

On new tests of fit for the geometric distribution based on a characterisation

M Le Roux

 **orcid.org 0000-0002-9612-2672**

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

Supervisor:	Dr SC Liebenberg
Co-Supervisor:	Prof IJH Visagie

Abstract

The *geometric* distribution is used in multiple research fields across various disciplines. Therefore, it is important to be able to test the assumption of whether a set of observed data arises from this distribution. The purpose of this dissertation is to review the literature on existing goodness-of-fit tests for the *geometric* distribution and to develop new goodness-of-fit tests based on a mean residual life characterization of the *geometric* distribution. The first test is a Kolmogorov-Smirnov type test, and the second and third tests are Cramér-von Mises type tests, with empirical and parametric weight functions, respectively. We compare the performance of these tests with existing goodness-of-fit tests for the *geometric* distribution in a Monte Carlo simulation study. These tests are compared considering several alternative distributions including mixture distributions to analyze the estimated powers. Lastly, the tests are applied to observed data. Considering the power performance of our tests, it is clear that these tests match or outperform the majority of existing tests considered in this study against the alternatives considered.

Key words: *Characterisation, Goodness-of-fit, Geometric distribution*

Acknowledgements

Have I not commanded you? Be strong and courageous. Do not be frightened, and do not be dismayed, for the Lord your God is with you wherever you go. -Joshua 1:9

I wish to convey my heartfelt appreciation and gratitude to the individuals mentioned below:

- Dr. Shawn Liebenberg and Prof. Jaco Visagie, my supervisors, for their continuous patience, insight and guidance. Thank you for sticking with me to the end.
- My parents and brother without whom I would not be where I am today. Thank you for your endless love and support. Especially my parents, for giving me such an outstanding education that has brought me this far.
- Ruan, the love of my life, thank you for your unwavering support, motivation and unconditional love. Thank you for struggling with me, for being happy with me and living life to the fullest with me.

Contents

Abstract	2
Acknowledgements	3
1 Introduction	8
1.1 Overview	8
1.2 Objectives	11
1.3 Dissertation outline	11
2 Reliability theory and residual life	13
2.1 Introduction	13
2.2 Reliability theory	14
2.3 Estimation procedures for mean residual life	16
2.4 Discrete failure time	21
3 Characterisations and existing tests	23
3.1 Introduction	23
3.2 Properties of the geometric distribution	24
3.3 Characterisations of the geometric distribution	25
3.3.1 Characterisations based on the characterisation function . . .	26
3.3.2 A characterisation based on conditional expectation	28
3.3.3 A characterisation based on independence	29

3.3.4	A characterisation based on moments of truncated life	30
3.3.5	A characterisation based on order statistics	31
3.3.6	A characterisation based on weak records	33
3.3.7	A characterisation based on conditional variance	34
3.3.8	A characterisation based on the geometric renewal process . .	36
3.4	Tests for the geometric distribution	37
3.4.1	A test statistic based on linear regression of order statistics .	37
3.4.2	A test based on the probability generating function	39
3.4.3	A test based on the empirical distribution function	40
3.4.4	A test based on characterisation identities of Stein's method .	41
3.4.5	A test based on the integrated distribution function	43
3.4.6	A test based on a differential equation	45
3.4.7	A test based on the empirical characteristic functions	48
4	New tests of fit for the geometric distribution	50
4.1	Introduction	50
4.2	New tests of fit	50
5	Simulations and results	57
5.1	Introduction	57
5.2	Simulation setting	57
5.3	Simulation results	60
5.4	Observed Data Application	63
6	Conclusion and future research	74
6.1	Introduction	74
6.2	Concluding remarks and future research	75
	Appendix A	79
A.1	Chapter 3	79

A.2 Chapter 4 80

A.3 Chapter 6 81

List of Tables

5.1	Probability mass functions for our choice of alternative distributions	59
5.2	Notation and section reference for the choice of test statistics	61
5.3	Estimated powers for alternative distributions of sample size $n = 20$.	64
5.4	Estimated powers for alternative distributions of sample size $n = 20$.	65
5.5	Estimated powers for mixed distributions of sample size $n = 20$	66
5.6	Estimated powers for alternative distributions of sample size $n = 40$.	67
5.7	Estimated powers for alternative distributions of sample size $n = 40$.	68
5.8	Estimated powers for mixed distributions of sample size $n = 40$	69
5.9	Estimated powers for alternative distributions of sample size $n = 60$.	70
5.10	Estimated powers for alternative distributions of sample size $n = 60$.	71
5.11	Estimated powers for mixed distributions of sample size $n = 60$	72
5.12	Bootstrap <i>p-values</i> for earthquake data	73
5.13	Bootstrap <i>p-values</i> for discentric data 1	73
5.14	Bootstrap <i>p-values</i> for discentric data 2	73
5.15	Bootstrap <i>p-values</i> for birth data - smokers	73

Chapter 1

Introduction

1.1 Overview

The *geometric* distribution can be considered the discrete analogue of the widely used exponential distribution seeing that the exponential distribution models time between events in continuous systems and the *geometric* distribution models the number of trials to achieve the first success in a sequence of independent Bernoulli trials. The *geometric* distribution is used in multiple research fields across a variety of disciplines. For instance, in a recent paper, Sorsa, Siikonen, Kuusinen and Hakonen (2021) suggested that the *geometric* distribution can be used to develop elevator traffic simulations when designing multi-story buildings. More specifically, Sorsa et al. (2021) studied the size of social groups amidst elevator passengers in three types of buildings over four countries. The authors found that the *geometric* distribution provided a better fit to the data over all building types compared to the logarithmic distribution.

Evans (1967) studied the theory of queues and found that the limiting distribution of the state of the system is in a variety of cases *geometric*. Evans' study was focused on contributing to queuing theory as the basis of production. The *geometric* distribution

also finds application in frequency distributions of vertebrate dispersal distances (see Buechner (1987) and Miller and Carroll (1989)).

Coşar and Bellotto (2020) proposed a new re-identification method for domestic service robots' systems, constructed based on the *geometric* distribution. Domestic service robots have an increasingly important role in present-day life. The human re-identification feature of these service robots provides help with tasks such as monitoring and assisting the elderly. The ability to perform this test can be complicated due to differing human postures and poor lighting conditions. Therefore, it is important for these robots to have excellent re-identification systems in order to complete their tasks successfully. Coşar and Bellotto (2020) suggested using thermal images and compiled a thermal dictionary for each person in the study. A classifier was trained using this thermal dictionary which incorporates the *geometric* distribution. Coşar and Bellotto (2020) showed the good performance of their re-identification method and proved that it can be used for service robots assisting the elderly in daily life.

Zhang, Tang, Huang and Lee (2021) defined arc flash as a plasma discharge caused by a current flowing through a non-conductive medium which results in excessive heat exposure in the form of rapid energy release. Zhang et al. (2021) developed a new pre-emptive arc detection and location method using *geometric* distribution analysis on magnetic flux and current density. The proposed method was proved to have a good ability to detect the early stages of the arc current.

Radio frequency identification (RFID) is becoming more popular in the use of identifying and tracking objects in the manufacturing process because of its low cost, easy implementation and durability. Bueno-Delgado, Ferrero, Gandino, Pavon-Marino and Rebaudengo (2012) studied RFID systems generally consisting of numerous tags and at least one reader. The reader uses electromagnetic waves to communicate with the tags. However, these RDIF systems have several reader collision problems.

Bueno-Delgado et al. (2012) proposed the *geometric* distribution reader anti-collision (GDRA) method that uses the Sift *geometric* probability distribution function to minimize reader collision problems. The authors found that GDRA outperforms single-channel mechanisms and multichannel mechanisms.

Lastly, Elderton and Wood (1945) used the geometric distribution to model batting scores in cricket. The authors considered four approaches to their study. A) Aggregating the cumulative scores of large groups of batsmen across many innings, B) Compiling the cumulative scores of 61 batsmen throughout their careers (i.e., batsmen that were considered "great" during the period 1890 - 1939), C) Scoring the performance of batsmen, who batted at Numbers 1 to 7 in the batting order for every match played in England during the period 1939 - 1939, D) Considering all occurrences with scores 100 or above by English batsmen during the period 1933 - 1939. The authors organised the outcomes of each method into a comprehensive table and upon careful consideration they concluded that this table appears to provide strong evidence that a a geometric progression is mainly suitable to cricket scores.

Following the discussion above it is evident that there are many cases where it is necessary to test if the data has originated from the *geometric* distribution. Goodness-of-fit tests are developed to test this hypothesis and the study of goodness-of-fit tests for the *geometric* distribution has been undertaken by several authors, see Jiménez-Gamero and Alba-Fernández (2021) and Klar (1999). However, this list is still sparse when compared to similarly popular distributions within the discrete data framework such as the binomial distribution (see Tarone (1979), Beltrán-Beltrán and O'Reilly (2019) and ?). In this study, we suggest new tests for the *geometric* distribution based on the estimate mean residual life (MRL).

1.2 Objectives

This thesis aims to develop new goodness-of-fit tests for the *geometric* distribution and perform a Monte Carlo study in order to test the performance of the new goodness-of-fit test in comparison to the performance of existing goodness-of-fit tests for the *geometric* distribution.

The main objectives are as follows:

- Provide an overview of the *geometric* distribution with emphasis on its characterisations and existing goodness-of-fit tests.
- Introduce new goodness-of-fit tests for the *geometric* distribution.
- Perform a Monte Carlo simulation study to compare the empirical powers of the newly proposed goodness-of-fit tests to those of existing tests.
- Apply the new goodness-of-fit tests to observed data obtained from Jiménez-Gamero and Alba-Fernández (2021).

1.3 Dissertation outline

In Chapter 2 we discuss reliability theory and the properties of the MRL function since our goodness-of-fit tests are based on the MRL. We also include estimations of the MRL in the continuous case by various authors and discuss the MRL in terms of discrete failure times.

Chapter 3 contains the literature study on the *geometric* distribution with topics on the *geometric* distribution's characterisations, its goodness-of-fit tests and its relation to other well-known discrete distributions.

In Chapter 4 we describe three new goodness-of-fit tests for the *geometric* distribution and outline the construction of the new goodness-of-fit tests as well as the

mathematical basis that forms the motivation of these tests.

In Chapter 5 we conduct a Monte Carlo investigation by comparing the finite-sample performance of the newly introduced goodness-of-fit tests for the *geometric* distribution to that of the existing goodness-of-fit tests of the *geometric* distribution. We will also apply the newly introduced goodness-of-fit tests for the *geometric* distribution to observed data sets to illustrate how to use these tests in practical application.

Chapter 6 is the final conclusion and a discussion on future research.

Chapter 2

Reliability theory and residual life

2.1 Introduction

In the current literature, research on the estimation of the mean residual life (MRL) in the discrete case is notably scarce. Most existing research on estimating the MRL primarily focuses on the continuous case. Therefore, it is unclear how the estimators will perform when applied to discrete data. This chapter will introduce the concept of MRL and related concepts for continuous and discrete cases. In addition, some of the estimation techniques for the continuous case will be briefly outlined. Ideally, these methods could be directly applied to discrete data and developed into tests for the *geometric* distribution similar to the approach used in Rao Jammalamadaka and Taufer (2006) for the exponential distribution. The newly suggested goodness-of-fit tests for the *geometric* distribution will circumvent the direct estimation of the MRL by following the approach of Baringhaus and Henze (1992). Subsequently, we conducted a comprehensive numerical study, presented in Chapter 5, to evaluate the effectiveness of our newly developed test for the *geometric* distribution. A detailed exposition of our methodology is presented in Chapter 4. This chapter therefore serves a twofold purpose; to introduce the concept and properties of MRL on which the new tests are built and to pave the way for future research.

The outline of this chapter follows. In Section 2.2 we define the MRL function in terms of continuous failure times and give several existing methods to estimate the MRL function. In Section 2.3 we define the MRL function in terms of discrete failure times.

2.2 Reliability theory

Reliability theory is a branch of engineering, statistics, and mathematics that examines the ability of a system or component to fulfil its intended function in a consistent and dependable manner over a specified time period. This discipline is concerned with calculating the probability of failure and the analysis of system performance and its underlying components. For further details see Chinnam and Baruah (2004), Zhao, Makis, Chen and Li (2017) and Lai and Xie (2006).

In the context of reliability theory, there is interest in reliability measures that aid in describing the performance and dependability of a system or component. Some commonly calculated functions used in reliability theory are (i) Hazard rate (ii) Cumulative hazard rate and (iii) Mean residual life. These are described first for the continuous and then the discrete case.

To define the MRL function we start by denoting X as the lifetime of an object, i.e., the time to the first failure. Let F be the distribution function of X . We use f in order to denote the density of X in the continuous case. With a slight abuse of notation, we denote the probability mass function (PMF) of X by f in the discrete case as well.

The survival function, S , of a random variable X is,

$$S(t) = \mathbb{P}(X > t) = 1 - F(t). \quad (2.1)$$

Lai and Xie (2006) defined the hazard rate, i.e., the instantaneous rate of failure as

$$h(t) = \frac{f(t)}{S(t)},$$

for t such that $S(t) > 0$.

The cumulative hazard rate, H , can then be defined as

$$\begin{aligned} H(t) &= -\log S(t) \\ &= \int_0^t h(x) dx. \end{aligned}$$

Singpurwalla (2006) interpreted H in such a way that it gives insight into the relationship between the hazard rate and the cumulative hazard rate. We therefore have that,

$$\mathbb{P}(T > t) = e^{-H(t)}, t \geq 0.$$

The above equation is particularly useful since H lacks interpretation on its own. Furthermore, it should be noted that $e^{-H(t)}$ is the survival function of a random variable with standard exponential distribution, evaluated at $H(t)$.

Lai and Xie (2006) defined $S(x|t)$ as the conditional probability that a unit of age t will survive for an additional x units of time, also known as the residual lifetime,

$$S(x|t) = \frac{S(t+x)}{S(t)}, \quad x, t \geq 0.$$

It then follows that the MRL, at age t , is $m(t) = \mathbb{E}(X - t | X > t)$ also written as,

$$\begin{aligned} m(t) &= \frac{\int_t^\infty S(x) dx}{S(t)} \\ &= \int_t^\infty S(x|t) dx. \end{aligned} \tag{2.2}$$

MRL describes the expected remaining lifespan of a system or component. It can be conceptualised as the average amount of time that a system or component can be expected to last before it fails, given that it has already survived up to a certain point in time.

A function can be defined as an MRL function if it has the following properties as identified by Swartz (1973):

1. $m(t) > 0, \forall t > 0$,
2. $m(0) = \mathbb{E}(X)$,
3. If $S(t)$ is continuous, then $m'(t) \geq -1$,
4. $\int_0^\infty \frac{1}{m(t)} dt$ diverges,
5. $S(t) = \frac{m(0)}{m(t)} \exp \left\{ - \int_0^t \frac{1}{m(s)} ds \right\}$,

where $\mathbb{E}(x) = \tau$ is the finite first moment.

A detailed treatment of the MRL function can be found in Guess and Proschan (1988).

2.3 Estimation procedures for mean residual life

In the following section, we will give several existing methods to estimate the MRL function in the continuous case. Each section is named according to the authors who proposed the estimation method.

Abdous and Berred (2005)

Abdous and Berred (2005) defined a natural estimate for m as its empirical version, $\hat{m}$. Let $X_1, X_2, \dots, X_n$ be a sample from F , then $\hat{m}$ is traditionally defined to be,

$$\hat{m}(t) = \frac{\int_t^\infty S_n(y) dy}{S_n(t)} \mathbb{1}_{[X_{(n)} > t]}$$

with $X_{(n)} = \max_{1 \leq i \leq n} X_i$ and $S_n(t)$ the empirical survival function defined as,

$$S_n(t) = 1 - F_n(t). \quad (2.3)$$

F_n is then expressed as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t}. \quad (2.4)$$

with $\mathbb{1}(\cdot)$ the indicator function.

The authors however opted to use kernel density estimation as a smoothing method for m .

Abdous and Berred (2005) suggested this local linear fitting technique to estimate the function by denoting $K(\cdot)$ as the arbitrary probability density function and its corresponding survival function as $\mathbb{K}(t) = \int_t^\infty K(u) du$. The authors move on to define $\lambda = \lambda_n > 0$ to be a sequence of smoothing parameters such that $\lambda_n \rightarrow 0$ as $n \rightarrow \infty$. Now, because m is a ratio of two functions, one option for an estimate is,

$$\hat{m}(t) = \frac{\int_t^\infty \sum_{i=1}^n \mathbb{K}\left(\frac{u-X_i}{h}\right) du}{\sum_{i=1}^n \mathbb{K}\left(\frac{t-X_i}{h}\right)}.$$

Alternatively, the convolution of $\hat{m}$ can be taken with $K_h(\cdot) = \frac{K(\frac{\cdot}{h})}{h}$ and the MRL

estimator can be redefined to be,

$$\hat{m}(t) = \int_{-\infty}^{+\infty} K_h(t-u) \hat{m}(u) du.$$

The authors deal with the effect of boundary bias by introducing an estimate based on local polynomial fitting. This estimate seeks to minimise,

$$\int_{-\infty}^{U_F} K_h(y-t) [\hat{m}(y) - a_0 - a_1(y-t)]^2 dy,$$

where U_F is known, K denotes a symmetric kernel density function supported on $[-1, 1]$ and, $\hat{a}_{0n}$ and $\hat{a}_{1n}$, are the coefficients of the polynomial. The functions $\hat{a}_{0n}$ and $\hat{a}_{1n}$ are defined in Abdous and Berred (2005).

Chaubey and Sen (1999)

Yang (1978) suggested an estimation method for m based on the empirical distribution function (EDF) F_n ,

$$\hat{m}(t) = \begin{cases} \frac{1}{n-k} \sum_{i=k+1}^n (X_{(i)} - t) & \text{for } X_{(k)} \leq t < X_{(k+1)}, k < n, \\ 0, & \text{for } t \geq X_{(n)}, \end{cases} \quad (2.5)$$

where $(0 =) X_{(0)} \leq X_{(1)} \leq \dots \leq X_{(n)} < X_{(n+1)} (= +\infty)$ denotes the ordered random variables associated with the observed random variables $X_1, X_2, \dots, X_n$.

A smooth estimator for m is suggested by Chaubey and Sen (1999). Recall the empirical survival function defined in (2.3) and the EDF defined in (2.4) written in the form below by Chaubey and Sen (1999):

$$S_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i > t\}, t \in \mathbb{R}^+,$$

which results in $S_n(t) = \frac{n-k}{n}$ for $X_{(k)} \leq t < X_{(k+1)}$, $k = 0, 1, 2, \dots, n$ with X_i , $0 \leq i \leq (n+1)$ the order statistics.

A smooth estimator for $S(\cdot)$ then takes the form

$$\tilde{S}_n(t) = \sum_{k=0}^n W_{nk}(t\lambda_n) S_n\left(\frac{k}{\lambda_n}\right), t \in \mathbb{R}^+,$$

where $\lambda_n > 0$ is chosen such that $\lambda_n \rightarrow \infty$ but $n^{-1}\lambda_n \rightarrow 0$ as $n \rightarrow \infty$ for every $t \in \mathbb{R}^+$ and $n \geq 1$. Lastly, Chaubey and Sen (1999) defined W_{nk} as a sequence of non-negative weight functions. The density, cumulative hazard and hazard functions can now be defined in terms of $\tilde{S}_n$;

$$\begin{aligned} \tilde{f}_n(t) &= \lambda_n \{ \tilde{S}_n(t) [1 - w_{nn}(t\lambda_n)] - \sum_{k=0}^n w_{nk}(t\lambda_n) S_n((k+1)/\lambda_n) \}, \\ \tilde{H}_n(t) &= \log \tilde{S}_n(t), \\ \tilde{h}_n(t) &= \tilde{f}_n(t) / \tilde{S}_n(t). \end{aligned}$$

The choice of λ_n is important for the asymptotic properties for $\tilde{S}_n(t)$, $\tilde{f}_n(t)$, $\tilde{H}_n(t)$ and $\tilde{h}_n(t)$.

Ultimately, $m(t)$ is estimated by replacing $S_n(t)$ with $\tilde{S}_n(t)$ in (2.5) resulting in,

$$\tilde{m}_n(t) = \frac{1}{\tilde{S}_n(t)} \sum_{k=1}^n S_n\left(\frac{n-k}{\lambda_n}\right) \int_t^\infty \mathcal{W}_{n(n-k)}(\lambda_n x) dx.$$

Rao Jammalamadaka and Taufer (2006)

Rao Jammalamadaka and Taufer (2006) based their estimation on the characterisation that $\mathbb{E}(X - t | X > t)$ is a constant for $t \in [0, \infty)$ if and only if X is exponential. This characterisation is realised by Shanbhag (1970) and will be formally stated in Chapter 3, Section 3.3.2.

The characterisation by Shanbhag (1970) ultimately states that,

$$m(t) = \theta, \quad \forall t > 0$$

if and only if F is exponential with mean θ . Therefore, $m(t)$ can be shown to be equivalent to,

$$\begin{aligned} \int_t^\infty S(x)dx &= S(t)\theta, \quad \forall t > 0 \quad \text{or} \\ \mathbb{E}[\min(X, t)] &= F(t)\theta, \quad \forall t > 0. \end{aligned} \tag{2.6}$$

To estimate the MRL, Rao Jammalamadaka and Taufer (2006) defines a sample MRL for $k+1, k+2, \dots$ as,

$$\begin{aligned} \bar{X}_{>k} &= \frac{1}{n-k+1} \sum_{i=k+1}^{n+1} (X_{(i)} - X_{(k)}) \\ &= \frac{1}{n-k+1} \sum_{i=k+1}^{n+1} Y_i, \end{aligned} \tag{2.7}$$

where

$$Y_i = (n-i+2) (X_{(i)} - X_{(i-1)}), \quad i = 1, \dots, n+1,$$

are the normalised spacings and $\bar{X}$ is the sample mean. The authors use the estimator in (2.7) for the MRL as well as the characterisation by Shanbhag (1970) to construct a goodness-of-fit test for the exponential distribution.

The approach by Rao Jammalamadaka and Taufer (2006) follows similar lines than the approach by Baringhaus and Henze (1992) which serves as starting point for the newly proposed test outlined in Chapter 4.

2.4 Discrete failure time

It is important to find a lifetime distribution that can accurately describe discrete failure data. Lai and Xie (2006) gave examples of common situations that are described by discrete failure data:

- Field failure reports are routinely gathered on a regular basis, both weekly and monthly. These reports typically focus on the count of failures, without detailed information about when these failures occur.
- In some cases, equipment undergoes repeated cycles of operation, and researchers monitor the number of successful cycles completed before a failure event. For instance, a common example is a photocopier, where the total number of copies it can produce represents its operational life. Similarly, the number of on/off cycles of a switch before it malfunctions is another instance of this scenario.
- Additionally, researchers often employ a process of discretisation or grouping when dealing with continuous data to simplify their analysis and draw meaningful conclusions from the data.

Suppose X is a random variable defined on $\mathcal{N}^+ = \{1, 2, \dots\}$. Suppose further that X represents the discrete lifetime of a component with the probability of failure occurring at time t as,

$$f(t) = \mathbb{P}\{X = t\}, t = 1, 2, \dots$$

In the discrete case, the survival function of X is,

$$\begin{aligned} S(t) &= \mathbb{P}(X > t) \\ &= \sum_{j=t+1}^{\infty} f(j), \quad t = 1, 2, \dots \end{aligned}$$

Analogous to the continuous case Lai and Xie (2006) defined the failure rate, h ,

$$\begin{aligned} h(t) &= \mathbb{P}(X = t \mid X \geq t) \\ &= \frac{f(t)}{S(t-1)} \\ &= \frac{S(t-1) - S(t)}{S(t-1)}. \end{aligned}$$

The discrete failure rate function gives the conditional probability of the failure of a device at time t , given that it has not failed by $t-1$. Shaked, Shanthikumar and Valdez-Torres (1995) defined the conditions for a sequence $\{h(t), t > 1\}$ to be a failure rate:

- (a) $h(k) = 1$ for some $k \in \mathcal{N}^+$, and $h(t) = 1$ for $t > k$.
- (b) $0 \leq h(t) < 1$ for $t \in \mathcal{N}^+$ and $\sum_{i=1}^{\infty} h(i) = \infty$.

It then follows that the MRL at time t can be defined comparably to equation (2.2). That is m can be written as

$$\begin{aligned} m(t) &= \mathbb{E}(X - t \mid X \geq t) = \mathbb{E}(X \mid X \geq t) - t \\ &= \frac{\int_t^{\infty} S(t) dt}{S(t-1)}. \end{aligned}$$

where $\mathbb{E}(X)$ is the expected value of X . It should be noted that similar to the continuous case, the discrete MRL uniquely determines a discrete distribution.

We are not aware of research on the estimation of MRL in the discrete case at the time this analysis was undertaken. However, the MRL estimation techniques should be applicable to the discrete case with some limitations or modifications. A future study could be conducted to investigate this application.

Chapter 3

Characterisations and existing tests

3.1 Introduction

The purpose of this chapter is to explore the properties, characterisations, and existing goodness-of-fit tests of the *geometric* distribution. In Section 3.2, we delve into the properties of this distribution, while Section 3.3 focuses on its characterisations. Section 3.4 provides an overview of the current goodness-of-fit tests available for the *geometric* distribution.

In this chapter, we address two distinct parameterisations for the *geometric* distribution. Due to the absence of a consensus on which parameterisation to use in the literature, we use both in this dissertation. The characterisations presented in this chapter are formulated based on either one of these parameterisations. The selection of a specific parameter definition for each characterisation depends on the ease of its derivation within that particular context of the *geometric* distribution. In some cases, the characterisation is simple for one parameter definition but becomes complicated for the other parameter definition.

3.2 Properties of the geometric distribution

The *geometric* distribution can be defined as the number of trials up to and including the first success. Let X be a *geometric* random variable with probability of success on $(0, 1)$. The PMF for this case can be expressed as:

$$p(j) = \mathbb{P}(X = j) = p(1 - p)^{j-1} \text{ for } j = 1, 2, 3, \dots \quad (3.1)$$

The *geometric* distribution can also be defined as the number of failures before the first success. The PMF can be expressed as

$$p(j) = \mathbb{P}(X = j) = p(1 - p)^j \text{ for } j = 0, 1, 2, 3, \dots \quad (3.2)$$

It is further noted, that the *geometric* distribution is a special case of the negative binomial distribution with the PMF of the negative binomial distribution defined as

$$\mathbb{P}(X = j) = \binom{j-1}{r-1} p^r (1 - p)^{j-r},$$

where X denotes the total number of trials needed to achieve r successes ?. If $r = 1$ then the negative binomial distribution reduces to the *geometric* distribution defined in (3.1). The PMF of the negative binomial distribution can also be defined as,

$$\mathbb{P}(X = j) = \binom{j+r-1}{r-1} p^r (1 - p)^j.$$

If $r = 1$ then the negative binomial distribution reduces to the *geometric* distribution defined in (3.1). The *geometric* distribution has a unique property called the "non-aging" property or the "Markovian" property. This property characterises the *geometric* distribution among all distributions restricted to the non-negative integers.

This property can be defined as,

$$\mathbb{P}[X = x + j | X \geq j] = \mathbb{P}[X = x],$$

with $j \in 1, 2, \dots$ or $j \in 0, 1, \dots$ depending on the parameterisation used.

3.3 Characterisations of the geometric distribution

In this section, we give some of the characterisations of the *geometric* distribution. A characterisation of a probability distribution states that it is the only probability distribution that satisfies the specified conditions of the characterisation theorem. Characterisations are thus useful instruments in developing goodness-of-fit tests because if the observed data do not fit the conditions specified by the characterisation theorem we conclude that the distribution does not fit the data and we need to consider alternative distributions when modelling the data.

The cumulative distribution function of the *geometric* distribution in (3.1) is

$$F(j) = \mathbb{P}(X \leq j) = 1 - (1 - p)^j, \quad j \in 1, 2, \dots,$$

and the cumulative distribution function of the *geometric* distribution in (3.2) is defined as follows,

$$F(j) = 1 - (1 - p)^{j+1}, \quad j \in 0, 1, \dots$$

Further properties of the *geometric* distributions defined in (3.1) and (3.2) are as follows: For (3.1) the moment-generating function of the *geometric* distribution is

defined as,

$$M_X(t) = \mathbb{E}[e^{tX}] = \begin{cases} \frac{pe^t}{1-(1-p)e^t}, & \text{for } t \leq -\ln(1-p), \\ \frac{p}{1-(1-p)e^t}, & \text{for } t > -\ln(1-p). \end{cases}$$

For (3.2) the moment-generating function of the *geometric* distribution is defined as,

$$M_X(t) = \mathbb{E}[e^{tX}] = \frac{p}{1-(1-p)e^t}, \text{ for } t > -\ln(1-p).$$

Furthermore, the maximum likelihood estimator of p for (3.1) can be defined as,

$$\hat{p} = \frac{1}{\bar{j}},$$

while, the maximum likelihood estimator of p for (3.2) is given by,

$$\hat{p} = \frac{1}{\bar{j} + 1}.$$

3.3.1 Characterisations based on the characterisation function

The characteristic function of the *geometric* distribution defined in (3.1) can be expressed as follows

$$\phi_x(t) = \mathbb{E}[e^{itX}] = \frac{pe^{it}}{1-(1-p)e^{it}},$$

where i is the imaginary unit and $t \in \mathbb{R}$. The characteristic function of the *geometric* distribution as defined in (3.2) can then be expressed as follows

$$\phi_x(t) = \mathbb{E}[e^{itX}] = \frac{p}{1-(1-p)e^{it}}.$$

Maiti and Biswas (2007)

The characteristic function of a *geometric* distributed random variable, X , with distribution function as defined in (3.2) is

$$\begin{aligned}\phi(t) &= \mathbb{E}(\exp(itX)) \\ &= (1 - \theta) \left[\sum_{j=0}^{\infty} \theta^j \cos jt + i \sum_{j=0}^{\infty} \theta^j \sin jt \right] \\ &= C(t) + iS(t),\end{aligned}$$

where $C(t) = \mathbb{E}(\cos tX)$ and $iS(t) = \mathbb{E}(\sin tX)$ which are the real and imaginary terms of $\phi(t)$. Meintanis and Iliopoulos (2003) illustrated that for an exponential distribution the ratio $\frac{S(t)}{C(t)}$ is linear in t . Maiti and Biswas (2007) then based their characterisation on $\frac{S(t)}{C(t)}$ for the *geometric* distribution in the discrete setting.

Theorem 3.1. *Among all distributions of non-negative integer-valued random variables, the geometric distribution is the only one for which*

$$S(t) = \frac{\theta \sin t}{1 - \theta \cos t} C(t) \quad \text{for all } t.$$

See proof of the theorem in Maiti and Biswas (2007).

Paulson and Uppuluri (1972)

Paulson and Uppuluri (1972) based their characterisation on the *geometric* distribution as defined in (3.2).

Theorem 3.2. *Assume that $\psi(t)$ and $\phi(t)$ are characteristic functions and assume that V is a random variable with distribution function $G(v)$ then the characteristic-functional equation*

$$\phi(t) = \psi(t) \mathbb{E}[\phi(tV)],$$

characterises the geometric distribution, if and only if, $\psi(t)$ is the characteristic function of a geometric distribution with $G(v)$ such that $P[V = 0] = a$ and $P[V = 1] = b$ where $a + b = 1$ and $0 < a \leq 1$,

See proof in Paulson and Uppuluri (1972).

3.3.2 A characterisation based on conditional expectation

Recall that $\mathbb{E}(X | Y = y)$ is defined as follows: If X and Y are discrete random variables, the conditional expectation of X given Y is

$$\begin{aligned}\mathbb{E}(X | Y = y) &= \sum_x x \mathbb{P}(X = x | Y = y) \\ &= \sum_x x \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)},\end{aligned}$$

where $\mathbb{P}(X = x, Y = y)$ is the joint PMF of X and Y .

Osaki and Li (1988)

Consider the following characterisation for the negative binomial distribution.

Theorem 3.3. *If X is a positive discrete random variable with PMF, $\mathbb{P}(X)$ and mean μ . Then X is negative binomial distributed with PMF, $\mathbb{P}(X; \theta, r)$ if and only if*

$$\mathbb{E}(X | X > t) = \mu + (t + 1 - r) \frac{h(t + 1)}{\theta},$$

for all integers $t \geq r - 1$, where $\mu = \frac{r}{\theta}$ and h the hazard rate.

By setting $r = 1$, Theorem 3.3 becomes a characterisation for the *geometric* distribution as expressed in (3.1), given by Shanbhag (1970) in (2.6).

Theorem 3.4. *Let y and t denote positive numbers and let t be such that $\mathbb{P}(X > t) > 0$. Define T as a non-negative random value which is then distributed according to a*

geometric distribution with a modified initial term such that the initial term is zero if and only if

$$\mathbb{E}(T > y) = [y] + \frac{\mathbb{E}(T)}{1 - \mathbb{P}(T = 0)} \quad \text{for all } y,$$

where $[y]$ denotes the integer part of y .

Therefore, the non-negative random variable X is *geometrically* distributed if and only if

$$\mathbb{E}(X|X > t) = [t] + 1 + \mathbb{E}(X),$$

for all positive integers t .

3.3.3 A characterisation based on independence

Two discrete random variables X and Y are independent if and only if their joint distribution function can be written as

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \mathbb{P}(Y = y).$$

Ferguson (1965)

Ferguson (1965) based their characterisation on the *geometric* distribution as defined in (3.2).

Theorem 3.5. *Let X and Y be independent, nondegenerate discrete random variables. Then, $U = \min(X, Y)$ and $V = X - Y$ are independently distributed random variables if, and only if, both X and Y are geometric distributions.*

See the proof in Ferguson (1965).

3.3.4 A characterisation based on moments of truncated life

Su and Hung (2018) defined the moments,

$$(X - t)_+ = \max\{X - t, 0\},$$

with the n^{th} moment $\mathbb{E}[(X - t)_+^n]$ described as the n^{th} degree stop-loss transform of X .

Su and Hung (2018)

Su and Hung (2018) based a characterisation on the *geometric* distribution defined in (3.1).

Theorem 3.6. *Suppose that X is an independent, positive discrete random variable where c_1 and c_2 are constants and $\mathcal{N}$ denotes the non-negative integers. Assume that*

$$\mathbb{E}[(X - t)_+] = c_1 c_2^t, \quad \forall t \in \mathcal{N},$$

then $c_1 = \mu_1 > 1$, $c_2 = 1 - 1/\mu_1 > 0$, and X is geometrically distributed with parameter $\frac{1}{c_1}$.

Su and Hung (2018) gave a second characterisation in the theorem below.

Theorem 3.7. *Suppose that*

$$\mathbb{E}[(X - t)_+^2] = c_1 c_2^t, \quad \forall t \in \mathcal{N},$$

and suppose that

$$\mu_2 = 2\mu_1^2 - \mu_1,$$

then $c_1 = \mu_2 > 1$, $c_2 = 1 + (1 - 2\mu_1)/\mu_2 > 0$ and X is geometrically distributed with

parameter $\frac{(1+c_1-c_1c_2)}{2}$.

See the proof in Su and Hung (2018). The authors further also show that the *geometric* distribution can subsequently be characterised by Theorem 3.8 below.

Theorem 3.8. *Let X be a positive integer-valued random variable, and denote $\rho(\theta)$ as the Laplace-Stieltjes transform of its probability mass function (PMF):*

$$\begin{aligned}\rho(\theta) &= \mathbb{E} \left[\exp^{-\theta x} \right] \\ &= \sum_{x=1}^{\infty} \exp^{-\theta x} \mathbb{P}(X = x), \quad \theta \geq 0.\end{aligned}$$

Additionally, let ρ' and ρ'' represent the first and second derivatives of ρ , respectively, and let $\mu_n = \mathbb{E}(X^n)$ for $n \in \mathbb{N}^+$.

If X follows a geometric distribution with parameter p , then the following expressions hold:

$$\begin{aligned}\mathbb{E}[(X-t)_+] &= \frac{1}{p} (1-p)^t, \quad \forall t \in \mathbb{N}, \\ \mathbb{E}[(X-t)_+^2] &= \frac{2-p}{p^2} (1-p)^t, \quad \forall t \in \mathbb{N}.\end{aligned}$$

3.3.5 A characterisation based on order statistics

Define the order statistics of a set of independent and identically distributed (i.i.d) random variables, $X_1, \dots, X_n$, as $X_{(1)} \leq \dots \leq X_{(n)}$. We can define the distribution of the k^{th} order statistic in terms of the first-order statistic by

$$F_{(k)}(x) = \sum_{j=k}^n \binom{n}{j} F(x)^j [1 - F(x)]^{n-j},$$

The above expression is a general distribution of the k^{th} order statistic not only for the *geometric* distribution.

López-Blázquez and Miño (1998)

Consider the following theorem for the *geometric* distribution.

Theorem 3.9. *If $X_1, \dots, X_n$ are i.i.d positive discrete random variables, with $\mathbb{P}(X_1 = j) = p_j > 0$, for $j \geq 1$, then*

$$\mathbb{E}[X_{(2)} | X_{(1)} = j] = j + a,$$

for all $j \geq 1$ and $a > 0$, where q is the unique solution of the equation

$$\frac{n}{a} = \frac{1 - q^n}{1 - q} \frac{1 - q^{n-1}}{q^{n-1}},$$

if and only if X has the PMF given in (3.1).

See the proof in López-Blázquez and Miño (1998).

Arnold (1980)

Arnold (1980) developed a characterisation of the *geometric* distribution using the definition provided in (3.1).

Theorem 3.10. *If and only if X_1 is geometrically distributed then the conditional distribution of $X_{(k+1)} - X_{(k)}$ given $X_{(k+1)} - X_{(k)} > 0$ is the same as the unconditional distribution of $X_{(1)}$ for some pair (k, n) , where $1 \leq k < n$.*

Arnold (1980) gave a second characterisation as well. This characterisation is also based on the *geometric* distribution as defined in (3.1):

Theorem 3.11. *If and only if X_1 is geometrically distributed $\mathbb{P}(X_i = 1)$, $\mathbb{P}(X_i = 2)$ and $\mathbb{P}(X_i > 2)$ are strictly positive and the random variable $X_{(2)} - X_{(1)}$ and the event $[X_{(1)} = 1]$ are independent.*

See the proofs for both theorems in Arnold (1980).

Govindarajulu (1980)

In their characterisation of the *geometric* distribution, Govindarajulu (1980) employed the definition given in (3.1).

Theorem 3.12. *Assume that $D(i) = \mathbb{P}(X \geq i)$, and $d(i) = D(i+1)/D(i)$ for $i = 1, 2, \dots$. Then $X_{(1)}$ and $\{X_{(k)} \geq X_{(1)} \geq m\}$ are independent for some fixed constant k where $2 \leq k \leq n$ and some $m \geq 1$ if and only if $D(i) = (1-p)^{i-1}$, for $i = 1, \dots$ given that $D(l) = (1-p)^{l-1}$, for $l = 1, 2, \dots, m+1$.*

See proof in Govindarajulu (1980).

The *geometric* distribution is also characterised by the independence of the distribution of $X_{(1)}$ and the event $\{X_{(k)} - X_{(1)} = 0\}$ for some arbitrarily fixed k where $(2 \leq k \leq n)$. This is shown by setting $m = 1$ and considering the complement of the event $\{X_{(k)} - X_{(1)} \geq 1\}$.

As noted by Ferguson (1965) (see Theorem (3.5)), $\min(X_1, X_2)$ and $|X_1 - X_2|$ are independent if and only if the X_i are *geometrically* distributed and X_1 and X_2 are independent non-negative discrete random variables. Govindarajulu (1980) extended this result to show that the smallest of two independent non-negative discrete random variables as well as the sample range are independent if and only if the random variables are *geometrically* distributed.

Furthermore, if $D_j(l) = (1-p)_j^{l-1}$ for $l = 1, 2, \dots$, and $j = 1, 2, \dots, n$ then $X_{(1)}$ and the event $\{X_{(k)} - X_{(1)} = m\}$ are independent for some arbitrarily fixed k where $2 \leq k \leq n$ and $m \geq 1$.

3.3.6 A characterisation based on weak records

Suppose that $X_1 \in \{0, 1, \dots, N \leq \infty\}$. Then, a sequence of upper weak record times, $U_{\mathcal{W}}(n)$, and upper weak record values, $X_{U_{\mathcal{W}}}(n)$, can be defined as $U_{\mathcal{W}}(1) = 1$ and $U_{\mathcal{W}}(n+1) = \min\{j > U_{\mathcal{W}}(n) : X_j \geq X_{U_{\mathcal{W}}}(n)\}$, for $n = 1, 2, \dots$

Ahsanullah and Aliev (2011)

Ahsanullah and Aliev (2011) based their characterisation on the *geometric* distribution as defined in (3.1).

Theorem 3.13. *Assume that $\{X_n, n \geq 1\}$ is a series of non-degenerate i.i.d random variables which is concentrated on $0, 1, \dots, N \leq \infty$ with distribution function F and $p(1) \geq 0$, where $p(1)$ is the geometric distribution as in (3.1). Then, if and only if*

$$X_{U_w(m)} \stackrel{d}{=} X_1 + \dots + X_m,$$

X_1 has a geometric distribution, as defined in (3.1), with parameter $p \geq 0$ for all $m \geq 2$.

See proof in Ahsanullah and Aliev (2011).

3.3.7 A characterisation based on conditional variance

Let $\mathcal{B}$ be the set of all real-valued functions of X . For any $q(x) \in \mathcal{B}$ such that $\mathbb{E}[q^2(X)] < \infty$, we define the following conditional expectations:

1. The conditional mean of $q(x)$ given $X > x$ is denoted by $m_q(x)$ and is defined as $m_q(x) = \mathbb{E}[q(X) \mid X > x]$.
2. The conditional mean of $h(x)$ given $X \leq x$ is denoted as $r_q(x)$ and is defined as $r_q(x) = \mathbb{E}[q(X) \mid X \leq x]$.

We also define the mean of $q(X)$ as μ and the variance of $q(X)$ as σ_q^2 .

Sudheesh and Nair (2010)

Sudheesh and Nair (2010) based their characterisation on the *geometric* distribution as defined in (3.2). The following helping theorem is given before the characterisation.

Theorem 3.14. Denote $c(x)$ as functions in $\mathcal{B}$ and suppose that $q(x)$ and $c(x)$ satisfies

- (i) $\Delta c(x) \neq 0 \quad \forall \quad x, \quad c(x)q(x) > 0.$
- (ii) $\mathbb{E}(|\Delta c(X)|g_q(X)), \quad \mathbb{E}(c(X)q(X)), \quad \mathbb{E}(q^2(X)) \quad \text{are all} \quad < \infty.$
- (iii) $f(0) > 0 \quad \text{and the support of } X \text{ is an integer interval.}$

Then, X has PMF

$$\frac{f(x+1)}{f(x)} = \frac{\sigma_h g_q(x)}{\sigma_h g_q(x+1) - \mu_q + q(x+1)},$$

where $r_q(x) = \mu_q - \sigma_q g_q(x)\lambda(x)$ with $\sigma_q g_q(0) = \mu_q - q(0)$.

If and only if for all $x = 0, 1, 2, \dots$ the following holds,

$$\begin{aligned} \text{Cov}(c(X), q(X) \mid X > x) &= \sigma_q \mathbb{E}(\Delta c(X) g_q(X) \mid X > x) \\ &+ (\mu_q - m_q(x))(m_c(x) - c(x+1)), \end{aligned}$$

where $m_c(x) = \mathbb{E}(c(X) \mid X > x)$ for some $g_q(x) \in \mathcal{B}$.

See proof in Sudheesh and Nair (2010). The theorem for the *geometric* distribution becomes:

Theorem 3.15. Define the mean and variance as $\mathbb{E}(q(X))$ and $\mathbb{V}()$,

$$m(x) = \mu + x + 1 \quad \text{and}$$

$$\mathbb{V}(X \mid X > x) = \mathbb{V}(X),$$

then in terms of the MRL function, we have that

$$\mathbb{V}(X \mid X > x) = (m(x) - x)(m(x) - x - 1),$$

if and only if X has the geometric distribution.

See the derivation in Sudheesh and Nair (2010).

3.3.8 A characterisation based on the geometric renewal process

Let $S_0 = 0$, $S_m = \sum_{i=1}^m X_i$, $m \geq 1$ and $A(t) = \sup \{m \mid S_m \leq t\}$, $t \geq 0$. Then denote by $\{A(t), t \geq 0\}$ a renewal process and by $\gamma_t = S_{A(t)+1} - t$ the residual life at time t .

Su and Hung (2018)

Su and Hung (2018) based their characterisation on the *geometric* distribution as expressed in (3.1).

Theorem 3.16. *If*

$$\mathbb{E}(\gamma_t^2) = c\mathbb{E}^2(\gamma_t), \quad \forall t \in \mathcal{N},$$

with c a constant and γ_t the moments of residual life. If

$$\mathbb{E}(\gamma_1) = \mu_1.$$

then $1 < c = 2 - 1/\mathbb{E}(\gamma_1) < 2$ and X_1 is geometrically distributed with parameter $2 - c$.

See the proof in Su and Hung (2018). The authors developed a second characterisation as well; see Theorem 3.17 below.

Theorem 3.17. *Suppose that*

$$\mathbb{E}(\gamma_t^2) = 2\mathbb{E}^2(\gamma_t) - \mathbb{E}(\gamma_t), \quad \forall t \in \mathcal{N},$$

then X_1 is geometrically distributed with parameter $\frac{1}{\mathbb{E}(\gamma_1)}$.

See the proof in Su and Hung (2018).

3.4 Tests for the geometric distribution

In this section, we give short descriptions of some of the existing goodness-of-fit tests for the *geometric* distribution.

3.4.1 A test statistic based on linear regression of order statistics

Jiménez-Gamero and Alba-Fernández (2021) constructed a test statistic based on a linear regression of order statistics where the starting point is the characterisation of López-Blázquez and Miño (1998).

The authors defined the null hypothesis: $H_0 : X \sim \text{Geo}(p)$, $\forall p \in (0, 1)$ as in (3.2) with $\text{Geo}(p)$ the *geometric* distribution evaluated with parameter p .

Recall the characterisation as defined in Theorem 3.9, Jiménez-Gamero and Alba-Fernández (2021) gave an equivalent characterisation. Suppose that $M = \min(X_1, X_2)$ and $D = |X_1 - X_2|$ then,

$$\mathbb{E}(D|M = j) = \gamma, \quad (3.3)$$

if and only if $X_1 \sim \text{Geo}(p)$ with $n \geq 2$ where $\gamma = 2 \frac{q}{1-q^2}$.

The test statistic is constructed using the property described in (3.3), as established by Bierens (1982). If Y is a random variable and W is a random vector of dimension r , then $\mathbb{E}(Y|W) = 0$ if and only if $\mathbb{E}(Y e^{it^T W}) = 0$ for all $t \in \mathbb{R}^r$.

This relationship can be written as

$$JG = \int_0^\infty |\tilde{S}_n(t)|^2 w(t) dt, \quad (3.4)$$

where

$$\tilde{S}_n(t) = \frac{1}{n(n-1)} \sum_{1 \leq j \neq k \leq n} (D_{jk} - \hat{\gamma}) e^{itM_{jk}}, t \in R,$$

with $D_{jk} = |X_j - X_k|$, $M_{jk} = \min(X_j, X_k)$ for $1 \leq j \neq k \leq n$, and $\hat{\gamma} = \frac{2\hat{q}}{1-\hat{q}^2}$ which reduces to the maximum likelihood estimator of the *geometric* distribution. In addition, $w(t)$ is a weight function.

Another form of (3.4) is obtained through $G_{n,\gamma} = \|\tilde{S}_n\|^2$,

$$G_{n,\gamma} = \frac{1}{n^2(n-1)^2} \sum_{j \neq k, u \neq v} (D_{jk} - \hat{\gamma})(D_{uv} - \hat{\gamma}) I_w(M_{jk} - M_{uv}),$$

with

$$I_w(t) = \int \cos(ts) w(s) ds,$$

and chosen weight function $w(s)$.

It is noted in Jiménez-Gamero and Alba-Fernández (2021) to use a numerical integration technique to calculate the test statistic since the exact calculation of $G_{n,\gamma}$ involves a fourfold sum and is very time-consuming. The authors also provided a derivation for the null distribution and consistency of the test in the setting of a Hilbert space. In the simulation study, they used both the weighted and parametric bootstrap in the estimation of the powers.

An observed data example relating to pregnant women with planned pregnancies was also included. Only women who got pregnant within 12 cycles of trying were considered in the study and they were also classified as smokers and non-smokers. It was found that the author's test rejects the null hypothesis for the group of non-smokers and the combined population and the test does not reject the null hypothesis

for the group of smokers. Thus for the group of smokers the authors did not reject the null hypothesis that the number of cycles to get pregnant has a *geometric* distribution.

3.4.2 A test based on the probability generating function

Rueda and O'Reilly (1999) proposed a statistic for discrete distributions. This test is a quadratic statistic that uses the probability-generating function (PGF) of a particular discrete distribution and its empirical counterpart.

The empirical PGF of an observed dataset $X_1, X_2, \dots, X_n$ is defined to be

$$\mathcal{G}_n(t) = \frac{1}{n} \sum_{j=1}^n t^{X_j} \quad |t| \leq 1.$$

Now, consider the empirical process defined by Marques and Pérez-Abreu (1989)

$$\xi_n(t; \theta) = \sqrt{n} \{ \phi_n(t) - \phi(t; \theta) \},$$

and in the case where θ is unknown

$$\xi_n(t; \hat{\theta}_n) = \sqrt{n} \{ \phi_n(t) - \phi(t; \hat{\theta}_n) \},$$

where $\hat{\theta}_n$ is a consistent estimator for θ .

Rueda and O'Reilly (1999) proposed the test statistics

$$R_n(\theta) = \int_0^1 \xi_n^2(t; \theta) dt,$$

and

$$\hat{R}_n(\hat{\theta}_n) = \int_0^1 \xi_n^2(t; \hat{\theta}_n) dt.$$

Rueda and O'Reilly (1999) included the test statistic for the Poisson distribution,

$$R_n^P(\theta) = n \int_0^1 \left(\phi_n(t) - e^{\lambda(t-1)} \right)^2 dt,$$

where $\phi_\lambda(t) = e^{\lambda(t-1)}$ is the PGF of the Poisson distribution.

The test statistic for the negative binomial distribution is also given

$$R_n^{NB}(\theta) = n \int_0^1 \left[\phi_n(t) - \left(\frac{p}{1-qt} \right)^r \right]^2 dt,$$

where $\phi(t; p) = \frac{p}{1-qt}$ is the PGF of the negative binomial distribution with $p \in [0, 1]$, $t \in [-1, 1]$ and $q = 1 - p$.

The *geometric* distribution as in (3.2) is a special case of the negative binomial distribution and consequently a test for the *geometric* distribution can be obtained by setting

$$R_n = \int_0^1 \left| \phi_n(t) - \frac{t\hat{p}_n}{1 - \hat{q}_n t} \right|^2 dt,$$

where $\hat{q}_n = 1 - \hat{p}_n$ and $\hat{p}_n$ is a consistent estimator for p .

Rueda and O'Reilly (1999) included a numerical study to show the accuracy of the asymptotic approximation of the distribution of R_n for finite samples. They found that for known λ , even for small sample sizes, the approximation of the limiting distribution is very good and for unknown λ only large sample sizes obtain good approximations. This is the case for both the Poisson and the negative binomial distributions.

3.4.3 A test based on the empirical distribution function

Let $F(t) = \mathbb{P}(X \leq t)$ be the distribution function of a random variable X . Now, by estimating the unknown parameter θ with a consistent estimator, $\hat{\theta}_n$, the Kolmogorov-

Smirnov test statistic for the *geometric* distribution as in (3.2) can be written as

$$KS_n = \sqrt{n} \max_{0 \leq k \leq M} \left| F_n(k) - F(k, \hat{\theta}_n) \right|,$$

where $M := \max_{1 \leq j \leq n} X_j$.

The Cramér-von Mises test statistic, on the other hand, can be written as

$$C_n = n \sum_{k=0}^l \left\{ F_n(k) - F(k, \hat{\theta}_n) \right\}^2 f(k, \hat{\theta}_n) + R,$$

with $l \geq M$, $0 \leq R \leq n \{1 - F(l, \hat{\theta}_n)\}^3$ and $l := \min \{m \geq M : n \{1 - F(m, \hat{\theta}_n)\}^3 \leq 10^{-4}\}$.

Henze (1996) then gave an alternative Cramér-Von-Mises statistic in the form,

$$H_n = n \sum_{k=0}^{\infty} \left\{ F_n(k) - F(k, \hat{\theta}_n) \right\}^2 \left\{ F_n(k) - F_n(k-1) \right\}.$$

The test rejects for large values of the Kolmogorov-Smirnov statistic and the Cramér-von Mises statistic.

3.4.4 A test based on characterisation identities of Stein's method

Betsch, Ebner and Nestmann (2022) developed tests for discrete models and based their test statistic on the characterisation by Ley and Swan (2013). The starting point for these tests is the expectation

$$\mathbb{E} \left[\Delta^+ f(X) + \frac{\Delta^+ p(X)}{p(X)} f(X+1) \right] = 0. \quad (3.5)$$

For a large group of test functions f , where Δ^+ is denoted as the forward difference operator.

The authors showed that by restating the characterisation, X is regulated by the PMF, if and only if, p , satisfies

$$p(k) = \mathbb{E} \left[-\frac{\Delta^+ p(X)}{p(X)} \mathbb{1}\{X \geq k\} \right], \quad k \in \mathbb{N}_0.$$

The tests are then constructed as follows. Suppose that the above characterisation holds for some p_θ , then

$$\hat{\rho}_n(k) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}\{X_j = k\},$$

and

$$\frac{1}{n} \sum_{j=1}^n -\frac{\Delta^+ p_{\vartheta_n}(X_j)}{p_{\vartheta_n}(X_j)} \mathbb{1}\{X_j \geq k\},$$

should be close for each $k \in \mathbb{N}_0$.

Therefore, Betsch et al. (2022) proposed the following test statistic for a specified distribution, by using the characterisation

$$\sum_{k=0}^{\infty} \left(\frac{1}{n} \sum_{j=1}^n \mathbb{1}\{X_j = k\} + \frac{1}{n} \sum_{j=1}^n \frac{\Delta^+ p_{\vartheta_n}(X_j)}{p_{\vartheta_n}(X_j)} \mathbb{1}\{X_j \geq k\} \right)^2,$$

with $\hat{\vartheta}_n$ a constant estimator for ϑ . The authors noted that the test rejects the null hypothesis for large values of the test statistic.

Betsch et al. (2022) suggested that for the unknown $\vartheta_0 \in \Theta$, where Θ denotes the parameter space, the following estimator for θ_0 should be used

$$\hat{\vartheta}_n = \operatorname{argmin}_{\vartheta \in \Theta} \sum_{k=0}^{\infty} \left(\frac{1}{n} \sum_{j=1}^n \mathbb{1}\{X_j = k\} + \frac{1}{n} \sum_{j=1}^n \frac{\Delta^+ p_{\vartheta}(X_j)}{p_{\vartheta}(X_j)} \mathbb{1}\{X_j \geq k\} \right)^2.$$

The authors outlined a test statistic for two classic discrete distributions. The first being, the Poisson distribution and the second being the negative binomial distribution. A test for the *geometric* distribution can then be constructed by using the fact that the *geometric* distribution is a special case of the negative binomial distribution. The authors also found that in both cases their new method leads to excellent approaches for these distributions.

3.4.5 A test based on the integrated distribution function

Klar (1999) proposed an omnibus test for discrete distributions based on the difference between the integrated distribution function (IDF) and its empirical counterpart. Define the IDF as,

$$\Psi(t) = \int_t^{\infty} (1 - F(x)) dx.$$

If X is a positive discrete random variable with PMF, f , the IDF can be written as

$$\begin{aligned} \Psi(t) &= \sum_{k=[t]+1}^{\infty} (k - t) f(k) \\ &= ([t] + 1 - t) \bar{F}([t]) \sum_{k=[t]+1}^{\infty} S(k), \end{aligned}$$

where S as defined in (2.1) and $[t]$ denotes the integer part of t . The empirical integrated distribution function (EIDF) is defined as

$$\begin{aligned} \Psi_n(t) &= \int_t^{\infty} \bar{F}_n(x) dx \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - t) \mathbb{1}\{X_i > t\}, \end{aligned}$$

and the estimated IDF is defined as

$$\widehat{\Psi}_n(t) = \bar{F}\left(x, \widehat{\vartheta}_n\right) dx,$$

with $\widehat{\vartheta}_n = \widehat{\vartheta}_n(X_1, \dots, X_n)$ an estimator for the parameter vector ϑ .

A Kolmogorov-Smirnov-type test statistic based on the quantities defined above is then

$$K_n = \sup_{t \geq 0} \sqrt{n} \left| \Psi_n(t) - \widehat{\Psi}(t) \right|.$$

Note that K_n can also be written as a functional of the estimated discrete empirical process, $\mathcal{L}_n = (Z_{n,k})_{k \geq 0}$ where $Z_{n,k} = \sqrt{n} \left(F_n(k) - F(k, \widehat{\vartheta}_n) \right)$, i.e.,

$$\Psi_n(k) - \widehat{\Psi}(k) = \sum_{j=k}^{\infty} \left(F(k, \widehat{\vartheta}_n) - F_n(k) \right)$$

A suitable closed form for the test statistic was suggested by Klar (1999) in order to calculate the test in practice,

$$K_n = \sup_{1 \leq k \leq M} \sqrt{n} \left| \bar{X}_n - \mathbb{E}_{\widehat{\vartheta}_n}(X) + \sum_{j=0}^{k-1} \left(F(j, \widehat{\vartheta}_n) - F_n(j) \right) \right|,$$

with $M = \max_{1 \leq j \leq n} X_j$.

Now, it is important to note that in the case of the *geometric* distribution, we have $\mathbb{E}_{\widehat{\vartheta}_n}(X) = \bar{X}_n$ and the test statistic for the *geometric* distribution, as in (3.2), becomes,

$$K_n = \sup_{1 \leq k \leq M} \sqrt{n} \left| \sum_{j=0}^{k-1} \left(F(j, \widehat{\vartheta}_n) - F_n(j) \right) \right|.$$

Klar (1999) included a bootstrap version of the test for common lattice models and found that even for small samples the statistic has accurate error rates and high powers compared to other procedures over a large range of alternatives and rejects the null hypothesis for large values of K_n .

3.4.6 A test based on a differential equation

Milošević, Jiménez-Gamero and Alba-Fernández (2021) proposed a test statistic for the *geometric* distribution based on characterisation in terms of a differential equation concerning its PGF.

A random variable X is *geometrically* distributed as in (3.1) if, and only if its probability PGF, $\mathcal{G}(t) = \mathbb{E}(t^X)$, satisfies the differential equation

$$pt^2\mathcal{G}'(t) - \mathcal{G}^2(t) = 0,$$

where $t \in [0, 1]$, $p \in (0, 1)$ and $\mathcal{G}'(t) = \frac{d}{dt}\mathcal{G}(t)$.

In order to proceed we replace $\mathcal{G}, \mathcal{G}'$ and p by suitable estimators. The empirical PGF, $\mathcal{G}_n(t) = \frac{1}{n} \sum_{j=1}^n t^{X_j}$, and its derivative, $\mathcal{G}'_n(t) = \frac{d}{dt}\mathcal{G}_n(t)$, can be consistently estimated by the empirical PGF obtaining,

$$D_n(t, p) = pt^2\mathcal{G}'_n(t) - \mathcal{G}_n^2(t),$$

where p is estimated by a consistent estimator, $\hat{p} = 1/\bar{X}$.

Next, Milošević et al. (2021) expressed $D_n(t, p)$ as a power series,

$$\begin{aligned} D_n(t, \hat{p}) &= pt^2 \frac{1}{n} \sum_{j=1}^n X_j t^{X_j-1} - \frac{1}{n^2} \sum_{i,j} t^{X_i+X_j} \\ &= \frac{1}{n} \hat{p} \sum_{j \geq 1} t^{j+1} j n_j - \frac{1}{n^2} \sum_{j \geq 2} t^j \sum_{k=1}^{j-1} n_k n_{j-k} \end{aligned}$$

$$= \sum_{j \geq 2} \beta_j(\hat{p}) t^j,$$

where $\hat{p}_i = \frac{n_i}{n}$ denotes the relative frequency, n_i denotes the observed frequency of i in the underlying sample and

$$\beta_j(p) = p(j-1)\hat{p}_{j-1} - \sum_{k=1}^{j-1} \hat{p}_k \hat{p}_{j-k}, \quad j \geq 2.$$

D_n should be small for each t if the null hypothesis is true. This closeness infers that all the coefficients, $\beta_j(p)$ should be near zero. Therefore, Milošević et al. (2021) suggests as a first test statistic

$$M^{(1)} = \sum_{j \geq 2} \beta_j^2(\hat{p}).$$

The second test statistic is based on the L^2 - norm and can be expressed as

$$\begin{aligned} M^{(2)} &= \int_0^1 D_n^2(t, \hat{p}) w(t) dt \\ &= \|D_n\|_w^2, \end{aligned}$$

where $w(\cdot)$ is a non-negative weight function. The third proposed test statistic is based on the supremum norm and takes the form

$$\begin{aligned} M^{(3)} &= \sup_{t \in [0,1]} |D_n(t, \hat{p})| \\ &= \|D_n\|_\infty. \end{aligned}$$

Milošević et al. (2021) proposed a second set of statistics based on the same characterisation, however they considered the difference

$$\hat{D}_n(t, p) = pt^2 \mathcal{G}'_n(t) - g_n(t) \mathcal{G}(t, p),$$

where $\mathcal{G}(t, \hat{p})$ is the PGF of the *geometric* distribution.

The authors suggested however to use the following difference equation

$$\begin{aligned}\tilde{D}_n(t, \hat{p}) &= \frac{(1 - \hat{q}t)}{\hat{p}t} \hat{D}_n(t, \hat{p}) \\ &= (1 - \hat{q}t) \mathfrak{t}\mathcal{G}'_n(t) - \mathcal{G}_n(t).\end{aligned}$$

Following similar reasoning as for $M^{(2)}$ and $M^{(3)}$ and based on $\tilde{D}_n(t, \hat{p})$ they defined

$$\widetilde{M}^{(2)} = \left\| \tilde{D}_n \right\|_w^2, \quad \widetilde{M}^{(3)} = \left\| \tilde{D}_n \right\|_\infty.$$

To construct $\widetilde{M}^{(1)}$ they first used the power series representation of $\tilde{D}_n(t, p)$ and since

$$\tilde{D}_n(t, p) = \sum_{j \geq 2} \tilde{\beta}_j(p) t^j,$$

where $\tilde{\beta}_j(p) = (j-1) \{ \hat{p}_j - (1-p) \hat{p}_{j-1} \}$ the test statistic is then

$$\begin{aligned}\widetilde{M}^{(1)} &= \sum_{j \geq 2} \tilde{\beta}_j^2(\hat{p}) \\ &= \left\| \tilde{\beta}(\hat{p}) \right\|_{\ell^2}^2,\end{aligned}$$

with $\tilde{\beta}(p) = (\tilde{\beta}_2(p), \tilde{\beta}_3(p), \dots)$.

Closed form expressions for $M^{(2)}$ and $\widetilde{M}^{(2)} = B_{n,\gamma}$ follow by choosing the weight function to be $w(t) = t^\gamma, \gamma > 0$,

$$\begin{aligned}M^{(2)} &= \hat{p}^2 \frac{1}{n^2} \sum_{k,j} \frac{X_k X_j}{X_k + X_j + 3 + \gamma} + \frac{1}{n^4} \sum_{k,l,s,r} \frac{1}{X_k + X_l + X_r + X_s + 1 + \gamma} \\ &\quad - 2\hat{p} \frac{1}{n^3} \sum_{k,l,s} \frac{X_k}{X_k + X_l + X_s + 2 + \gamma},\end{aligned}$$

$$\begin{aligned}
B_{n,\gamma} = & \frac{1}{n^2} \sum_{i,j} \left(\frac{(X_j - 1)(X_i - 1)}{X_i + X_j + \gamma + 1} - (1 - \hat{p}) \frac{2X_i X_j - X_i - X_j}{X_i + X_j + \gamma + 2} \right. \\
& \left. + (1 - \hat{p})^2 \frac{X_i X_j}{X_i + X_j + \gamma + 3} \right).
\end{aligned}$$

Milošević et al. (2021) noted that the exact calculation of $M^{(2)}$ is a fourfold sum, which is time-consuming for large sample sizes and suggested a numerical integration method for faster implementation. It is also noted that these tests reject the null hypothesis for large values of the test statistics. The authors also included a study on observed data and concluded that these tests are very comparable with existing goodness of fit tests.

3.4.7 A test based on the empirical characteristic functions

Jiménez-Gamero and Batsidis (2017) based their test on the empirical characteristic function and followed an approach similar to Baringhaus and Henze (1992). The starting point is writing the test by Baringhaus and Henze (1992) for the d -variable case as

$$n \int \left| c_n(t) - \psi(t; \hat{\mu}, \hat{\Sigma}) \right|^2 g(t; \hat{\mu}, \hat{\Sigma}) dt,$$

with $c_n(t)$ the empirical characterisation function,

$$c_n(t) = \frac{1}{n} \sum_{j=1}^n \exp(it' X_j),$$

where the prime denotes the transpose, $\psi(t; \hat{\mu}, \hat{\Sigma})$ denotes the characteristic function, μ is defined as the mean and Σ is denoted as the variance matrix. Jiménez-Gamero and Batsidis (2017) suggested a test statistic of the form

$$J_n(\hat{\theta}) = n \int \left| c_n(t) - c_0(t; \hat{\theta}) \right|^2 dG(t),$$

with $c_0(t; \hat{\theta})$ the characteristic function of and $\hat{\theta}$ the consistent estimator for θ and G the distribution function. By assuming that the weight function of $J_n(\hat{\theta})$ is completely specified. Jimenez-Gamero (2009) expressed the test statistic in the form,

$$J_n(\hat{\theta}) = \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^n h(X_j, X_k; \hat{\theta}),$$

with

$$\begin{aligned} h(x, y, \theta) &= u(x - y) - u_0(x; \theta) - u_0(y; \theta) + u_{00}(\theta), \\ u_0(x; \theta) &= \int u(x - y) dF(y; \theta), \\ u_{00}(\theta) &= \int u(x - y) dF(x; \theta) dF(y; \theta), \end{aligned}$$

where $u(t) = \int \cos(x't) dG(x)$, which is the real part of the characteristic function of G .

The authors showed that a bootstrap simulation can be used to approximate the null distribution of the test statistic consistently. The asymptotic results and an in-depth discussion on the weight function were also provided. In their work Jiménez-Gamero and Batsidis (2017), included an observed dataset and concluded that the p -value derived from the test statistic does not provide grounds for rejecting the null hypothesis, which posits that the data follows a geometric distribution.

Chapter 4

New tests of fit for the geometric distribution

4.1 Introduction

In this chapter, we define and explain the motivation and theory behind our new tests of fit for the *geometric* distribution. We adopt a methodology analogous to that of Baringhaus and Henze (1992), while utilising the characterisation used in Rao Jammalamadaka and Taufer (2006), in order to construct our test statistics; tailored specifically to the *geometric* distribution as defined in (3.1). This choice was motivated by the scarcity of research concerning the MRL for the discrete case. We include all necessary proofs and definitions.

4.2 New tests of fit

Call to mind the characterisation of Shanbhag (1970) which states that if X has a *geometric* distribution with distribution function as defined in (3.1) then

$$\mathbb{E}[X - t | X > t] = \mathbb{E}[X], \quad (4.1)$$

if and only if the MRL is constant. By following the approach of Baringhaus and Henze (1992) we start with the construction of our test statistics. We give the following theorem as starting point.

Theorem 4.1. *The random variable X is geometrically distributed with PMF $f(x) = p(1-p)^{x-1}$ with $x > 0$ and $p \in [0, 1]$ if and only if the MRL is constant. That is,*

$$\mathbb{E}[\min(X, t)] = \mathbb{E}[X] F(t). \quad (4.2)$$

Proof.

$$\begin{aligned} \mathbb{E}[\min(X, t)] &= \sum_{x=0}^{\infty} \min(x, t) f(x) \\ &= \sum_{x=0}^t x f(x) + t \sum_{x=t+1}^{\infty} f(x) \\ &= \sum_{x=0}^{\infty} x f(x) - \sum_{x=t+1}^{\infty} x f(x) + t \sum_{x=t+1}^{\infty} f(x) \\ &= \mathbb{E}[X] - (1 - F(t)) \frac{\sum_{x=t+1}^{\infty} (x - t) f(x)}{1 - F(t)} \\ &= \mathbb{E}[X] - (1 - F(t)) \mathbb{E}[X - t | X > t] \\ &= \mathbb{E}[X] - (1 - F(t)) \mathbb{E}[X] \\ &= \mathbb{E}[X] F(t), \end{aligned}$$

where we used the property of the *geometric* distribution that $\mathbb{E}[X - t | X > t] = \mathbb{E}[X]$.

□

Theorem 4.1 has an interesting connection to the characterisation by Su and Hung (2018) which states that,

$$\mathbb{E}[(X - t)_+] = \frac{1}{p}(1 - p)^t,$$

with $(X - t)_+ = \max(X - t, 0)$ if X is *geometrically* distributed. We show that,

$$\min(X, t) - X = -\max(X - t, 0). \quad (4.3)$$

Start by writing the left and right side of (4.3) as,

$$\min(X, t) = \begin{cases} 0 & \text{if } t \leq X, \\ t - X & \text{if } t > X, \end{cases}$$

and

$$-\max(X - t, 0) = \begin{cases} -(X - t) & \text{if } t < X, \\ 0 & \text{if } t \geq X, \end{cases}$$

therefore $\min(X, t) - X$ and $-\max(X - t, 0)$ are equal. This leads to a connection between $\mathbb{E}[\min(X, t)]$ and $\mathbb{E}[\max(X - t)_+]$.

We opted to follow the approach by Baringhaus and Henze (1992) for our test statistic rather than the estimation techniques proposed in Chapter 2 since the MRL estimators are not very well studied in the discrete case. However, further tests can be based on MRL estimates for the discrete case.

Let ζ_n be the empirical version of the difference between the left-hand side and right-hand of side (4.3) given as

$$\begin{aligned} \zeta_n(t) &= \frac{1}{n} \sum_{j=1}^n \min(X_j, t) - (1 - (1 - \hat{p})^t) \bar{X} \\ &= \frac{1}{n} \sum_{j=1}^n \min(X_j, t) - \bar{X} \left(1 - \left(\frac{\bar{X} - 1}{\bar{X}} \right)^t \right). \end{aligned}$$

Now, we want to calculate how this difference at the extreme order statistics, i.e.,

the minimum data point and the maximum of the data. Therefore, consider the formulation at the minimum of the data,

$$\begin{aligned}\zeta_n(1) &= \frac{1}{n} \sum_{j=1}^n \min(X_j, 1) - \bar{X} \left(1 - \left(\frac{\bar{X} - 1}{\bar{X}} \right)^1 \right) \\ &= \frac{1}{n} \sum_{j=1}^n 1 - \bar{X} \frac{1}{\bar{X}} \\ &= 0.\end{aligned}$$

Denoting the sample maximum, $m = \max[X_1, X_2, \dots, X_n]$, we consider $\zeta_n(m)$,

$$\begin{aligned}\zeta_n(m) &= \frac{1}{n} \sum_{j=1}^n \min(X_j, m) - \bar{X} \left(1 - \left(\frac{\bar{X} - 1}{\bar{X}} \right)^m \right) \\ &= \frac{1}{n} \sum_{j=1}^n X_j - \bar{X} \left(1 - \left(\frac{\bar{X} - 1}{\bar{X}} \right)^m \right) \\ &= \bar{X} \left(\frac{\bar{X} - 1}{\bar{X}} \right)^m, m \geq 0.\end{aligned}$$

We find that the difference in the point m has a value of $\bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^m$ larger than zero.

In the same way, it follows that, for every integer $t > m$, we have,

$$\zeta_n(t) = \bar{X} \left(\frac{\bar{X} - 1}{\bar{X}} \right)^t, t \geq 0.$$

We will be interested in the behaviour of ζ_n for large values of t , therefore we define

$$\Delta \zeta_n(t+1) = \zeta_n(t+1) - \zeta_n(t),$$

and we consider the behaviour of $\Delta \zeta_n(t+1)$ when $t \geq m$,

$$\begin{aligned}\Delta \zeta_n(t+1) &= \zeta_n(t+1) - \zeta_n(t) \\ &= \bar{X} \left(\frac{\bar{X} - 1}{\bar{X}} \right)^{t+1} - \bar{X} \left(\frac{\bar{X} - 1}{\bar{X}} \right)^t\end{aligned}$$

$$= \bar{X} \left(\frac{\bar{X} - 1}{\bar{X}} \right)^t (1 - \bar{X}) \leq 0,$$

with the equality holding if, and only if, $\bar{X} = 1$ (which is the situation where $X_1 = X_2 = \dots = X_n = 1$). This implies that ζ_n is positive and decreasing for $t > m$, and allows considering the domain only up to the maximum of the data. Following the calculations provided above we can now begin with the construction of our new test statistics.

We propose the first of three tests. The first is a Kolmogorov-Smirnov type test based on ζ_n with formulation,

$$\begin{aligned} T_n^{(1)} &= \max_{t \in 1, 2, \dots} |\zeta_n(t)| \\ &= \max_{t \in 1, 2, \dots, m} \left| \frac{1}{n} \sum_{j=1}^n \min(X_j, t) - \bar{X} \left(1 - \left(\frac{\bar{X} - 1}{\bar{X}} \right)^t \right) \right|. \end{aligned}$$

Now, for the second of our tests, we consider the Cramér-von Mises type test which is to be defined,

$$T_n^{(2)} = \frac{1}{n} \sum_{t=2}^m \zeta_n^{(2)} f_n(t), \quad \text{since } f_n(t) = 0 \quad \forall \quad t > m,$$

where $f_n(t) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}(X_j = t)$.

Note that the above is based on an empirical weight function. We then consider using a parametric weight function for the third test, where the parametric weight function can be defined as the estimated *geometric* mass function $f_{\hat{p}}(t) = \left(1 - \frac{1}{\bar{X}}\right)^t \frac{1}{\bar{X}}$.

Our third and final test is therefore defined to be

$$T_n^{(3)} = \frac{1}{n} \sum_{t=1}^{\infty} \zeta_n^{(2)} \left(1 - \frac{1}{\bar{X}}\right)^t \frac{1}{\bar{X}}.$$

The computation of $T_n^{(3)}$ requires the calculation of an infinite sum. So, in order to use this test, a calculable form is required.

Now consider $T_n^{(3)}$ written as,

$$\begin{aligned} T_n^{(3)} &= \frac{1}{n} \sum_{t=2}^{\infty} \zeta_n^{(2)} f_{\hat{p}}(t) \\ &= \frac{1}{n} \sum_{t=2}^m \zeta_n^{(2)} f_{\hat{p}}(t) + \frac{1}{n} \sum_{t=m+1}^{\infty} \zeta_n^{(2)} f_{\hat{p}}(t). \end{aligned}$$

The first summation above is finite and can easily be computed. The second summation becomes,

$$\begin{aligned} \sum_{t=m+1}^{\infty} \zeta_n^{(2)} f_{\hat{p}}(t) &= \sum_{t=m+1}^{\infty} \left[\bar{X} \left(1 - \frac{1}{\bar{X}}\right)^t \right]^2 \left(1 - \frac{1}{\bar{X}}\right)^{t-1} \frac{1}{\bar{X}} \\ &= \sum_{t=m+1}^{\infty} \bar{X} \left(\frac{\bar{X}-1}{\bar{X}}\right)^{3t-1}, \end{aligned}$$

which can be written as the difference between an infinite and a finite *geometric* series.

Consider the following,

$$S_m = \frac{a(1 - r^{m+1})}{(1 - r)}. \quad (4.4)$$

Assume, for a moment, that $0 < r < 1$, then, we have that $S_{\infty} = \frac{a}{1-r}$. Now we have that,

$$\sum_{j=m+1}^{\infty} ar^{j-1} = \sum_{j=1}^{\infty} ar^{j-1} - \sum_{j=1}^m ar^{j-1}$$

$$= \frac{ar^{m+1}}{1-r}.$$

Set $a = \bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^2$ and $r = \left(\frac{\bar{X}-1}{\bar{X}} \right)^3$ we have that,

$$\begin{aligned} \sum_{j=m+1}^{\infty} ar^{j-1} &= \sum_{j=m+1}^{\infty} \bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^2 \left(\frac{\bar{X}-1}{\bar{X}} \right)^{3j-3} \\ &= \frac{\bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^3 \left(\frac{\bar{X}-1}{\bar{X}} \right)^{3m+2}}{1 - \left(\frac{\bar{X}-1}{\bar{X}} \right)^3} \\ &= \frac{\bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^{3m+5}}{1 - \left(\frac{\bar{X}-1}{\bar{X}} \right)^3}. \end{aligned}$$

As a result we can express $T_n^{(3)}$ as,

$$T_n^{(3)} = \frac{1}{n} \sum_{t=2}^{\infty} \zeta_n^{(2)}(s)^{t-1} \frac{1}{\bar{X}} + \frac{\bar{X} s^{3m+5}}{n(1-s^3)},$$

with

$$s = \frac{\bar{X}-1}{\bar{X}}.$$

See Appendix A.2 for a more thorough treatment of some of the derivations. It is also important to note that all three tests reject the null hypothesis for large values. In Chapter 5 we will present the results of our newly proposed tests compared to the existing tests.

Chapter 5

Simulations and results

5.1 Introduction

In this chapter, we compare the performance of the newly proposed tests to the performance of the existing tests. In Section 5.2 we discuss the simulation setting and in Section 5.3 we provide the power estimates of the newly proposed tests against alternative tests. Lastly, in Section 5.4, we apply the new and existing tests to observed data examples.

5.2 Simulation setting

In this chapter, we analyse the performance of our newly proposed tests compared to existing tests against various alternative distributions. The power estimates were obtained through a Monte Carlo study with 10,000 independent replications drawn from various alternative distributions. These alternative distributions are listed in Table 5.1. The tests and comparisons were performed at a 5% significance level. We calculated and reported the power estimates for sample sizes 20, 40 and 60 which can be found in Tables 5.3 to 5.11. All the simulations and calculations were done in R ? using the Monte Carlo package ?. The critical values were obtained with the

bootstrap and $B = 1000$ iterations.

All the distributions discussed below are detailed in Table 5.1. An explanation of some of the notation follows. To generate random variates for the *geometric-geometric* distribution a vector, denoted as U is used. This vector consists of random uniform variables ranging between 0 and 1, and a chosen value y as a constant. A uniform vector is also employed by the *geometric-Poisson* distribution. The components of the *zero-inflated Poisson* distribution can be explained as follows. $\delta(x, 0)$ represents the Kronecker delta function, which equals 1 when x is zero and 0 otherwise. π denotes the probability of zero inflation, i.e., the probability of observing a zero. In contrast, p represents the parameter of the *Poisson* distribution. In the context of the *mixture Poisson* distribution, there exist several unknown values, defined as α_1 and α_2 , which respectively denote the mixing weights of the first and second *Poisson* distributions, and p_1 and p_2 , representing the parameters of the first and second *Poisson* distributions. The *beta-binomial* distribution, uses the *beta* function expressed as:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Lastly, we simulate random samples from the *Weibull-Poisson* distribution. This function takes key parameters such as the sample size n , the Poisson parameter λ , and *Weibull* distribution parameters a and b . To generate random variables from this distribution, we begin by initialising essential variables and computing weights for a range of values based on a quadratic function. To ensure the accuracy of our generated data, we employ an iterative process. Inside this loop, we gradually extend the range of values and recalculate weights and probabilities until the cumulative probability closely approximates 1. Once this criterion is met, we proceed to compute the CDF. Subsequently, we generate random values from this CDF using uniform random variables. The output is a vector of random values, which represents a sample from the *Weibull-Poisson* distribution.

Table 5.1: Probability mass functions for our choice of alternative distributions

Distribution	$f(x)$	Notation
Geometric	$(1-p)^x p + 1$	$G(p) + 1$
Geometric Geometric	$(U < y) (1-p)^x p + (U > y) (1-p)^x p + 1$	$GG(p_1, p_2, y) + 1$
Poisson	$px \frac{e^p}{x!} + 1$	$Poi(p) + 1$
Zero Inflated Poisson	$(1-\pi)e^{-p} \frac{p^{x_1}}{x_1!} + \pi \delta(x, 0) + 1$	$ZIP(\pi, p) + 1$
Mixture Poisson	$\alpha_1 p_1 x \frac{e^{p_1}}{x!} + (1-\alpha_2) p_2 x \frac{e^{p_2}}{x!} + 1$	$MP(\alpha_1, \alpha_2, p_1, p_2) + 1$
Geometric Poisson	$(U < y) (1-p)^x p + (U > y) px \frac{e^p}{x!} + 1$	$GP(p_1, p_2, y) + 1$
Binomial	$\binom{n}{x} p^x (1-p)^{(n-x)} + 1$	$B(n, p) + 1$
Negative Binomial	$\binom{x+r-1}{x} \frac{p^r (1-p)^x}{B(x+\alpha, n-x+\beta)} + 1$	$NB(r, p) + 1$
Beta-Binomial	$\binom{n}{x} \frac{p^x (1-p)^{(n-x)}}{B(\alpha, \beta)} + 1$	$BB(n, \alpha, \beta) + 1$
Weibull	$\frac{p}{\lambda} \left(\frac{x}{\lambda}\right)^{p-1} \exp\left(-\frac{x}{\lambda}\right)^p + 1$	$W(p, \lambda) + 1$
Weibull Poisson	See Section 5.2	$WP(\lambda, a, b) + 1$
Uniform	$\frac{1}{b-a} + 1$	$U(a, b) + 1$

The existing tests that are used for comparison were discussed in Chapter 3 and are presented in Table 5.2, which displays the "Test Statistics", "Section", "Section heading", and "Reference" for the test statistics used in our simulation study. The estimated powers of the $B_{n,\gamma}$ test statistic depend on a tuning parameter γ . We calculated the powers for this test statistic using two fixed values of γ , namely, $\gamma = 5$ and $\gamma = 1$. The test statistic $G_{n,\gamma}$ also depends on γ . In our simulation, we followed the suggestion of the authors and chose $\gamma = 5$.

5.3 Simulation results

In this section, we will discuss the results of the estimated powers of the different tests considered.

In this chapter, we will focus on the results of our newly proposed test statistics namely T_n^1, T_n^2 and T_n^3 from Chapter 4. For easy comparison, only the maximum powers for each distribution are made bold.

Taking an overall view of the tables we observe that our newly proposed tests have relative higher power results against alternative distributions in most cases with the exception of the mixed distributions. Furthermore, we note that increasing sample sizes leads to better power estimates. In particular, T_n^3 has notably better results for sample size 60 compared to its results for sample size 20.

Considering our newly proposed tests in isolation they perform very well compared to the existing tests. This is evident for T_n^3 . Looking at each table individually it is clear that T_n^3 and T_n^2 outperform the existing tests for almost all the alternative distributions.

In fact, it becomes evident that T_n^3 consistently performs exceptionally well, exhibiting the highest statistical power across most scenarios. The exception is the $ZIP(2, 0.8)+1$

Test Statistic	Section	Details	Reference
$G_{n,\gamma}$	3.4.1.	Test statistic based on linear regression of order statistics	Jiménez-Gamero and Alba-Fernández (2021)
R_n	3.4.2.	Test based on the probability-generating function	Rueda and O'Reilly (1999)
H_n	3.4.3.	Test based on empirical distribution function	Henze (1996)
K_n	3.4.5.	Test based on the integrated distribution function	Klar (1999)
$B_{n,\gamma}$	3.4.6.	Test based on the probability generating function	Milošević et al. (2021)
J_n	3.4.7.	Test based on empirical characteristic functions	Jiménez-Gamero and Batsidis (2017)

Table 5.2: Notation and section reference for the choice of test statistics

distribution. For this specific distribution, both T_n^2 and T_n^3 fall short, failing to provide any substantial results for sample sizes of $n = 20$ and $n = 40$.

It is noted that the H_n test is the weakest performing test on the alternative distributions by consistently obtaining the lowest estimated powers. Looking at Tables 5.6 and 5.7, which are the estimated powers for alternative distributions of sample size $n = 40$, we observe that test H_n and test J_n have the lowest estimated powers. J_n is also the weakest performing test in Tables 5.9 and 5.10, which are the estimated powers for alternative distributions of sample size $n = 60$.

However, if we consider the mixed distributions our newly proposed tests exhibit more diverse results. In Table 5.5, the estimated powers for mixed distributions of sample size $n = 20$ the proposed tests T_n^1, T_n^2 and T_n^3 mostly do not discern the mixture *geometric-geometric* distributions.

Considering Table 5.8, the estimated powers for mixed distributions of sample size $n = 40$, we observe that only T_n^1 has very low powers for the mixture *geometric-geometric* distributions. Observing Table 5.11, the estimated powers for mixed distributions of sample size $n = 60$, it is clear that T_n^1 performs slightly better.

Our newly proposed tests are the weakest performing tests on the mixture distributions but surprisingly H_n is one of the strongest performing tests. H_n often obtains the maximum power estimates for all the mixture distribution tables.

Considering all the tables it can clearly be noted that all tests reach nominal significance levels. However, it is interesting to note that J_n fails to reach the nominal significance level for $G(0.8)$ and $G(0.9)$ over all sample sizes.

In conclusion, our newly proposed tests consistently achieve the expected nominal significance levels, particularly for larger sample sizes, as is demonstrated in Table 5.9. Generally, statistical power increases with sample size, as evident when comparing

Table 5.3 to Table 5.9. Our tests outperform several of the existing tests, even with smaller sample sizes, as shown in Tables 5.3 and 5.4. However, we recommend using our newly proposed tests for sample sizes greater than $n = 40$.

5.4 Observed Data Application

In this section, we apply our newly proposed test statistics as well as the existing test statistics to observed data sets and calculate the bootstrap *p-values* in order to demonstrate the use of the tests. The hypothesis to be tested is that the data are *geometrically* distributed and the *p-values* are given in Tables 5.12 to 5.15.

The observed data sets for the tests were taken from examples in Jiménez-Gamero and Alba-Fernández (2021). We calculated the *p-values*, ρ , with the bootstrap. The calculations we performed by applying our proposed test statistics and the existing test statistics to the observed data set, $X_1, \dots, X_n$. The test statistic values were then defined by $t_n = T_n(X_1, \dots, X_n)$. Data was generated from the *geometric* distribution and denoted, $X_1^*, \dots, X_n^*$. The newly proposed test statistics and the existing test statistics were calculated on the newly generated sample and denoted $T_n^* = T_n(X_1^*, \dots, X_n^*)$. We repeated the previous steps for $MC = 1000$ simulations to obtain a vector of test statistics, $T_{n,1}^*, \dots, T_{n,MC}^*$. We calculated the *p-values* using the following formula:

$$\begin{aligned} \rho &= P(T_n > t_n \mid H_0 \text{ is True}) \\ \rho_{MC} &= \frac{1}{MC} \sum_{m=1}^{MC} \mathbb{1}(T_{n,m}^* \leq t_n), \end{aligned}$$

where ρ_{MC} is the Monte Carlo *p-values*. The *p-values* for the observed data sets are listed in Tables 5.12 to 5.15 below.

The tables consistently indicate high *p-values* for all test statistics, surpassing the

Table 5.3: Estimated powers for alternative distributions of sample size $n = 20$.

	R_n	H_n	K_n	J_n	$B_{n,5}$	$B_{n,1}$	$G_{n,5}$	T_n^1	T_n^2	T_n^3
$Poi(0.5) + 1$	0.09	0.14	0.1	0.12	0.07	0.08	0.11	0.1	0.21	0.19
$Poi(1) + 1$	0.39	0.31	0.29	0.36	0.36	0.37	0.39	0.3	0.45	0.47
$Poi(1.5) + 1$	0.66	0.42	0.5	0.57	0.65	0.66	0.64	0.5	0.68	0.7
$NB(5, 0.89) + 1$	0.09	0.13	0.09	0.12	0.08	0.09	0.12	0.09	0.18	0.17
$NB(5, 0.83) + 1$	0.24	0.19	0.16	0.22	0.21	0.22	0.24	0.16	0.28	0.3
$NB(5, 0.75) + 1$	0.44	0.23	0.28	0.36	0.42	0.43	0.41	0.28	0.44	0.47
$B(10, 0.05) + 1$	0.12	0.18	0.13	0.15	0.09	0.11	0.14	0.13	0.25	0.23
$B(10, 0.1) + 1$	0.49	0.42	0.39	0.46	0.47	0.49	0.5	0.39	0.56	0.58
$B(10, 0.15) + 1$	0.8	0.58	0.67	0.72	0.79	0.81	0.79	0.67	0.83	0.84
$B(2, 0.5) + 1$	0.94	0.93	0.93	0.94	0.93	0.94	0.94	0.93	0.99	0.99
$B(4, 0.25) + 1$	0.68	0.61	0.58	0.65	0.66	0.67	0.68	0.57	0.75	0.78
$B(1, 0.5) + 1$	0.71	0.88	0.81	0.79	0.66	0.68	0.76	0.81	0.95	0.94
$W(0.4, 1.4) + 1$	0.11	0.17	0.12	0.14	0.09	0.1	0.13	0.12	0.23	0.21
$W(0.6, 1.4) + 1$	0.24	0.19	0.17	0.23	0.22	0.23	0.24	0.17	0.29	0.31
$W(0.25, 2) + 1$	0.03	0.11	0.06	0.06	0.02	0.03	0.05	0.06	0.22	0.2
$W(0.75, 2) + 1$	0.82	0.71	0.7	0.78	0.8	0.81	0.81	0.71	0.84	0.86
$ZIP(2, 0.8) + 1$	0.4	0.48	0.37	0.29	0.39	0.4	0.26	0.36	0	0
$ZIP(1, 0.1) + 1$	0.22	0.19	0.17	0.21	0.2	0.21	0.23	0.17	0.3	0.32
$ZIP(2, 0.2) + 1$	0.23	0.15	0.23	0.22	0.24	0.24	0.23	0.22	0.38	0.38
$BB(9, 1, 2) + 1$	0.25	0.12	0.23	0.18	0.27	0.26	0.24	0.24	0.43	0.48
$BB(9, 0.5, 1) + 1$	0.04	0.08	0.06	0.11	0.04	0.04	0.06	0.07	0.17	0.16
$BB(3, 1, 2) + 1$	0.27	0.2	0.25	0.24	0.27	0.27	0.26	0.24	0.48	0.49

Table 5.4: Estimated powers for alternative distributions of sample size $n = 20$.

	R_n	H_n	K_n	J_n	$B_{n,5}$	$B_{n,1}$	$G_{n,5}$	T_n^1	T_n^2	T_n^3
$U(1, 5) + 1$	0.64	0.6	0.77	0.68	0.67	0.68	0.68	0.76	0.92	0.93
$U(1, 10) + 1$	0.55	0.5	0.75	0.56	0.63	0.6	0.64	0.75	0.91	0.91
$U(3, 10) + 1$	1	0.98	1	1	1	1	1	1	1	1
$MP(0.5, 1.5, 0.25, 0.75) + 1$	0.36	0.24	0.26	0.31	0.35	0.36	0.35	0.26	0.42	0.44
$MP(0.5, 1.5, 0.5, 0.5) + 1$	0.18	0.15	0.12	0.17	0.16	0.17	0.18	0.12	0.23	0.25
$MP(0.5, 1.5, 0.75, 0.25) + 1$	0.11	0.12	0.08	0.12	0.09	0.1	0.12	0.09	0.17	0.16
$MP(0.33, 1, 0.25, 0.75) + 1$	0.2	0.19	0.15	0.2	0.17	0.18	0.2	0.15	0.27	0.28
$MP(0.33, 1, 0.5, 0.5) + 1$	0.09	0.11	0.09	0.11	0.07	0.09	0.1	0.08	0.17	0.16
$MP(0.33, 1, 0.75, 0.25) + 1$	0.05	0.08	0.05	0.07	0.04	0.04	0.05	0.06	0.13	0.11
$WP(0.5, 0.75, 0.25) + 1$	0.42	0.34	0.33	0.39	0.4	0.42	0.42	0.33	0.52	0.54
$WP(0.5, 0.5, 0.5) + 1$	0.41	0.35	0.33	0.4	0.39	0.4	0.42	0.33	0.49	0.53
$WP(1, 0.75, 0.25) + 1$	0.89	0.72	0.82	0.83	0.9	0.9	0.88	0.81	0.91	0.92
$WP(1, 0.5, 0.5) + 1$	0.87	0.66	0.76	0.79	0.88	0.87	0.85	0.76	0.88	0.9
$WP(1.5, 0.75, 0.25) + 1$	0.99	0.91	0.96	0.96	0.99	0.99	0.98	0.96	0.98	0.99

Table 5.5: Estimated powers for mixed distributions of sample size $n = 20$.

	R_n	H_n	K_n	J_n	$B_{n,5}$	$B_{n,1}$	G_n	T_n^1	T_n^2	T_n^3
$G(0.1) + 1$	0.05	0.05	0.04	0.05	0.06	0.05	0.04	0.04	0.03	0.05
$G(0.2) + 1$	0.06	0.05	0.04	0.05	0.06	0.06	0.07	0.04	0.04	0.05
$G(0.3) + 1$	0.05	0.05	0.04	0.05	0.06	0.05	0.02	0.04	0.05	0.05
$G(0.4) + 1$	0.06	0.06	0.04	0.05	0.05	0.05	0.02	0.04	0.05	0.05
$G(0.5) + 1$	0.05	0.05	0.04	0.05	0.05	0.05	0.09	0.04	0.05	0.05
$G(0.6) + 1$	0.04	0.06	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.05
$G(0.7) + 1$	0.03	0.05	0.03	0.02	0.03	0.03	0.01	0.03	0.04	0.04
$G(0.8) + 1$	0.02	0.03	0.02	0.01	0.02	0.02	0	0.02	0.03	0.02
$G(0.9) + 1$	0.01	0.01	0.01	0	0.01	0.01	0	0.02	0	0
$GG(0.3, 0.6, 0.1) + 1$	0.09	0.1	0.08	0.05	0.1	0.09	0.08	0.08	0.03	0.03
$GG(0.3, 0.6, 0.2) + 1$	0.12	0.12	0.11	0.06	0.13	0.13	0.14	0.11	0.02	0.03
$GG(0.3, 0.6, 0.3) + 1$	0.14	0.14	0.12	0.07	0.14	0.13	0.08	0.11	0.02	0.02
$GG(0.3, 0.6, 0.4) + 1$	0.13	0.14	0.11	0.07	0.13	0.13	0.08	0.11	0.02	0.02
$GG(0.3, 0.6, 0.5) + 1$	0.12	0.14	0.1	0.06	0.12	0.12	0.07	0.1	0.01	0.02
$GG(0.3, 0.6, 0.6) + 1$	0.1	0.13	0.08	0.06	0.11	0.1	0.14	0.08	0.02	0.02
$GG(0.3, 0.6, 0.7) + 1$	0.08	0.11	0.07	0.05	0.09	0.08	0.06	0.06	0.02	0.02
$GG(0.3, 0.6, 0.8) + 1$	0.07	0.09	0.05	0.05	0.07	0.07	0.05	0.05	0.03	0.03
$GG(0.3, 0.6, 0.9) + 1$	0.06	0.07	0.04	0.05	0.06	0.06	0.03	0.04	0.04	0.04
$GP(0.3, 1, 0.1) + 1$	0.29	0.24	0.21	0.29	0.26	0.27	0.22	0.21	0.31	0.32
$GP(0.3, 1, 0.2) + 1$	0.2	0.18	0.16	0.22	0.2	0.19	0.19	0.15	0.2	0.21
$GP(0.3, 1, 0.3) + 1$	0.16	0.14	0.11	0.18	0.14	0.15	0.09	0.12	0.15	0.14
$GP(0.3, 1, 0.4) + 1$	0.12	0.12	0.09	0.13	0.12	0.12	0.12	0.09	0.1	0.11
$GP(0.3, 1, 0.5) + 1$	0.11	0.09	0.08	0.11	0.09	0.09	0.11	0.08	0.08	0.08
$GP(0.3, 1, 0.6) + 1$	0.07	0.08	0.06	0.09	0.08	0.08	0.04	0.06	0.06	0.06
$GP(0.3, 1, 0.7) + 1$	0.07	0.07	0.05	0.07	0.07	0.07	0.09	0.05	0.05	0.05
$GP(0.3, 1, 0.8) + 1$	0.06	0.06	0.04	0.06	0.07	0.06	0.03	0.04	0.05	0.05

Table 5.6: Estimated powers for alternative distributions of sample size $n = 40$.

	R_n	H_n	K_n	J_n	$B_{n,5}$	$B_{n,1}$	G_n	T_n^1	T_n^2	T_n^3
$Poi(0.5) + 1$	0.23	0.28	0.24	0.19	0.21	0.21	0.26	0.24	0.34	0.37
$Poi(1) + 1$	0.71	0.64	0.63	0.65	0.7	0.7	0.72	0.64	0.73	0.75
$Poi(1.5) + 1$	0.94	0.89	0.88	0.88	0.94	0.94	0.94	0.89	0.91	0.94
$NB(5, 0.89) + 1$	0.2	0.23	0.19	0.18	0.19	0.2	0.24	0.2	0.28	0.31
$NB(5, 0.83) + 1$	0.46	0.4	0.38	0.41	0.44	0.44	0.47	0.38	0.49	0.51
$NB(5, 0.75) + 1$	0.75	0.62	0.63	0.62	0.75	0.75	0.74	0.64	0.68	0.73
$B(10, 0.05) + 1$	0.29	0.35	0.32	0.25	0.27	0.27	0.34	0.3	0.41	0.44
$B(10, 0.1) + 1$	0.83	0.78	0.78	0.78	0.83	0.83	0.83	0.79	0.85	0.88
$B(10, 0.15) + 1$	0.99	0.96	0.97	0.96	0.99	0.99	0.98	0.97	0.98	.99
$B(2, 0.5) + 1$	1	1	1	1	1	1	1	1	1	1
$B(4, 0.25) + 1$	0.96	0.93	0.94	0.93	0.96	0.95	0.96	0.94	0.97	0.98
$B(1, 0.5) + 1$	0.99	1	1	0.97	0.98	0.99	0.99	1	1	1
$W(0.4, 1.4) + 1$	0.26	0.32	0.27	0.22	0.23	0.24	0.29	0.27	0.37	0.39
$W(0.6, 1.4) + 1$	0.48	0.41	0.38	0.42	0.45	0.46	0.48	0.4	0.5	0.52
$W(0.25, 2) + 1$	0.18	0.34	0.28	0.07	0.13	0.15	0.2	0.28	0.59	0.58
$W(0.75, 2) + 1$	0.99	0.98	0.98	0.97	0.99	0.99	0.99	0.98	0.99	0.99
$ZIP(2, 0.8) + 1$	0.77	0.85	0.8	0.56	0.75	0.76	0.7	0.8	0	0.01
$ZIP(1, 0.1) + 1$	0.45	0.38	0.4	0.39	0.43	0.46	0.45	0.39	0.54	0.56
$ZIP(2, 0.2) + 1$	0.45	0.35	0.54	0.4	0.49	0.48	0.48	0.53	0.66	0.7
$BB(9, 1, 2) + 1$	0.47	0.35	0.61	0.32	0.53	0.52	0.51	0.62	0.85	0.86
$BB(9, 0.5, 1) + 1$	0.04	0.14	0.2	0.21	0.05	0.05	0.09	0.2	0.47	0.45
$BB(3, 1, 2) + 1$	0.57	0.48	0.62	0.49	0.59	0.58	0.56	0.61	0.9	0.9

Table 5.8: Estimated powers for mixed distributions of sample size $n = 40$.

	R_n	H_n	K_n	J_n	$B_{n,5}$	$B_{n,1}$	G_n	T_n^1	T_n^2	T_n^3
$G(0.1) + 1$	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.05
$G(0.2) + 1$	0.06	0.06	0.05	0.04	0.06	0.05	0.06	0.05	0.05	0.05
$G(0.3) + 1$	0.05	0.06	0.04	0.04	0.06	0.05	0.05	0.05	0.05	0.05
$G(0.4) + 1$	0.06	0.06	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.06
$G(0.5) + 1$	0.06	0.06	0.04	0.05	0.05	0.05	0.05	0.04	0.05	0.05
$G(0.6) + 1$	0.05	0.06	0.04	0.04	0.05	0.05	0.04	0.05	0.05	0.05
$G(0.7) + 1$	0.04	0.05	0.04	0.02	0.05	0.04	0.03	0.04	0.04	0.05
$G(0.8) + 1$	0.03	0.04	0.04	0	0.03	0.03	0.02	0.04	0.05	0.05
$G(0.9) + 1$	0.02	0.03	0.03	0	0.02	0.02	0	0.03	0.01	0.01
$GG(0.3, 0.6, 0.1) + 1$	0.13	0.11	0.13	0.06	0.15	0.14	0.1	0.13	0.02	0.03
$GG(0.3, 0.6, 0.2) + 1$	0.18	0.17	0.19	0.08	0.21	0.19	0.14	0.18	0.02	0.01
$GG(0.3, 0.6, 0.3) + 1$	0.2	0.2	0.2	0.1	0.23	0.21	0.17	0.2	0.01	0.01
$GG(0.3, 0.6, 0.4) + 1$	0.19	0.21	0.2	0.1	0.22	0.2	0.17	0.2	0.01	0.01
$GG(0.3, 0.6, 0.5) + 1$	0.19	0.19	0.17	0.09	0.2	0.19	0.16	0.17	0.01	0.01
$GG(0.3, 0.6, 0.6) + 1$	0.15	0.17	0.14	0.08	0.16	0.16	0.14	0.14	0.01	0.01
$GG(0.3, 0.6, 0.7) + 1$	0.12	0.14	0.1	0.06	0.12	0.12	0.11	0.1	0.01	0.02
$GG(0.3, 0.6, 0.8) + 1$	0.09	0.1	0.07	0.05	0.09	0.09	0.08	0.08	0.02	0.02
$GG(0.3, 0.6, 0.9) + 1$	0.06	0.08	0.06	0.05	0.07	0.06	0.06	0.06	0.03	0.04
$GP(0.3, 1, 0.1) + 1$	0.47	0.48	0.4	0.49	0.44	0.45	0.5	0.4	0.39	0.43
$GP(0.3, 1, 0.2) + 1$	0.3	0.34	0.27	0.36	0.28	0.29	0.34	0.26	0.22	0.25
$GP(0.3, 1, 0.3) + 1$	0.2	0.24	0.18	0.27	0.19	0.2	0.24	0.18	0.13	0.15
$GP(0.3, 1, 0.4) + 1$	0.15	0.18	0.13	0.2	0.14	0.14	0.17	0.13	0.09	0.09
$GP(0.3, 1, 0.5) + 1$	0.11	0.13	0.1	0.14	0.1	0.1	0.12	0.1	0.06	0.07
$GP(0.3, 1, 0.6) + 1$	0.08	0.09	0.08	0.1	0.08	0.09	0.09	0.08	0.05	0.05
$GP(0.3, 1, 0.7) + 1$	0.07	0.08	0.06	0.08	0.07	0.07	0.07	0.07	0.04	0.04
$GP(0.3, 1, 0.8) + 1$	0.06	0.07	0.06	0.06	0.06	0.06	0.07	0.06	0.04	0.04

Table 5.9: Estimated powers for alternative distributions of sample size $n = 60$.

	R_n	H_n	K_n	J_n	$B_{n,5}$	$B_{n,1}$	G_n	T_n^1	T_n^2	T_n^3
$Poi(0.5) + 1$	0.37	0.41	0.36	0.25	0.33	0.35	0.39	0.37	0.45	0.49
$Poi(1) + 1$	0.89	0.84	0.84	0.82	0.87	0.89	0.89	0.84	0.84	0.89
$Poi(1.5) + 1$	0.99	0.98	0.98	0.97	0.99	0.99	0.99	0.98	0.97	0.99
$NB(5, 0.89) + 1$	0.32	0.33	0.3	0.26	0.31	0.3	0.34	0.31	0.37	0.42
$NB(5, 0.83) + 1$	0.64	0.58	0.56	0.56	0.62	0.64	0.65	0.57	0.61	0.65
$NB(5, 0.75) + 1$	0.9	0.83	0.84	0.8	0.9	0.9	0.9	0.84	0.79	0.87
$B(10, 0.05) + 1$	0.47	0.49	0.45	0.34	0.44	0.46	0.49	0.47	0.54	0.6
$B(10, 0.1) + 1$	0.96	0.93	0.94	0.92	0.96	0.96	0.95	0.94	0.94	0.96
$B(10, 0.15) + 1$	1	1	1	1	1	1	1	1	1	1
$B(2, 0.5) + 1$	1	1	1	1	1	1	1	1	1	1
$B(4, 0.25) + 1$	1	0.99	0.99	0.99	1	1	1	0.99	1	1
$B(1, 0.5) + 1$	1	1	1	1	1	1	1	1	1	1
$W(0.4, 1.4) + 1$	0.42	0.44	0.42	0.31	0.4	0.4	0.44	0.41	0.48	0.53
$W(0.6, 1.4) + 1$	0.66	0.6	0.58	0.58	0.65	0.66	0.68	0.59	0.62	0.69
$W(0.25, 2) + 1$	0.42	0.58	0.54	0.08	0.38	0.39	0.43	0.54	0.76	0.78
$W(0.75, 2) + 1$	1	1	1	1	1	1	1	1	1	1
$ZIP(2, 0.8) + 1$	0.93	0.96	0.95	0.71	0.91	0.92	0.91	0.95	0	0.05
$ZIP(1, 0.1) + 1$	0.65	0.56	0.6	0.55	0.64	0.64	0.64	0.6	0.67	0.73
$ZIP(2, 0.2) + 1$	0.65	0.56	0.75	0.57	0.69	0.68	0.68	0.75	0.81	0.85
$BB(9, 1, 2) + 1$	0.66	0.59	0.85	0.47	0.74	0.72	0.71	0.85	0.98	0.98
$BB(9, 0.5, 1) + 1$	0.04	0.21	0.36	0.34	0.06	0.05	0.13	0.37	0.76	0.73
$BB(3, 1, 2) + 1$	0.79	0.69	0.86	0.71	0.82	0.81	0.77	0.86	0.99	0.99

Table 5.11: Estimated powers for mixed distributions of sample size $n = 60$.

[illegible]

significance level of 0.05. As a result, none of the tests reject the null hypothesis. We conclude that the *geometric* distribution is an appropriate model in all of the cases considered.

Table 5.12: Bootstrap *p-values* for earthquake data

	R_n	H_n	K_n	J_n	G_n	B_n	T_n^1	T_n^2	T_n^3
1	0.537	0.593	0.766	0.341	0.587	0.634	0.766	0.543	0.719

Table 5.13: Bootstrap *p-values* for discentric data 1

	R_n	H_n	K_n	J_n	G_n	B_n	T_n^1	T_n^2	T_n^3
1	0.654	0.52	0.641	0.529	0.874	0.686	0.641	0.718	0.75

Table 5.14: Bootstrap *p-values* for discentric data 2

	R_n	H_n	K_n	J_n	G_n	B_n	T_n^1	T_n^2	T_n^3
1	0.204	0.139	0.167	0.446	0.215	0.242	0.167	0.333	0.222

Table 5.15: Bootstrap *p-values* for birth data - smokers

	R_n	H_n	K_n	J_n	G_n	B_n	T_n^1	T_n^2	T_n^3
1	0.583	0.512	0.707	0.267	0.476	0.658	0.707	0.213	0.273

Chapter 6

Conclusion and future research

6.1 Introduction

The purpose of this dissertation was to develop new goodness of fit tests for the *geometric* distribution based on a characterisation and to perform a Monte Carlo investigation into the newly proposed tests in order to observe their performance compared to existing tests. Our newly proposed tests were developed and discussed in Chapter 4. We discussed the characterisations and theory of the existing tests in Chapter 3. We can summarize the main objectives and the outcomes of these objectives as follows:

- Research the literature on the *geometric* distribution in relation to its characterisations and goodness-of-fit tests.

The literature on the *geometric* distribution in relation to its characterisations and goodness-of-fit tests was discussed in Chapter 3.

- Introduce new goodness-of-fit tests for the *geometric* distribution.

We have presented our newly proposed tests in Chapter 4 and have included all necessary proofs and definitions.

- Perform a Monte Carlo simulation study to compare the empirical powers of

the newly proposed goodness-of-fit tests to those of existing tests. Our Monte Carlo study was performed on our newly proposed tests as well as on existing tests in Chapter 5. We have included all necessary results and graphs to clearly illustrate the performance of our newly proposed tests.

- Apply the new goodness-of-fit tests to observed data.

The newly proposed tests as well as the existing tests were applied to observed data sets in Chapter 5 Section 5.2.

6.2 Concluding remarks and future research

The most important aim of this study was to develop new goodness-of-fit tests for the *geometric* distribution and compare these tests' performance to existing goodness-of-fit tests for the *geometric* distribution. The Monte Carlo study revealed that T_n^2 and T_n^3 were the superior performing tests. Tests T_n^2 and T_n^3 continually obtained or matched the highest power estimates achieved by the existing tests.

Test H_n and J_n were the weakest performing tests. However, the exception to the norm came in analysing the tables with the mixed distributions where it is evident that test T_n^2 performs weakly. T_n^2 consistently obtained the lowest power estimates where test H_n was the highest performing test obtaining the highest power estimates.

The newly proposed tests perform best for large sample sizes. For future research, we might consider investigating the asymptotic theory of our newly proposed tests.

On a final note, a brief investigation was launched on the characterisation proposed by Maiti and Biswas (2007) as presented in Chapter 3. As one of the potential future tests setting this characterisation equal to zero, we obtain:

$$S(j) = \frac{p \sin(j)}{1 - p \cos(j)} C(j) \text{ for all } j$$

then with some manipulation of the formula we obtained,

$$S(j) - S(j)p \cos(j) - p \sin(j)C(j) = 0$$

Replacing the quantities on the left-hand side of this equation and integrating the result we obtain,

$$\delta_n = \int_0^\infty \left[\hat{S}(j) - \hat{S}(j)\hat{p} \cos(j) - \hat{p} \sin(j)\hat{C}(j) \right]^2 w(j) dj$$

where $w(j)$ is a weight function, $\hat{S}(j)$ and $\hat{C}(j)$ are the estimated functions of $\sin$ and $\cos$. Going forward we suppress the bounds on the integrand and derive the integrand.

$$\begin{aligned} \delta_n &= \int \left[\frac{1}{n} \sum_{k=1}^n \sin(jX_k) - \frac{1}{n} \sum_{k=1}^n \sin(jX_k) \hat{p} \cos(j) - \hat{p} \sin(j) \frac{1}{n} \sum_{k=1}^n \cos(jX_k) \right]^2 w(j) dj \\ &= \int \left[\frac{1}{n} \sum_{k=1}^n \sin(jX_k) - \hat{p} \frac{1}{n} \sum_{k=1}^n (\sin(jX_k) \cos(j) + \sin(j) \cos(jX_k)) \right]^2 w(j) dj \end{aligned}$$

Recalling that,

$$\sin(\alpha + \beta) = \sin(\alpha) \cos(\beta) + \cos(\alpha) \sin(\beta),$$

the integral can be written as,

$$\delta_n = \int \left[\frac{1}{n} \sum_{k=1}^n \sin(jX_k) - \hat{p} \frac{1}{n} \sum_{k=1}^n \sin(j + jX_k) \right]^2 w(j) dj.$$

A closed form of the integral is,

$$\delta_n = \int \left[\hat{S}(j) - \hat{p} \frac{1}{n} \sum_{k=1}^n \sin(j[X_k + 1]) \right]^2 w(j) dj$$

$$\begin{aligned}
&= \int \widehat{S}(j)^2 w(j) dj + \widehat{p}^2 \int \left[\frac{1}{n} \sum_{k=1}^n (\sin(j[X_k + 1])) \right]^2 \\
&- 2\widehat{p} \int \left[\widehat{S}(j) \frac{1}{n} \sum_{k=1}^n \sin(j[X_k + 1]) \right] w(j) dj \\
&= \int \left[\frac{1}{n} \sum_{k=1}^n \sin(jX_k) \right]^2 w(j) dj + \widehat{p}^2 \int \left[\frac{1}{n} \sum_{k=1}^n (\sin(t[X_j] + 1)) \right] w(j) dj \\
&- 2\widehat{p} \int \left[\frac{1}{n} \sum_{k=1}^n \sin(jX_k) \frac{1}{n} (\sin(j[X_k + 1])) \right] w(j) dj \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \int \sin(jX_i) \sin(jX_k) w(j) dj \\
&+ \widehat{p}^2 \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \int \sin(j[X_i + 1]) \sin(j[X_k + 1]) w(j) dj \\
&- 2\widehat{p} \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \int \sin(j[X_i]) \sin(j[X_k + 1]) w(j) dj
\end{aligned}$$

$$\delta_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \left[\begin{aligned} &\int \sin(jX_i) \sin(jX_k) w(j) dj \\ &+ \widehat{p}^2 \int \sin(j[X_i + 1]) \sin(j[X_k + 1]) w(j) dj \\ &- 2\widehat{p} \int \sin(j[X_i]) \sin(j[X_k + 1]) w(j) dj \end{aligned} \right].$$

Now let,

$$I_w(\alpha, \beta) = \int \sin(j\alpha) \sin(j\beta) w(j) dj, \quad (6.1)$$

then

$$\delta_n = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^n I_w(X_i, X_k) + \widehat{p}^2 I_w(X_i + 1, X_k + 1) - 2\widehat{p} I_w(X_i, X_k + 1).$$

Choosing different weight functions, $w(j)$, produces different integral estimates. For

example, choosing a Gaussian weight function,

$$w(j) = e^{\frac{-1}{2}\gamma j^2},$$

the term $I_w(\alpha, \beta)$ becomes,

$$I_w(\alpha, \beta) = \sqrt{\frac{\pi}{2\gamma}} \left[e^{\frac{-(\alpha-\beta)^2}{2\gamma}} - e^{\frac{-(\alpha+\beta)^2}{2\gamma}} \right],$$

therefore,

$$\delta_n^{(1)} = \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \left(\begin{aligned} & \sqrt{\frac{\pi}{2\gamma}} \left[e^{\frac{-(X_i - X_k)^2}{2\gamma}} - e^{\frac{-(X_i + X_k)^2}{2\gamma}} \right] \\ & + \hat{p}^2 \sqrt{\frac{\pi}{2\gamma}} \left[e^{\frac{-([X_i+1] - [X_k+1])^2}{2\gamma}} - e^{\frac{-([X_i+1] + [X_k+1])^2}{2\gamma}} \right] \\ & - 2\hat{p} \sqrt{\frac{\pi}{2\gamma}} \left[e^{\frac{-(X_i - [X_k+1])^2}{2\gamma}} - e^{\frac{-(X_i + [X_k+1])^2}{2\gamma}} \right] \end{aligned} \right).$$

Maiti and Biswas (2007) noted that the characterisation of the *geometric* distribution is not linear in j for $\frac{S(j)}{C(j)}$ since the *geometric* distribution is a periodic function with period 2π . This may have an effect on the performance of the test statistic and on the derivation of the closed form. This will require further investigation and for this reason, was left for future research. In addition, see Appendix A.3 for the result when $w(j)$ is chosen as the Laplace or Cauchy weight functions.

Appendix A

A.1 Chapter 3

The maximum likelihood estimator of the *geometric* distribution as in 3.1 can be derived as follows:

First, write the likelihood function as

$$L(p) = p(1-p)^{X_1-1} p(1-p)^{X_2-1} \dots p(1-p)^{X_n-1} = p^n (1-p)^{\sum_{i=1}^n X_i - n}.$$

Take logarithm on both sides,

$$\ln L(p) = n \ln p + \left(\sum_{i=1}^n X_i - n \right) \ln(1-p),$$

differentiating on both sides and setting it to zero we get

$$\frac{d \ln L(p)}{dp} = \frac{n}{\hat{p}} - \frac{(\sum_{i=1}^n X_i - n)}{1 - \hat{p}} = 0,$$

thus

$$\hat{p} = \frac{n}{\sum_{i=1}^n X_i - n} = \frac{1}{\bar{X} - 1}.$$

A.2 Chapter 4

Related to 4.4 consider the following:

$$\begin{aligned}
 S_m &= \sum_{j=1}^n ar^{j-1} = a + ar + \dots + ar^m \\
 rS_m &= ar + \dots + ar^m + ar^{m+1} \\
 \therefore S_m - rS_m &= a - ar^{m+1} \\
 \therefore S_m(1-r) &= a(1-r^{m+1}) \\
 \therefore S_m &= \frac{a(1-r^{m+1})}{(1-r)}.
 \end{aligned}$$

Now, we have that,

$$\begin{aligned}
 \sum_{j=m+1}^{\infty} ar^{j-1} &= \sum_{j=1}^{\infty} ar^{j-1} - \sum_{j=1}^m ar^{j-1} \\
 &= S_{\infty} - S_m \\
 &= \frac{a}{1-r} - \frac{a(1-r^{m+1})}{(1-r)} \\
 &= \frac{a - a(1-r^{m+1})}{1-r} \\
 &= \frac{a[1 - (1-r^{m+1})]}{1-r} \\
 &= \frac{ar^{m+1}}{1-r}.
 \end{aligned}$$

Continuing we set $a = \bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^2$ and $r = \left(\frac{\bar{X}-1}{\bar{X}} \right)^3$ we have that:

$$\begin{aligned}
 \sum_{j=m+1}^{\infty} ar^{j-1} &= \sum_{j=m+1}^{\infty} \bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^2 \left(\frac{\bar{X}-1}{\bar{X}} \right)^{3j-3} \\
 &= \sum_{j=m+1}^{\infty} \bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^{3j-1}, r = \left(\frac{\bar{X}-1}{\bar{X}} \right)^3 \in (0, 1)
 \end{aligned}$$

$$\begin{aligned}
&= \frac{\bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^3 \left(\frac{\bar{X}-1}{\bar{X}} \right)^{3m+2}}{1 - \left(\frac{\bar{X}-1}{\bar{X}} \right)^3} \\
&= \frac{\bar{X} \left(\frac{\bar{X}-1}{\bar{X}} \right)^{3m+5}}{1 - \left(\frac{\bar{X}-1}{\bar{X}} \right)^3}.
\end{aligned}$$

A.3 Chapter 6

Continuing the expansion on the weight functions of 6.1 we choose a second weight function as a Laplace weight function defined as,

$$w(j) = e^{-\gamma|t|},$$

then,

$$I_w(\alpha, \beta) = \frac{4\gamma\alpha\beta}{\gamma^4 + (\alpha^2 - \beta^2)^2 + 2\gamma^2(\alpha^2 + \beta^2)},$$

therefore,

$$\frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \left[\begin{aligned} &\frac{4\gamma(X_i)(X_k)}{\gamma^4 + (X_i^2 - X_k^2)^2 + 2\gamma^2(X_i^2 - X_k^2)} \\ &+ \hat{p}^2 \left(\frac{4\gamma[X_i+1][X_k+1]}{\gamma^4 + ([X_i+1]^2 - [X_k+1]^2)^2 + 2\gamma^2([X_i+1]^2 + [X_k+1]^2)^2} \right) \\ &- 2\hat{p} \left(\frac{4\gamma[X_i][X_k+1]}{\gamma^4 + ([X_i]^2 - [X_k+1]^2)^2 + 2\gamma^2([X_i]^2 + [X_k+1]^2)^2} \right) \end{aligned} \right].$$

Lastly, we choose a Cauchy weight function and defined it as,

$$w(j) = \left(1 + \frac{j^2}{\gamma^2} \right)^{-1},$$

then,

$$I_w(\alpha, \beta) = \frac{1}{2}\gamma\pi \begin{bmatrix} \cosh(\gamma|\alpha - \beta|) - \cosh(\gamma|\alpha + \beta|) \\ -\operatorname{sign}(\alpha - \beta) \sinh(\gamma(\alpha - \beta)) \\ +\operatorname{sign}(\alpha + \beta) \sinh(\gamma(\alpha + \beta)), \end{bmatrix}$$

therefore,

$$\begin{aligned} &= \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \frac{1}{2}\gamma\pi \begin{bmatrix} \cosh(\gamma|jX_i - jX_k|) - \cosh(\gamma|jX_i + jX_k|) \\ -\operatorname{sign}(jX_i - jX_k) \sinh(\gamma(jX_i - jX_k)) \\ +\operatorname{sign}(jX_i + jX_k) \sinh(\gamma(jX_i + jX_k)), \end{bmatrix} \\ &+ \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \frac{1}{2}\hat{p}^2\gamma\pi \begin{bmatrix} \cosh(\gamma|j[X_i + 1] - j[X_k + 1]|) \\ -\cosh(\gamma|j[X_i + 1] + j[X_k + 1]|) \\ -\operatorname{sign}(j[X_i + 1] - j[X_k + 1]) \sinh(\gamma(j[X_i + 1] - j[X_k + 1])) \\ +\operatorname{sign}(j[X_i + 1] + j[X_k + 1]) \sinh(\gamma(j[X_i + 1] + j[X_k + 1])) \end{bmatrix} \\ &- \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \hat{p}\gamma\pi \begin{bmatrix} \cosh(\gamma|j[X_i] - j[X_k + 1]|) \\ -\cosh(\gamma|j[X_i] + j[X_k + 1]|) \\ -\operatorname{sign}(j[X_i] - j[X_k + 1]) \sinh(\gamma(j[X_i] - j[X_k + 1])) \\ +\operatorname{sign}(j[X_i] + j[X_k + 1]) \sinh(\gamma(j[X_i] + j[X_k + 1])) \end{bmatrix} \end{aligned}$$

where sign is defined as a mathematical function which returns the sign of a real number X ,

$$-1 \quad \text{if } x < 0$$

$$0 \quad \text{if } x = 0$$

$$1 \quad \text{if } x > 0.$$

Bibliography

- Abdous, B. and Berred, A. (2005). Mean Residual Life Estimation, *Journal of Statistical Planning and Inference* **132**(1-2): 3–19.
- Ahsanullah, M. and Aliev, F. (2011). A Characterization of Geometric Distribution Based on Weak Records, *Statistical Papers* **52**(3): 651–655.
- Arnold, B. C. (1980). Two Characterizations of the Geometric Distribution, *Journal of Applied Probability* **17**(2): 570–573.
- Baringhaus, L. and Henze, N. (1992). A Goodness of Fit Test for the Poisson Distribution Based on the Empirical Generating Function, *Statistics & Probability Letters* **13**(4): 269–274.
- Beltrán-Beltrán, J. and O'Reilly, F. (2019). On Goodness of Fit Tests for the Poisson, Negative Binomial, and Binomial Distributions, *Statistical Papers* **60**: 1–18.
- Betsch, S., Ebner, B. and Nestmann, F. (2022). Characterizations of Non-Normalized Discrete Probability Distributions and Their Application in Statistics, *Electronic Journal of Statistics* **16**(1): 1303–1329.
- Bierens, H. J. (1982). Consistent Model Specification Tests, *Journal of Econometrics* **20**(1): 105–134.
- Buechner, M. (1987). A Geometric Model of Vertebrate Dispersal: Tests and Implications, *Ecology* **68**(2): 310–318.

- Bueno-Delgado, M. V., Ferrero, R., Gandino, F., Pavon-Marino, P. and Rebaudengo, M. (2012). A Geometric Distribution Reader Anti-Collision Protocol for Rfid Dense Reader Environments, *IEEE Transactions on Automation Science and Engineering* **10**(2): 296–306.
- Chaubey, Y. P. and Sen, P. K. (1999). On Smooth Estimation of Mean Residual Life, *Journal of Statistical Planning and Inference* **75**(2): 223–236.
- Chinnam, R. B. and Baruah, P. (2004). A Neuro-Fuzzy Approach for Estimating Mean Residual Life in Condition-Based Maintenance Systems, *International Journal of Materials and Product Technology* **20**(1-3): 166–179.
- Coşar, S. and Bellotto, N. (2020). Human Re-identification with a Robot Thermal Camera Using Entropy-Based Sampling, *Journal of Intelligent & Robotic Systems* **98**(1): 85–102.
- Elderton, W. and Wood, G. H. (1945). Cricket Scores and Geometrical Progression, *Journal of the Royal Statistical Society* **108**(1/2): 12–40.
- Evans, R. V. (1967). Geometric Distribution in Some Two-Dimensional Queuing Systems, *Operations Research* **15**(5): 830–846.
- Ferguson, T. S. (1965). A Characterization of the Geometric Distribution, *The American Mathematical Monthly* **72**(3): 256–260.
- Govindarajulu, Z. (1980). Characterization of the Geometric Distribution Using Properties of Order Statistics, *Journal of Statistical Planning and Inference* **4**(3): 237–247.
- Guess, F. and Proschan, F. (1988). 12 Mean Residual Life: Theory and Applications, *Handbook of Statistics* **7**: 215–224.
- Henze, N. (1996). Empirical-Distribution-Function Goodness-of-Fit Tests for Discrete Models, *Canadian Journal of Statistics* **24**(1): 81–93.

- Jiménez-Gamero, M.-D. and Alba-Fernández, M. (2021). A Test for the Geometric Distribution Based on Linear Regression of Order Statistics, *Mathematics and Computers in Simulation* **186**: 103–123.
- Jiménez-Gamero, M.-D. and Batsidis, A. (2017). Minimum Distance Estimators for Count Data Based on the Probability Generating Function with Applications, *Metrika* **80**(5): 503–545.
- Klar, B. (1999). Goodness-of-Fit Tests for Discrete Models Based on the Integrated Distribution Function, *Metrika* **49**(1): 53–69.
- Lai, C. D. and Xie, M. (2006). *Stochastic Ageing and Dependence for Reliability*, Springer Science & Business Media.
- Ley, C. and Swan, Y. (2013). Local Pinsker Inequalities via Stein’s Discrete Density Approach, *IEEE Transactions on Information Theory* **59**(9): 5584–5591.
- López-Blázquez, F. and Miño, B. S. (1998). A Characterization of the Geometric Distribution, *Statistical Papers* **39**(2): 231–236.
- Maiti, S. S. and Biswas, A. (2007). A New Characterization of Geometric Distribution, *Kybernetika* **43**(1): 97–102.
- Marques, M. and Pérez-Abreu, V. (1989). Law of Large Numbers and Central Limit Theorem for the Empirical Probability Generating Function of Stationary Random Sequences and Processes, *Aportaciones Mat., Notas Invest* **4**(2): 100–109.
- Meintanis, S. G. and Iliopoulos, G. (2003). Characterizations of the Exponential Distribution Based on Certain Properties of Its Characteristic Function, *Kybernetika* **39**(3): 295–298.
- Miller, G. L. and Carroll, B. W. (1989). Modeling Vertebrate Dispersal Distances: Alternatives to the Geometric Distribution, *Ecology* **70**(4): 977–986.

- Milošević, B., Jiménez-Gamero, M. D. and Alba-Fernández, M. V. (2021). Quantifying the Ratio-Plot for the Geometric Distribution, *Journal of Statistical Computation and Simulation* **91**(11): 2153–2177.
- Osaki, S. and Li, X.-x. (1988). Characterizations of Gamma and Negative Binomial Distributions (Reliability Application), *IEEE Transactions on Reliability* **37**(4): 379–382.
- Paulson, A. and Uppuluri, V. (1972). A Characterization of the Geometric Distribution and a Bivariate Geometric Distribution, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 297–300.
- Peng, C. and Chen, L. (2021). *Monte Carlo: Automatic Parallelization of Monte Carlo Simulations*. R package version 1.0.
URL: <https://CRAN.R-project.org/package=MonteCarlo>
- Rao Jammalamadaka, S. and Taufer, E. (2006). Use of Mean Residual Life in Testing Departures from Exponentiality, *Nonparametric Statistics* **18**(3): 277–292.
- Rice, J. A. (2006). *Mathematical Statistics and Data Analysis*, Cengage Learning.
- RStudio Team (2021). RStudio: Integrated Development Environment for R, <http://www.rstudio.com/>. Boston, MA: RStudio, PBC.
- Rueda, R. and O'Reilly, F. (1999). Tests of Fit for Discrete Distributions Based on the Probability Generating Function, *Communications in Statistics-Simulation and Computation* **28**(1): 259–274.
- Shaked, M., Shanthikumar, J. G. and Valdez-Torres, J. B. (1995). Discrete Hazard Rate Functions, *Computers & Operations Research* **22**(4): 391–402.
- Shanbhag, D. (1970). The Characterizations for Exponential and Geometric Distributions, *Journal of the American Statistical Association* **65**(331): 1256–1259.

- Singpurwalla, N. D. (2006). The Hazard Potential: Introduction and Overview, *Journal of the American Statistical Association* **101**(476): 1705–1717.
- Sorsa, J., Siikonen, M.-L., Kuusinen, J.-M. and Hakonen, H. (2021). A Field study and Analysis of Passengers Arriving at Lift Lobbies in Social Groups in Multi-Storey Office, Hotel, and Residential Buildings, *Building Services Engineering Research and Technology* **42**(2): 197–210.
- Su, N.-C. and Hung, W.-P. (2018). Characterizations of the Geometric Distribution via Residual Lifetime, *Statistical Papers* **59**(1): 57–73.
- Sudheesh, K. K. and Nair, N. U. (2010). Characterization of Discrete Distributions by Conditional variance, *Metron* **68**(1): 77–85.
- Swartz, G. B. (1973). The Mean Residual Lifetime Function, *IEEE Transactions on Reliability* **22**(2): 108–109.
- Tarone, R. E. (1979). Testing the Goodness of Fit of the Binomial Distribution, *Biometrika* **66**(3): 585–590.
- Yang, G. L. (1978). Estimation of a Biometric Function, *The Annals of Statistics* pp. 112–116.
- Zhang, Z., Tang, X., Huang, Q. and Lee, W.-J. (2021). Preemptive Medium-Low Voltage Arc Flash Detection with Geometric Distribution Analysis on Magnetic Field, *IEEE Transactions on Industry Applications* **57**(3): 2129–2137.
- Zhao, S., Makis, V., Chen, S. and Li, Y. (2017). Evaluation of Reliability Function and Mean Residual Life for Degrading Systems Subject to Condition Monitoring and Random Failure, *IEEE Transactions on Reliability* **67**(1): 13–25.