic proof of consistency for one of our new tests. The performance of these tests will be tested in the next section through the use of a Monte Carlo study.

Chapter 5

Simulation study and conclusions

In this chapter the performance of our newly proposed tests will be analysed by means of a power study. The powers used in this study are obtained through Monte Carlo simulation. All the tests and comparisons will be done using a significance level of 5%. The critical values are obtained through the use of 50 000 independent Monte Carlo replications drawn from a $Ral(1)$ distribution. Power estimates are calculated and reported for sample sizes $n = 20$ and $n = 30$ using 10 000 independent Monte Carlo replications for various alternative distributions. These include some 'local' alternatives as well as those given in Table 5.1. These alternative distributions were chosen since they are frequently used alternatives for the Rayleigh distribution, which has an increasing hazard rate. The hazard rates of the considered alternative distributions include constant hazard rates (CHR), decreasing hazard rates (DHR) and non-monotone hazard rates (NMHR). Other distributions that also have increasing hazard rates (IHR) are included as well.

Table 5.1: Probability density functions of the alternative distributions.

Alternative	f(x)	Notation
Gamma	$\frac{1}{\Gamma(\theta)}x^{\theta-1}\exp(-x)$	$\Gamma(\theta)$
Weibull	$\theta x^{\theta-1}\exp(-x^{\theta})$	$W(\theta)$
Power	$\frac{1}{\theta}x^{(1-\theta)/\theta}, 0 < x < 1$	$PW(\theta)$
Linear Failure Rate	$(1+\theta x)\exp\left(-x-\frac{\theta x^2}{2}\right)$	$LFR(\theta)$
Lognormal	$\frac{\exp\left\{-\frac{1}{2}\left(\frac{\log(x)}{\theta}\right)^2\right\}}{\{\theta x\sqrt{2\pi}\}}$	$LN(\theta)$
Inverse Gaussian	$\left(\frac{\theta}{2\pi x^3}\right)^{1/2}\exp\left\{\frac{-\theta(x-1)^2}{2x}\right\}$	$IG(\theta)$
Gompertz	$\exp(-\theta x)\exp\left\{-\left(\frac{1}{\theta}\right)(\exp(\theta x)-1)\right\}$	$GO(\theta)$
Exponential	$\theta\exp(-\theta x)$	$EXP(\theta)$
Extreme value	$\frac{1}{\theta}\exp\left(x+\frac{1-\exp(x)}{\theta}\right)$	$EV(\theta)$
Exponential geometric	$\frac{(1-\theta)\exp(-x)}{(1-\theta\exp(-x))^2}$	$EG(\theta)$

The tests that are used for comparison are the tests that were discussed in Chapter 3. The estimated powers of the $S_{n,a}$, $R_{n,a}$ and $EL_{n,a}$ test statistics are functions of a tuning parameter, a. In this mini dissertation results for two fixed values of a are reported, namely $a = 1$ and $a = 5$. The motivation for these choices of a will be provided as a part of the discussion of the results. All simulations and calculations are done in R (R Core Team 2019). We first consider some local power estimates. Here we consider two mixture distributions, the first one is obtained by sampling with probability p from an

$Exp(1)$ distribution and with probability $(1-p)$ from a $Ral(1)$ distribution. These estimate powers are given in Table 5.2. The second mixture is where we sample with probability p from a $\Gamma(2)$ distribution and with probability $(1-p)$ from a $Ral(1)$ distribution. These estimated powers are given in Table 5.3. The estimated powers for sample sizes 20 and 30 against every alternative distribution in Table 5.1 are given in Tables 5.4 and 5.5, respectively. The entries in these tables are the percentages of 10 000 independent Monte Carlo samples that resulted in the rejection of the null hypothesis. These estimated powers are rounded to the nearest integer. For the reader's comfort the highest power against each distribution is highlighted in the simulation results.

Table 5.2: Estimated local powers for the mixture of the Rayleigh and exponential distributions for various choices of the mixture parameter, p.

p	n	KS_n	CM_n	AD_n	$EL_{n,1}$	$EL_{n,5}$	CR_n	$S_{n,1}$	$S_{n,5}$	$R_{n,1}$	$R_{n,5}$
0	20	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
	30	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
0.05	20	6	5	7	8	7	7	6	**9**	7	**9**
	30	6	7	8	**10**	8	8	7	**10**	8	8
0.1	20	7	7	11	**14**	10	9	8	**14**	10	**14**
	30	8	9	13	16	12	12	10	**18**	11	15
0.15	20	10	10	17	20	14	13	11	**21**	15	20
	30	11	13	19	24	18	16	14	**26**	18	23
0.2	20	13	14	22	27	20	17	15	**28**	18	27
	30	15	17	27	33	24	20	20	**36**	24	33
0.25	20	16	18	28	34	25	21	20	**35**	25	34
	30	21	24	37	45	34	27	26	**46**	32	42
0.3	20	19	22	36	**42**	31	25	24	**42**	30	**42**
	30	25	30	44	52	41	33	32	**56**	40	51
0.35	20	24	27	42	49	38	30	29	**50**	36	49
	30	32	37	53	62	50	38	41	**64**	47	61
0.4	20	29	33	49	**57**	44	35	35	**57**	42	56
	30	38	45	61	69	58	44	49	**71**	56	69
0.45	20	34	38	57	**64**	51	40	41	63	50	63
	30	46	52	69	77	65	51	57	**78**	64	75
0.5	20	39	44	61	68	56	45	47	**70**	56	69
	30	54	60	76	82	72	58	63	**83**	71	81

We will now present some general conclusions regarding the tabulated estimated powers of the different tests considered. Note that since the performance of the tests are affected by the type of hazard rate of the alternative distribution, we will discuss the overall performance as well as the performance when the results are grouped according to the type of hazard rate. These hazard rate groups are classified as increasing, decreasing and non-monotone.

To begin, we will consider the estimated local powers, presented in Tables 5.2 and 5.3, where we will discuss the results for each of the two mixture distributions under investigation. In Table 5.2, the results of the mixture distributions consisting of the standard Rayleigh distribution and the standard exponential distribution are presented. We find that the KS_n, CM_n, and $S_{n,1}$ tests exhibit poor power performance,

Table 5.3: Estimated local powers for the mixture of the Rayleigh and Gamma distributions for various choices of the mixture parameter, p.

p	n	KS_n	CM_n	AD_n	$EL_{n,1}$	$EL_{n,5}$	CR_n	$S_{n,1}$	$S_{n,5}$	$R_{n,1}$	$R_{n,5}$
0	20	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
	30	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
0.05	20	9	9	10	8	11	**13**	9	10	11	9
	30	10	11	11	9	13	**16**	12	11	13	9
0.1	20	12	14	15	12	16	**20**	14	14	17	13
	30	15	18	18	14	22	**26**	18	18	21	14
0.15	20	16	18	19	16	22	**26**	18	19	22	16
	30	20	23	24	18	28	**35**	25	24	27	19
0.2	20	20	22	23	18	26	**30**	22	23	26	20
	30	24	28	29	22	34	**40**	30	30	33	24
0.25	20	22	25	27	22	31	**35**	26	27	30	23
	30	30	33	35	27	40	**45**	34	34	39	29
0.3	20	25	28	31	25	34	**38**	29	30	34	26
	30	33	37	38	31	44	**50**	39	40	44	32
0.35	20	28	31	33	27	37	**41**	32	33	37	29
	30	37	42	44	36	49	**54**	44	44	48	36
0.4	20	29	33	36	30	39	**43**	35	36	39	31
	30	40	45	47	39	53	**56**	47	49	53	41
0.45	20	32	36	39	32	42	**46**	36	38	43	34
	30	42	48	49	41	55	**58**	51	51	55	43
0.5	20	33	37	40	34	44	**46**	38	40	44	36
	30	45	50	52	45	57	**60**	53	54	57	46

Table 5.4: Estimated powers for general alternatives for sample size $n = 20$.

	KS_n	CM_n	AD_n	$EL_{n,1}$	$EL_{n,5}$	CR_n	$S_{n,1}$	$S_{n,5}$	$R_{n,1}$	$R_{n,5}$
CHR										
$EXP(1)$	87	90	96	**97**	95	89	92	**97**	95	**97**
IHR										
$\Gamma(1.5)$	58	64	73	75	73	64	68	**79**	73	76
$\Gamma(2)$	34	39	44	42	45	42	40	**50**	45	43
$W(1.2)$	64	70	81	85	79	69	72	**86**	78	85
$W(1.4)$	37	42	54	58	53	43	44	**62**	52	60
$PW(1)$	15	18	40	41	12	20	14	40	19	**42**
$LFR(2)$	39	43	62	69	56	42	46	**70**	56	69
$LFR(4)$	26	30	47	55	40	29	33	**57**	41	56
$EV(0.5)$	57	62	79	**85**	74	58	65	84	73	84
$EV(1.5)$	23	25	46	**56**	33	19	28	**56**	36	**56**
$GO(0.5)$	17	18	37	**46**	23	13	19	45	24	45
$GO(1.5)$	48	53	72	**79**	65	48	57	**79**	65	**79**
DHR										
$\Gamma(0.4)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$\Gamma(0.7)$	97	98	**100**	**100**	99	98	99	**100**	99	**100**
$W(0.8)$	98	98	**100**	**100**	99	98	99	**100**	99	**100**
$EG(0.2)$	91	94	**98**	**98**	97	93	95	**98**	97	**98**
$EG(0.5)$	96	97	99	99	99	97	98	**100**	99	99
$EG(0.8)$	99	99	**100**	**100**	**100**	99	**100**	**100**	**100**	**100**
NMHR										
$PW(2)$	88	90	**99**	**99**	94	85	91	**99**	95	**99**
$PW(3)$	99	99	**100**	**100**	**100**	99	**100**	**100**	99	**100**
$LN(0.8)$	67	71	73	66	75	74	72	74	**76**	68
$LN(1)$	90	92	94	93	94	91	92	**96**	94	94
$LN(1.5)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$IG(0.5)$	96	97	98	98	98	97	98	**99**	98	98
$IG(1.5)$	57	61	61	47	62	**64**	60	60	**64**	50

displaying the lowest powers among the tests for the majority of the choices of the mixture probability, p. The $S_{n,5}$ test outperforms all the other tests as the powers against the test is the highest for almost all of the alternatives considered. We note, however, that the $EL_{n,1}$ and $R_{n,5}$ are tied for the best test

Table 5.5: Estimated powers for general alternatives for sample size $n = 30$.

	KS_n	CM_n	AD_n	$EL_{n,1}$	$EL_{n,5}$	CR_n	$S_{n,1}$	$S_{n,5}$	$R_{n,1}$	$R_{n,5}$
CHR										
$EXP(1)$	97	98	99	**100**	99	97	99	**100**	99	**100**
IHR										
$\Gamma(1.5)$	76	82	88	89	88	79	84	**91**	87	89
$\Gamma(2)$	46	52	57	55	60	53	54	**64**	60	57
$W(1.2)$	81	86	92	94	92	82	88	**95**	91	94
$W(1.4)$	51	58	69	73	69	56	62	**77**	67	72
$PW(1)$	21	28	**53**	51	14	36	18	51	24	50
$LFR(2)$	53	59	77	83	72	53	64	**84**	71	82
$LFR(4)$	38	42	60	69	54	37	45	**71**	53	68
$EV(0.5)$	74	79	91	94	87	72	83	**95**	88	93
$EV(1.5)$	32	35	59	68	45	24	38	**70**	46	66
$GO(0.5)$	21	24	46	57	30	13	25	**58**	32	56
$GO(1.5)$	64	70	85	**91**	80	61	75	**91**	80	90
DHR										
$\Gamma(0.4)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$\Gamma(0.7)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$W(0.8)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$EG(0.2)$	98	99	**100**	**100**	**100**	98	99	**100**	**100**	**100**
$EG(0.5)$	99	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$EG(0.8)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
NMHR										
$PW(2)$	97	98	**100**	**100**	99	94	98	**100**	99	**100**
$PW(3)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99	**100**
$LN(0.8)$	82	86	87	80	88	86	87	88	**89**	82
$LN(1)$	97	98	**99**	**99**	**99**	98	**99**	**99**	**99**	**99**
$LN(1.5)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$IG(0.5)$	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$IG(1.5)$	73	78	77	61	78	79	78	76	**80**	65

in a handful of alternatives, making these two tests the closest competitors for $S_{n,5}$. The $S_{n,5}$ test also performs well when considering the global power when the standard exponential distribution, which has constant hazard rate, is the alternative, as shown in Figure 5.1.

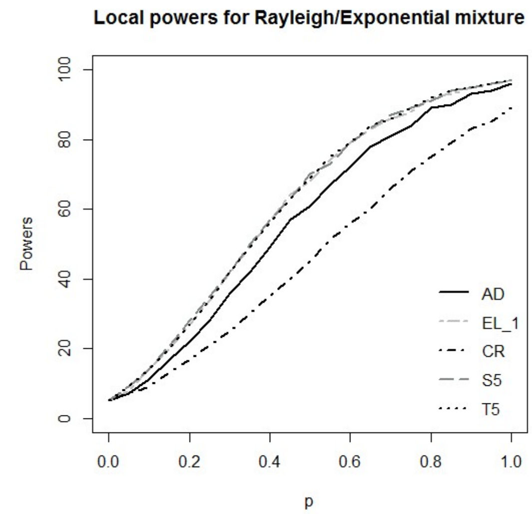

Fig. 5.1: Local powers for some of the tests over the entire range of mixture probabilities of the Rayleigh-exponential mixture distribution for n=20.

Next, we discuss the results obtained when the local powers are estimated using the mixture distribution consisting of the standard Rayleigh distribution and the gamma distribution with parameter $\theta = 2$. Here, the $EL_{n,1}$, KS_n and $R_{n,5}$ tests perform poorly; their powers compare very unfavourably with the other tests. The powers of these tests are marginally lower for almost all the mixture probabilities considered. The CR_n test completely outperforms all the other test for all values of p considered, yet, when considering the global power of the CR_n test with the $\Gamma(2)$ distribution as alternative, it is lower than a few of the other tests (see Figure 5.2). This implies that somewhere in the range of mixture probabilities ($p \in [0, 1]$) used for the alternative distributions, the power of the CR_n test achieves its maximum value for a lower value of p when compared to the other tests. Furthermore, we note that the powers of this test appear to decrease after achieving this maximum. Figure 5.2 illustrates this trend, it shows that the other tests start to outperform the CR_n test at $p = 0.75$, including the $S_{n,5}$ test.

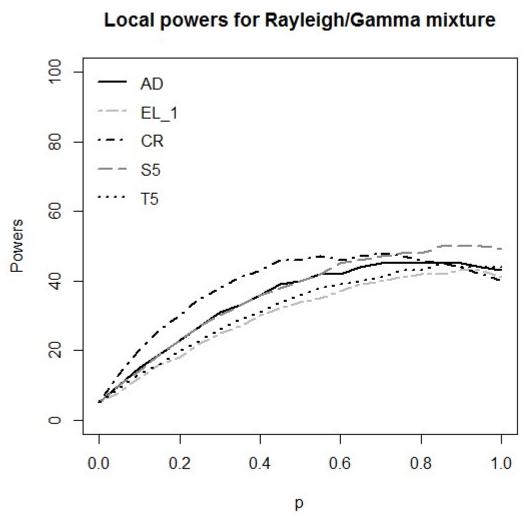

Fig. 5.2: Local powers for some of the tests over the entire range of mixture probabilities of the Rayleigh-gamma mixture distribution for n=20.

We will now consider the performance of the tests in general against all of the general alternative distributions listed in Table 5.1. These results are presented for the sample sizes $n = 20$ and $n = 30$ in Tables 5.4 and 5.5, respectively. From both Tables 5.4 and 5.5 we see that, in general, the powers of the KS_n and CR_n tests are lower for the majority of the alternatives considered and perform unfavourably in comparison to the other tests, for both sample sizes. On the other hand, the $EL_{n,1}$ and $R_{n,5}$ tests perform quite well for most of the alternatives and we find that the $S_{n,5}$ test outperforms the other tests, having the highest estimated power for the majority of the alternatives considered.

All tests considered perform quite well against the standard exponential distribution (which has a constant hazard rate) for both sample sizes.

Shifting our attention now to results associated with alternatives with increasing hazard rates, one finds from Tables 5.4 and 5.5, once again, that the KS_n and CR_n tests have lower powers for both sample sizes considered. For most of the alternatives in this category the $S_{n,5}$ test has the highest power, only being outperformed, or equaled, for a handful of these alternatives by $EL_{n,1}$ and $R_{n,5}$.

Moving our attention to the alternatives in the decreasing hazard rate category, we see that all the tests considered perform very well and, since there are such minor differences in the power performance between these tests, it is difficult to identify a single 'best' test for this set of alternatives. However, for the smaller sample size in Table 5.4, the KS_n test still shows powers that are slightly lower than the rest of the tests.

We now observe the results associated with the alternatives with non-monotone hazard rates in Tables 5.4 and 5.5. The $EL_{n,1}$ and $R_{n,5}$ tests seem to exhibit the lowest powers for a few of the alternatives, but have competitive powers for many of the other alternatives in this category. The tests that generally perform well are AD_n and $R_{n,1}$. The test that exhibits the highest power for the majority of the alternatives, for both sample sizes, is the $S_{n,5}$ test, making it the test that performs the best overall, closely followed by $R_{n,5}$.

To conclude, we provide a brief demonstration of how the choice of the tuning parameter, a, influences the powers of the two newly proposed tests. In order to visualise the behaviour of the powers for different values of a, Figures 5.3 and 5.4 present the powers for the $S_{n,a}$ and $R_{n,a}$ tests, respectively over a grid of a values and six different alternative distributions. These figures are also used to motivate the choice of a values included in the study.

The choice of $a = 1$ was made since it is the point where the powers for most of the alternative distributions start to plateau. The choice for $a = 5$ is due to the fact that it is the point where the powers for most of the alternative distributions reach their maximum value.

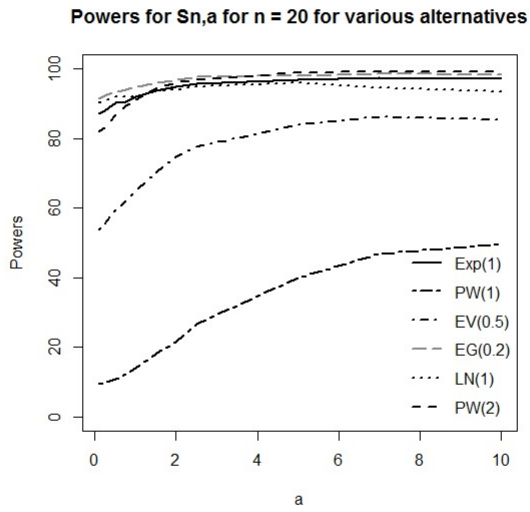

Fig. 5.3: Powers for $S_{n,a}$ for various alternatives.

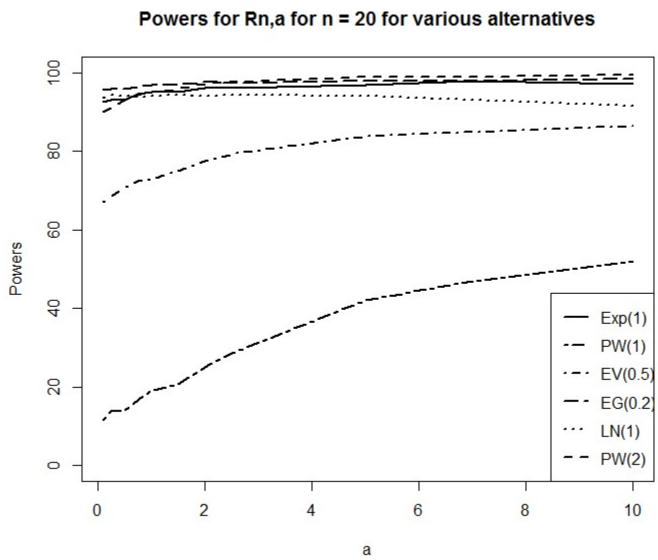

Fig. 5.4: Powers for $R_{n,a}$ for various alternatives.

In this dissertation two new goodness-of-fit test statistics specifically designed for the Rayleigh distribution were considered. The finite-sample performance of these newly suggested tests were studied via the use of a Monte Carlo simulation. From the results it is clear that this new tests are feasible when testing goodness-of-fit for the Rayleigh distribution. Not only are they feasible, they also outperform or equal competitor tests for the majority of the alternative distributions considered. The results that are obtained serves as evidence for the consistency of the test.

Appendix A

Motivation for the use of weight functions

Let

$$G = n\int_0^\infty (\frac{1}{n}\sum_{j=1}^{n}(Y_j - \frac{1}{Y_j})min(Y_j,t) - \frac{1}{n}\sum_{j=1}^{n} I(Y_j \leq t))^2 \mathrm{d}t.$$

In order to calculate G, one first needs the following:

$$\begin{aligned}
A &= \int_0^\infty \left\{\sum_{j=1}^{n}\left(Y_j - \frac{1}{Y_j}\right)\min(Y_j,t)\right\}^2 \mathrm{d}t \\
&= \int_0^\infty \left(\sum_{j=1}^{n}\left(Y_j - \frac{1}{Y_j}\right)^2 \{\min(Y_j,t)\}^2 + \sum\sum_{1\leq j<k\leq n}\left(Y_j - \frac{1}{Y_j}\right)\left(Y_k - \frac{1}{Y_k}\right)\min(Y_j,t)\min(Y_k,t)\right)\mathrm{d}t \\
&= \sum_{j=1}^{n}\left(Y_j - \frac{1}{Y_j}\right)^2 \int_0^\infty \{\min(Y_j,t)\}^2 \mathrm{d}t + 2\sum\sum_{1\leq j<k\leq n}\left(Y_j - \frac{1}{Y_j}\right)\left(Y_k - \frac{1}{Y_k}\right)\int_0^\infty \min(Y_j,t)\min(Y_k,t)\mathrm{d}t
\end{aligned}$$

Now,

$$\{\min(Y_j,t)\}^2 = \begin{cases} Y_j^2, & \text{if } Y_j \leq t, \\ t^2, & \text{if } Y_j > t. \end{cases}$$

Therefore,

$$\begin{aligned}
\int_0^\infty \{\min(Y_j,t)\}^2 &= \int_0^{Y_j} t^2 \mathrm{d}t + \int_{Y_j}^\infty Y_j^2 \mathrm{d}t \\
&= \frac{1}{3}Y_j^3 + Y_j^2 \lim_{s\to\infty} t|_{Y_j}^{s} \\
&= \infty.
\end{aligned}$$

Similarly,

$$\int_0^\infty \min(Y_j,t)\min(Y_k,t)\mathrm{d}t = \infty.$$

In the following calculation we use order statistics since we have $Y_{(j)} < Y_{(k)}$ if $j < k$. Thus,

$$\min(Y_{(j)}, t)\min(Y_{(k)}, t) = \begin{cases} t^2, & \text{if } 0 \leq t \leq Y_{(j)}, \\ Y_{(j)}t, & \text{if } Y_{(j)} < t \leq Y_{(k)}, \\ Y_{(j)}Y_{(k)}, & \text{if } Y_{(k)} \leq t. \end{cases}$$

Therefore,

$$\begin{aligned} \int_0^\infty \min(Y_{(j)}, t)\min(Y_{(k)}, t)\mathrm{d}t &= \int_0^{Y_{(j)}} t^2 \mathrm{d}t + \int_{Y_{(j)}}^{Y_(k)} Y_{(j)} t \mathrm{d}t + \int_{Y_{(k)}}^\infty Y_{(j)}Y_{(k)}\mathrm{d}t \\ &= \frac{1}{3}Y_{(j)}^3 + \frac{1}{2}Y_{(j)}\left(Y_{(k)}^2 - Y_{(j)}^2\right) + Y_{(j)}Y_{(k)} \lim_{s\to\infty} t|_{Y_{(k)}}^s \\ &= \infty. \end{aligned}$$

Thus to ensure that the integral in A exists one has to add an appropriate weight function in the integral, such as e^{-at}.

Appendix B

Derivation of a calculable form of our test statistic

To show the calculations in Chapter 4.2 we need to simplify the following two integrals

$$\frac{1}{n}\int_0^\infty \sum_{j=1}^{n}\left(\left\{Y_{(j)}-\frac{1}{Y_{(j)}}\right\}^2 \min\left\{Y_{(j)},t\right\}^2 - 2\left\{Y_{(j)}-\frac{1}{Y_{(j)}}\right\}\min\left\{Y_{(j)},t\right\}\mathbb{1}\left\{Y_{(j)}\le t\right\} \right.$$
$$\left. +\mathbb{1}\left\{Y_{(j)}\le t\right\}\right)e^{-at}\mathrm{d}t \tag{B.1}$$

and

$$\frac{2}{n}\int_0^\infty \sum_{1\le j<k\le n}\left(\left\{Y_{(j)}-\frac{1}{Y_{(j)}}\right\}\left\{Y_{(k)}-\frac{1}{Y_{(k)}}\right\}\min\left\{Y_{(j)},t\right\}\min\left\{Y_{(k)},t\right\}\right.$$
$$-\left\{Y_{(j)}-\frac{1}{Y_{(j)}}\right\}\min\left\{Y_{(j)},t\right\}\mathbb{1}\left\{Y_{(k)}\le t\right\}$$
$$-\left\{Y_{(k)}-\frac{1}{Y_{(k)}}\right\}\min\left\{Y_{(k)},t\right\}\mathbb{1}\left\{Y_{(j)}\le t\right\}$$
$$\left.+\mathbb{1}\left\{Y_{(j)}\le t\right\}\mathbb{1}\left\{Y_{(k)}\le t\right\}\right)e^{-at}\mathrm{d}t. \tag{B.2}$$

First, to derive an expression for equation (B.1). We have that

$$\{\min(Y_{(j)},t)\}^2 = \begin{cases} Y_{(j)}^2, & \text{if } Y_{(j)}\le t,\\ t^2, & \text{if } Y_{(j)}>t.\end{cases}$$

Therefore,

$$\int_0^\infty \{\min(Y_{(j)},t)\}^2 e^{-at}\mathrm{d}t = \int_0^{Y_{(j)}} t^2 e^{-at}\mathrm{d}t + \int_{Y_{(j)}}^\infty Y_{(j)}^2 e^{-at}\mathrm{d}t.$$

Similarly,

$$\min(Y_{(j)},t)\mathbb{1}\{Y_{(j)}\le t\} = \begin{cases} Y_{(j)}, & \text{if } Y_{(j)}\le t,\\ 0, & \text{if } Y_{(j)}>t,\end{cases}$$

which results in

$$\int_0^\infty \min(Y_{(j)},t)\mathbb{1}\{Y_{(j)}\le t\}e^{-at}\mathrm{d}t = \int_{Y_{(j)}}^\infty Y_{(j)}e^{-at}\mathrm{d}t,$$

and lastly

$$\int_0^\infty \mathbb{1}\{Y_{(j)} \leq t\}e^{-at}\mathrm{d}t = \int_{Y_{(j)}}^\infty e^{-at}\mathrm{d}t.$$

To calculate the interval in (B.2) we have $Y_{(j)} < Y_{(k)}$ since $j < k$ in the double sum. Therefore,

$$\min(Y_{(j)}, t)\min(Y_{(k)}, t) = \begin{cases} t^2, & \text{if } 0 \leq t \leq Y_{(j)} \\ Y_{(j)}t, & \text{if } Y_{(j)} < t \leq Y_{(k)} \\ Y_{(j)}Y_{(k)}, & \text{if } Y_{(k)} \leq t \end{cases},$$

which results in

$$\int_0^\infty \min(Y_{(j)}, t)\min(Y_{(k)}, t)e^{-at}\mathrm{d}t = \int_0^{Y_{(j)}} t^2e^{-at}\mathrm{d}t + \int_{Y_{(j)}}^{Y_{(k)}} Y_{(j)}te^{-at}\mathrm{d}t + \int_{Y_{(k)}}^\infty Y_{(j)}Y_{(k)}e^{-at}\mathrm{d}t.$$

Similarly, we have

$$\min(Y_{(j)}, t)\mathbb{1}\{Y_{(k)} \leq t\} = \begin{cases} 0, & \text{if } Y_{(k)} > t, \\ Y_{(j)}, & \text{if } Y_{(k)} \leq t, \end{cases}$$

and

$$\min(Y_{(k)}, t)\mathbb{1}\{Y_{(j)} \leq t\} = \begin{cases} 0, & \text{if } Y_{(j)} > t, \\ t, & \text{if } Y_{(j)} \leq t < Y_{(k)}, \\ Y_{(k)}, & \text{if } Y_{(k)} \leq t. \end{cases}.$$

This results in

$$\int_0^\infty \min(Y_{(j)}, t)\mathbb{1}\{Y_{(k)} \leq t\}e^{-at}\mathrm{d}t = \int_{Y_{(k)}}^\infty Y_{(j)}e^{-at}\mathrm{d}t,$$

and

$$\int_0^\infty \min(Y_{(k)}, t)\mathbb{1}\{Y_{(j)} \leq t\}e^{-at}\mathrm{d}t = \int_{Y_{(j)}}^{Y_{(k)}} te^{-at}\mathrm{d}t \int_{Y_{(k)}}^\infty Y_{(k)}e^{-at}\mathrm{d}t.$$

Lastly, we have

$$\begin{aligned} &\int_0^\infty \mathbb{1}\{Y_{(j)} \leq t\}\mathbb{1}\{Y_{(k)} \leq t\}e^{-at}\mathrm{d}t \\ &= \int_0^\infty \mathbb{1}\{Y_{(k)} \leq t\}e^{-at}\mathrm{d}t \\ &= \int_{Y(k)}^\infty e^{-at}\mathrm{d}t. \end{aligned}$$

Expressions for these integrals can now easily be found which results in

$$\frac{1}{n}\sum_{j=1}^{n}\left(-\frac{1}{a}e^{-aY_{(j)}}\left[\left\{Y_{(j)}-\frac{1}{Y_{(j)}}\right\}^2\left\{\frac{2}{a}Y_{(j)}+\frac{2}{a^2}\right\}+2Y_{(j)}^2-3\right]+\frac{2}{a^3}\left[Y_{(j)}^2-2+\frac{1}{Y_{(j)}^2}\right]\right)$$

$$+\frac{2}{n}\sum_{1\le j<k\le n}\left(\left\{Y_{(j)}-\frac{1}{Y_{(j)}}\right\}\left\{Y_{(k)}-\frac{1}{Y_{(k)}}\right\}\left(-\frac{1}{a}e^{-aY_{(j)}}\left\{\frac{1}{a}Y_{(j)}+\frac{2}{a^2}\right\}+\frac{2}{a^3}-\frac{Y_{(j)}}{a^2}e^{-aY_{(k)}}\right)\right.$$

$$+\left\{Y_{(j)}-\frac{1}{Y_{(j)}}\right\}\left\{-\frac{Y_{(j)}}{a}e^{-aY_{(k)}}\right\}$$

$$\left.+\left\{Y_{(k)}-\frac{1}{Y_{(k)}}\right\}\left\{\frac{1}{a^2}e^{-aY_{(k)}}-\frac{1}{a}e^{-aY_{(j)}}\left(Y_{(j)}+\frac{1}{a}\right)\right\}+\frac{1}{a}e^{-aY_{(k)}}\right).$$

Bibliography

Ahrari, V., Baratpour, S., Habibirad, A. & Fakoor, V. (2019). Goodness of fit tests for Rayleigh distribution based on quantiles, *Communications in Statistics-Simulation and Computation* pp. 1–17.

Ahsanullah, M. & Shakil, M. (2013). Characterizations of Rayleigh distribution based on order statistics and record values, *Bull. Malays. Math. Sci. Soc.(2)* **36**: 625–635.

Baratpour, B. S. & Khodadadi, F. (2012). A cumulative residual entropy characterization of the Rayleigh distribution and related goodness-of-fit test, *Journal of Statistical Research of Iran* **9**(2): 115–1294.

Baringhaus, L. & Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform, *Annals of the Institute of Statistical Mathematics* **43**(3): 551–564.

Best, D. J., Rayner, J. C. & Thas, O. (2010). Easily applied tests of fit for the Rayleigh distribution, *Sankhya B* **72**(2): 254–263.

Betsch, S. & Ebner, B. (2018a). Characterizations of continuous univariate probability distributions with applications to goodness-of-fit testing, *arXiv preprint arXiv:1810.06226* .

Betsch, S. & Ebner, B. (2018b). Testing normality via a distributional fixed point property in the Stein characterization, *TEST* pp. 1–34.

Betsch, S. & Ebner, B. (2019). A new characterization of the Gamma distribution and associated goodness-of-fit tests, *Metrika* **82**(7): 779–806.

Brummer, M., Mersereau, R., Eisner, R. & Lewine, R. (1993). Automatic detection of brain contours in MRI data sets, *IEEE transactions on medical imaging* **12**(2): 153–166.

Dyer, D. & Whisenand, C. (1973). Best linear unbiased estimator of the parameter of the Rayleigh distribution, *IEEE Transactions on Reliability* **22**(4): 229–231.

Goldstein, L. & Reinert, G. (1997). Stein's method and the zero bias transformation with application to simple random sampling, *The Annals of Applied Probability* **7**(4): 935–952.

Gulati, S. (2011). Goodness of fit test for the Rayleigh and the Laplace distributions, *International Journal of Applied Mathematics and Statistics* **24**: 74–85.

Hirano, K. (1986). *Rayleigh distributions*, Wiley, New York.

Jahanshahi, S. M. A., Rad, A. H. & Fakoor, V. (2016). A goodness-of-fit test for Rayleigh distribution based on Hellinger distance, *Annals of Data Science* **3**(4): 401–411.

Liebenberg, S. C. & Allison, J. S. (2018). A goodness-of-fit test for the Rayleigh distribution based on a lesser-known characterisation, *Annual Proceedings of the South African Statistical Association Conference*, Vol. 2018, South African Statistical Association (SASA), pp. 17–24.

Meintanis, S. G. (2008). A new approach of goodness-of-fit testing for exponentiated laws applied to the generalized Rayleigh distribution, *Computational Statistics & Data Analysis* **52**(5): 2496–2503.

Meintanis, S. G. & Iliopoulos, G. (2003). Tests of fit for the Rayleigh distribution based on the empirical Laplace transform, *Annals of the Institute of Statistical Mathematics* **55**(1): 137–151.

Nanda, A. (2010). Characterization of distributions through failure rate and mean residual life functions, *Statistics & Probability Letters* **80**(9–10): 752–755.

Polovko, A. (1968). *Fundamentals of reliability theory*, Academic Press.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rajan, J., Poot, D., Juntu, J. & Sijbers, J. (2010). Segmentation based noise variance estimation from background MRI data, *ICIAR 2010 Part I LNCS*, Vol. 6111, Springer-Verlag Berlin Heidelberg, pp. 62–70.

Rao, M., Chen, Y., Vemuri, B. & Wang, F. (2004). Cumulative residual entropy: A new measure of information, *IEEE Transactions on Information Theory* **50**(6): 1220–1228.

Rayleigh, F. R. S. (1880). Xii. on the resultant of a large number of vibrations of the same pitch and of arbitrary phase, *The London, Edinburgh, and Dublin Philisophical Magazine and Journal of Science* **10**(60): 73–78.

Safavinejad, M., Jomhoori, S. & Alizadeh Noughabi, H. (2015). A density-based empirical likelihood ratio goodness-of-fit test for the Rayleigh distribution and power comparison, *Journal of Statistical Computation and Simulation* **85**(16): 3322–3334.

Sijbers, J., Poot, D., den Dekker, A. & Pintjens, W. (2007). Automatic estimation of the noise variance from the histogram of a magnetic resonance image, *Physics in Medicine and Biology* **52**: 1335–1348.

Stein, C. M. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, The Regents of the University of California.

Zamanzade, E. & Mahdizadeh, M. (2017). Goodness of fit tests for Rayleigh distribution based on Phi-divergence, *Revista Colombiana de Estadística* **40**(2): 279–290.